ACADÉMIE D'AIX-MARSEILLE
UNIVERSITÉ D'AVIGNON ET DES PAYS DE VAUCLUSE

THÈSE

Approche Statistique pour la Reconnaissance Automatique du Locuteur : Informations Dynamiques et Normalisation Bayesienne des Vraisemblances

Présentée et soutenue publiquement le 24 octobre 2000
pour obtenir le grade de Docteur en Sciences
de l'Université d'Avignon et des Pays de Vaucluse

SPÉCIALITÉ : INFORMATIQUE

par

Corinne Fredouille

Composition du jury :

M	Joseph Mariani	DR, Limsi, Paris	Président du Jury
MM	Jean-Paul Haton	PR, LORIA, Nancy	Rapporteur
	Christian Wellekens	PR, Eurecom, Sophia-Antipolis	Rapporteur
MM	Frédéric Bimbot	CR, IRISA, Rennes	Examinateur
	Gérard Chollet	DR, ENST, Paris	Examinateur
MM	Henri Méloni	PR, LIA, Avignon	Directeur de thèse
	Jean-François Bonastre	MC, LIA, Avignon	Co-Directeur de thèse

Laboratoire Informatique d'Avignon

Remerciements

Prologue

Ces trois années de thèse ont été l'occasion de rencontrer et de dialoguer avec un grand nombre de personnes. Ces échanges ont été, pour une grande part, des plus enrichissants tant au niveau scientifique qu'humain. Aussi, je désirerais associer à cette thèse tout ce "petit monde" qui a contribué, de près ou de loin, à faire de ces trois ans une "époque formidable".

ACTE I

Je tiens à remercier chacun des membres de mon jury pour leur présence et leur participation lors de la soutenance de thèse : Monsieur Joseph Mariani pour avoir accepté de présider ce jury de thèse, Messieurs Jean-Paul Haton et Christian Wellekens en tant que rapporteurs de ce travail, pour le temps qu'ils ont consacré à la lecture de ce manuscrit malgré la période estivale ainsi que Messieurs Frédéric Bimbot et Gérard Chollet pour leur participation active à cette soutenance et les critiques constructives qu'ils ont pu émettre.

Je tiens également à remercier Monsieur Henri Méloni, directeur de cette thèse, qui, malgré sa lourde tâche de Président de l'Université d'Avignon et des Pays de Vaucluse, a suivi et soutenu ce travail tout au long de ces trois années.

ACTE II

Je voudrais exprimer toute ma gratitude à Jean-François Bonastre qui a largement contribué à l'accomplissement de ce travail. En tant que co-directeur de cette thèse, Jean-François a démontré toutes les capacités requises pour diriger des recherches, tant au niveau de ses directions scientifiques, de ses conseils ou de sa disponibilité. Il a toujours été présent et motivé pour relancer ce travail dans des moments difficiles.
Je n'ai qu'un seul regret à exprimer : qu'il n'ait pas eu le temps de soutenir son habilitation à diriger des recherches (HDR) avant ma soutenance de thèse.

ACTE III

Au cours de ma deuxième année de thèse, j'ai eu l'opportunité d'intégrer pour une durée de quatre mois l'équipe Speech du centre de recherche UBILAB/UBS à Zurich. Ce séjour des plus enrichissants a été rendu possible grâce à un certain nombre de personnes

auxquelles je voudrais exprimer toute ma reconnaissance : Messieurs Henri Méloni, Jean-François Bonastre et Marc El-Bèze, directeur du LIA en 1999, d'avoir accepté de me libérer durant ces quatre mois ainsi que Monsieur Hans Peter Frei et toute son équipe de m'avoir si bien accueillie au sein du laboratoire de recherche UBILAB/UBS. Finalement, je voudrais remercier tout particulièrement Messieurs Cédric Jaboulet et Jean Hennebert de m'avoir offert cette opportunité, pour leur accueil si chaleureux au sein du groupe Speech et pour la confiance qu'ils m'ont accordée en me laissant travailler sur le projet européen PICASSO.

ACTE IV

Des pauses cafés studieuses aux *brainstorming* fertiles ... ce travail de thèse, c'est aussi le fruit d'une collaboration étroite au sein d'une équipe de RALleurs. De Laurent dit la besace (expatrié à Grenoble) à Olivier, notre petit dernier, en passant par les Teva et Sylvain, sans oublier notre chère Carole qui s'est parfois arrachée les cheveux en relisant nos articles anglophones, tous ont contribué à faire avancer (ou reculer) cette thèse et j'en profite pour leur dire un grand merci !

Je voudrais au passage faire un clin d'œil à tous les membres du Consortium ELISA (passés et présents) avec qui j'ai pris et je prendrai encore un grand plaisir à travailler.

ACTE V

Cette thèse, c'est aussi trois ans d'une vie passés dans une ambiance formidable ... entre le LIA et l'IUP, c'est comme une grande famille où se fêtent les thèses soutenues, les allées et venues des collègues, les naissances, les mariages...
Un grand merci à tout ce petit monde qui m'a fait passé des moments inoubliables !

Dernier ACTE

Je dédie ces dernières lignes à toute ma petite famille, mes parents pour leur soutien et leur amour, mes deux grands frères pour avoir toujours pris soin de leur sœurette, mes adorables belles-sœurs, mes p'tits bout'd'choux Sandie, Bastien, Robby et Camille et finalement Joël pour m'avoir supportée et chouchoutée pendant cette longue période de rédaction.

Résumé

De nos jours, un grand nombre d'applications nécessitent une phase d'authentification de l'utilisateur. Cette authentification peut être réalisée au moyen d'une carte à puce et d'un code confidentiel (retrait à un guichet automatique), au travers d'empreintes digitales (accès sécurisé à des locaux) ou d'empreintes génétiques (domaine juridique) ou encore grâce à la voix (serrure vocale). Dès lors qu'une application est accessible à distance (par le réseau téléphonique par exemple), la voix reste le seul élément disponible pour authentifier une personne.

Cette thèse s'inscrit dans le cadre de la Reconnaissance Automatique du Locuteur dont l'objectif principal est de reconnaître une personne par l'analyse de sa voix. Le premier thème abordé dans ce travail concerne l'utilisation d'informations dynamiques, considérées comme une source potentielle d'informations pour caractériser le locuteur. Un panorama rapide des approches proposées dans la littérature pour le traitement des informations dynamiques met en évidence les limites d'une grande part de ces techniques. Ces limites portent notamment sur l'incapacité de prendre en compte de larges fenêtres temporelles nécessaires à une exploitation correcte des informations dynamiques.

Dans l'objectif de pallier ce problème, nous proposons une approche dynamique originale qui repose sur la concaténation de trames successives de signal de parole et sur la sélection de la part d'information utile spécifique du locuteur. Par sa simplicité de mise en œuvre cette approche permet de manipuler de larges fenêtre temporelles (de l'ordre de 100 milli-secondes) tout en conservant une complexité de calcul raisonnable.

Testée sur une base de données de très bonne qualité (TIMIT : parole lue, mono-session, milieu calme), l'approche dynamique et notamment la procédure de sélection se sont révélées très satisfaisantes conduisant à des améliorations significatives des performances, comparées à l'utilisation d'informations statiques. À l'opposé, testée sur une base de données de moindre qualité (Switchboard : parole conversationnelle, multi-session, milieu téléphonique), la procédure de sélection s'est avérée particulièrement inopérante. Néanmoins, le simple fait de concaténer des trames de signal de parole a permis, sur la base de données Switchboard, d'apporter un gain de performances du système de reconnaissance. Finalement, des expériences mettant en jeu un protocole particulier – basé sur un mélange aléatoire des trames temporelles – a permis de mettre en évidence la prise en compte effective d'informations de nature dynamique par l'approche proposée.

Le deuxième volet de cette thèse s'intéresse au processus de décision de la tâche de Vérification Automatique du Locuteur (VAL). Ce processus qui permet de décider d'accepter ou de rejeter l'identité d'une personne à l'aide de sa voix repose sur la comparaison d'une mesure de vraisemblance à un seuil de décision. Il nécessite une phase

de normalisation de la mesure de vraisemblance afin d'autoriser l'utilisation d'un seuil de décision indépendant du locuteur et fixé *a priori*, préconisé dans le domaine applicatif.

Nous proposons dans cette thèse une technique de normalisation, appelée World+MAP. L'originalité de cette approche repose sur la projection des mesures de vraisemblance à normaliser dans un espace probabiliste. Cette projection permet de doter le seuil de décision d'une signification directement interprétable. Par ailleurs, l'approche World+MAP offre l'avantage de faciliter ostensiblement l'étape de fusion des scores produits par plusieurs reconnaisseurs dans le cadre d'une architecture multi-reconnaisseur. Ce dernier point a pu être vérifié au cours d'une étude préliminaire mettant en jeu une architecture multi-bande.

Testée sur la base de données Switchboard, l'approche World+MAP a démontré des capacités comparables à celles de techniques classiques de normalisation des mesures de vraisemblances tout en proposant des probabilités comme scores de décision.

Table des matières

Abréviations

ACP	Analyse en Composantes Principales
ARV	Auto-Régressif Vectoriel
BT	Bande Totale
BTD	Bande Totale Dynamique
BTS	Bande Totale Statique
CMN	Cepstral Mean Normalization
CMS	Cepstral Mean Subtraction
COR/ROC	Caractéristique opérationnelle du recepteur - Receiver Operating Characteristic
DET	Detection Error Trade-off
DTW	Dynamic Time Warping
EER	Equal Error Rate
EM	Expectation-Maximization
EMV	Estimation du Maximum de Vraisemblance
FA	Fausse Acceptation
FR	Faux Rejet
GMM	Gaussian Mixture Model
HMM	Hidden Markov Model
HTER	Half Total Error Rate
IAL	Identification Automatique du Locuteur
LFCC	Linear Frequency Cepstrum Coefficient
LFSC	Linear Frequency Spectrum Coefficient
LPC	Linear Predictive Coefficient
LPCC	Linear Predictive Cepstrum Coefficient
LVQ	Learning Vector Quantization
MAP	Maximum A Posteriori
MFCC	Mel Frequency Cepstrum Coefficient
MFSC	Mel Frequency Spectrum Coefficient
MLLR	Maximum Likelihood Linear Regression
MLP	Multi-Layer Perceptron
MSSO	Méthodes Statistiques du Second Ordre
RAL	Reconnaissance Automatique du Locuteur
RAP	Reconnaissance Automatique de la Parole
RBF	Radial Basis Frequency
SB	Sous-Bande
SBD	Sous-Bande Dynamique
STM	Stochastic Trajectory Model
TDNN	Time Delay Neural Network
TFPC	Time Frequency Principal Component
TTM	Temporal Trajectory Model
VAL	Vérification Automatique du Locuteur
VQ	Vector Quantization

Notations

$p(a\|b)$	Probabilité de l'événement a sachant l'événement b
$P(a)$	Probabilité *a priori* de l'événement a
X	Un locuteur/client de la base
$\mathcal{X}$	Modèle du locuteur/client X
$\overline{\mathcal{X}}$	Modèle du non-locuteur
$\mathcal{W}$	Modèle du monde
$\mathcal{I}$	Modèle d'un imposteur
N	Nombre de trames dans un segment de parole
$\{y_t\}_{1\leq t\leq N} \equiv y_N$	Séquence de N vecteurs de coefficients instantanés
$\overline{y^{(t)}}$	Vecteur moyen estimé sur $\{y_t\}_{1\leq t\leq N}$
$\mathcal{Y}_{(t)}$	Matrice de covariance estimée sur $\{y_t\}_{1\leq t\leq N}$
$L_{\mathcal{X}}(y_N)$	Vraisemblance pour que le signal de parole y_N soit émis par le modèle $\mathcal{X}$
$LL_{\mathcal{X}}(y_N)$	Log-vraisemblance pour que le signal de parole y_N soit émis par le modèle $\mathcal{X}$
$LR_{\mathcal{X}}(y_N) \equiv LR_{\mathcal{X}}$	Rapport de vraisemblances impliquant les vraisemblances $L_{\mathcal{X}}(y_N)$ et $L_{\mathcal{W}}(y_N)$
$p(X = Y\|LR_{\mathcal{X}})$	Probabilité d'être en présence d'un accès client connaissant le rapport de vraisemblances $LR_{\mathcal{X}}$
$p(X \neq Y\|LR_{\mathcal{X}})$	Probabilité d'être en présence d'un accès imposteur connaissant le rapport de vraisemblances $LR_{\mathcal{X}}$
$P(X = Y)$	Probabilité *a priori* d'un accès client
$P(X \neq Y)$	Probabilité *a priori* d'un accès imposteur
$p(FA)$	Taux de fausse acceptation
$p(FR)$	Taux de faux rejet
$C(FA)$	Coût d'une erreur de fausse acceptation
$C(FR)$	Coût d'une erreur de faux rejet
H_0	Représentation de l'événement : le signal de parole est produit par le modèle client
H_1	Représentation de l'événement : le signal de parole est produit par un autre locuteur

$\mathcal{F}_{X=Y}$	Distribution de probabilités des rapports de vraisemblances calculés lors d'accès clients
$\mathcal{F}_{X\neq Y}$	Distribution de probabilités des rapports de vraisemblances calculés lors d'accès imposteurs
F_{WMap}	Fonction de normalisation relative à l'approche World+MAP
B_i	Bloc temporel
$\{y_t\}_{i\times T+1\leq t<(i+1)\times T}$	Séquence de vecteurs de coefficients composant le bloc B_i
$C_{Client}(p_{F_{WMap}})$	Nombre d'attributions de la probabilité $p_{F_{WMap}}$ à un test client
$C_{Imp}(p_{F_{WMap}})$	Nombre d'attributions de la probabilité p_{FWMap} à un test imposteur
f_T	Fenêtre temporelle ou empan temporel
T	Taille (en nombre de trames) de la fenêtre temporelle
$\{y_k\}_{t\leq k<t+T}$	Séquence de vecteurs instantanés qui composent la fenêtre temporelle
d	Déplacement (en nombre de trames) de la fenêtre temporelle le long du signal de parole
g	Fonction de compression
C	Ensemble de coefficients
$P_i(C)$	Sous-ensemble de coefficients issus de C
$\mathcal{P}$	Espace des sous-ensembles de coefficients $P_i(C)$
$P_{Best}(C)$	Meilleur sous-ensemble de coefficients
$\mathcal{A}$	Algorithme de sélection
$\mathcal{S}$	Critère de sélection du meilleur sous-ensemble
$\mathcal{E}$	Fonction d'évaluation des sous-ensembles

Table des figures

Liste des tableaux

Chapitre 1

Introduction

La parole est, depuis tout temps, le moyen privilégié de communication de l'Homme. Avec la révolution des machines et notamment des ordinateurs, chercheurs et industriels se sont efforcés d'étendre l'usage de la parole à la communication homme - machine. Si les systèmes actuels basés sur le langage naturel sont loins d'assimiler toutes les finesses d'une langue, le Traitement Automatique de la Parole a considérablement progressé ces dix dernières années, notamment dans les domaines de la Reconnaissance Automatique de la Parole (RAP) et de la synthèse.

Le message linguistique, matière première des systèmes de RAP, n'est pas la seule information véhiculée par la parole ou plus exactement par le signal de parole. Des indices sur l'identité, la personnalité, l'état émotionnel (tristesse) ou pathologique (rhume) d'un individu sont également perçus par l'être humain, au travers de la communication parlée. Parallèlement à la RAP, les chercheurs se sont penchés sur le problème de la caractérisation du locuteur à l'aide de sa voix et, en particulier, de la Reconnaissance Automatique du Locuteur (RAL).

La RAL consiste à reconnaître l'identité d'une personne par analyse de sa voix. Néanmoins, d'autres éléments que la voix peuvent être utilisés pour authentifier une personne tels que les empreintes digitales ou les empreintes génétiques. Contrairement à la voix, ces derniers éléments sont une composante du corps humain, ils ne varient pas ou très peu dans le temps et ne peuvent être modifiés sciemment par un individu. De par ces propriétés, le terme de *biométrie* est souvent employé pour les définir et souligne leur très grande fiabilité. Les gestes de parole ne sont, en aucune façon, un élément du corps humain et ne sont pas reproductibles à l'identique dans le temps. Dans ce sens, les dénominations de *biométrie* ou d'*empreinte vocale* ne sont pas appropriées pour caractériser la voix. Cependant, la voix reste pour certaines applications (services accessibles par réseau téléphonique) le seul élément disponible pour authentifier l'utilisateur.

Depuis les premiers travaux dédiés à la RAL, de nombreuses approches ont été proposées dans la littérature – approches analytique, connexionniste, prédictive, statistique, etc. De ce large panel, seule l'approche statistique demeure au premier plan des publications (et des systèmes de RAL) des récentes années. Offrant d'excellentes performances aux systèmes de RAL, elle est généralement considérée comme l'état de l'art dans le domaine. Ainsi, ce travail de thèse s'appuie essentiellement sur l'approche statistique.

Le premier volet de cette thèse (partie I – chapitre 3 à 6) est consacré aux informations dynamiques caractéristiques du locuteur. Les informations de nature dynamique – par opposition aux informations de nature statique – véhiculées par le signal de parole sont une source pertinente pour la Reconnaissance Automatique du Locuteur. Classiquement, le traitement des informations dynamiques s'appuie sur une fenêtre temporelle glissant le long du signal de parole. Cette configuration particulière requiert de la part des approches dynamiques une mise en œuvre plus recherchée.
Une grande part des approches actuelles rencontrent des difficultés majeures liées essentiellement à la complexité de calcul engendrée par les informations dynamiques et à la prise en compte d'une fenêtre temporelle suffisante pour exploiter correctement ces informations.
Dans ce cadre, nous proposons une approche "dynamique" originale et montrons de quelle manière cette approche, très simple à mettre en œuvre, permet de prendre en compte

l'ensemble des informations dynamiques présentes au sein d'une large fenêtre temporelle et de sélectionner, au sein de cet ensemble, la part spécifique du locuteur.

La caractérisation du locuteur ainsi que les différentes variétés d'informations spécifiques du locuteur et véhiculées par le signal de parole sont présentées dans le chapitre 3.
Un panorama des divers traitements des informations dynamiques, issus de la littérature, est proposé dans le chapitre 4. Deux grandes approches sont discutées selon que les informations dynamiques sont prises en compte lors de la phase de paramétrisation ou au niveau de la modélisation. Les limites des méthodes actuelles sont mises en évidence au cours de cette discussion.
L'approche "dynamique", proposée comme une alternative à ces limites, est présentée dans le chapitre 5.
Le comportement de cette approche est évalué au travers de séries d'expériences menées sur deux bases de données différentes – TIMIT et Switchboard – et reportées au chapitre 6.

Le deuxième volet de cette thèse (partie II – chapitre 8 à 11) s'intéresse au processus de décision de la tâche de Vérification Automatique du Locuteur et plus particulièrement à la phase de normalisation des mesures de vraisemblance.
Les mesures de vraisemblance – composantes dominantes du processus de décision – entre les signaux de parole et les modèles clients présentent une très grande variabilité, à l'origine de graves perturbations au sein des systèmes de VAL. Cette variabilité est intrinsèquement liée à la variabilité du signal (due au locuteur, à l'environnement, au message linguistique, etc.). Des techniques de normalisation sont nécessaires pour contrôler cette variabilité des mesures et améliorer ainsi les performances des systèmes.

Le travail exposé dans cette thèse est motivé par deux problématiques différentes.
D'une part, les techniques de normalisation proposées dans la littérature, malgré leur efficacité à réduire la variabilité des mesures de vraisemblance, ne permettent pas de doter le seuil de décision d'une valeur facilement interprétable. L'espace des valeurs de seuil de décision varie alors de 0 à l'infini.
La deuxième motivation concerne les architectures multi-reconnaisseurs, très souvent intégrées au sein des systèmes de RAL. Dans ce type d'architectures, l'une des difficultés majeures repose sur l'étape de fusion des mesures de vraisemblance, produites par chacun des reconnaisseurs, en vue de la décision finale. La présence de mesures dotées d'une signification – tenant compte par exemple de la qualité intrinsèque de chaque reconnaisseur – simplifierait amplement l'étape de fusion.

L'approche originale de normalisation que nous proposons dans cette seconde partie de thèse, intitulée World+MAP, répond à ces trois objectifs : réduction de la variabilité des mesures de vraisemblance, interprétation du seuil et signification des mesures des reconnaisseurs. Basée sur deux concepts fondamentaux – le test d'hypothèses et l'approche Bayesienne – cette normalisation permet de "projeter" les mesures de vraisemblance dans un espace de probabilités dotant chaque mesure d'une interprétation directe dans le domaine probabiliste.

Le chapitre 8 est consacré à la description du processus de décision de la tâche de VAL et des différents acteurs liés à ce processus – seuil de décision et mesures de vraisemblance. Un état de l'art des techniques de normalisation est présenté au

chapitre 9. L'approche originale World+MAP est détaillée au chapitre 10. Finalement, des expériences présentées dans le chapitre 11 ont pour objectif d'évaluer le comportement de l'approche World+MAP au sein d'un système de VAL et notamment sa capacité à réduire la variabilité des mesures de vraisemblance.

Le chapitre 13 présente des résultats de validation des différents travaux exposés dans les parties I et II de cette thèse – approche "dynamique" et approche World+MAP – dans le contexte des campagnes d'évaluations NIST 1999 et 2000.

Enfin, un ensemble de conclusions et de perspectives, exposées dans le chapitre 14, clôturent ce travail de thèse.

Chapitre 2

La Reconnaissance Automatique du Locuteur : des principes aux évaluations des systèmes

---

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

---

1 La Reconnaissance Automatique du Locuteur

1.1 Généralités

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

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

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

Finalement, les applications de la RAL sont principalement liées aux problèmes d'authentification ou de confidentialité.

Niveau de dépendance au texte

Une première classification des systèmes de RAL repose sur le niveau de dépendance au texte. En premier lieu, on distingue généralement les systèmes dépendants du texte des systèmes indépendants du texte. En mode dépendant du texte, la reconnaissance d'une personne est réalisée sur la base d'un message dont le contenu linguistique (mot de passe, phrase...) est connu du système. En mode indépendant du texte, le système de reconnaissance n'a aucune connaissance sur le message linguistique prononcé par la personne.

Concernant le mode dépendant du texte, une terminologie plus fine peut être donnée à un système suivant l'application visée. Celle-ci est inspirée de la littérature ainsi que des travaux de [Eagles, 1995] :

- systèmes à messages fixés : la personne est contrainte de prononcer un message, qu'elle aura fixé au préalable (mots de passe personnalisés : [Jacob et al., 2000], [Kharroubi et al., 2000]) ou qui sera imposé par le système.
- systèmes à messages promptés : un message, différent à chaque nouvelle session de reconnaissance, est imposé par le système sous forme visuelle [Matsui et al., 1994b] ou auditive [Lindberg et al., 1997]. Ces systèmes ont pour première motivation de se protéger des attaques de personnes malveillantes (imposteurs) qui disposeraient d'un enregistrement de la voix d'une personne.
- systèmes à unités segmentales fixées : la personne doit prononcer un message comportant soit une séquence de mots (séquence de chiffres), soit des traits phonétiques (séquence de phonèmes) connus du système.

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

1.2 Différentes Tâches en RAL

L'Identification Automatique du Locuteur et la Vérification Automatique du Locuteur sont les tâches pionnières du domaine de la RAL [Atal, 1976], [Doddington, 1985], [O'Shaughnessy, 1986], [Rosenberg et al., 1991], [Naik, 1994], [Furui, 1994], [Furui, 1997], [Doddington, 1998]. Plus récemment, les besoins applicatifs ont fait naître de nouvelles tâches comme l'Indexation par Locuteur de flux audio [Johnson, 1999], [Delacourt, 2000] ou le Suivi de Locuteurs (ou speaker tracking) [Rosenberg et al., 1998a], [Sonmez et al., 1999], [Bonastre et al., 2000a], [Bonastre et al., 2000b], [Martin et al., 2000] ou de nouvelles variantes telles que la détection d'un locuteur dans une conversation [Przybocki et al., 1999], [Martin et al., 2000].

1.2.1 Identification Automatique du Locuteur

L'Identification Automatique du Locuteur (IAL) est le processus qui consiste à déterminer, parmi une population de locuteurs connus, la personne ayant prononcé un message donné.
D'un point de vue schématique (voir figure 2.1), une séquence de parole est donnée en entrée du système d'IAL. Pour chaque locuteur connu du système, la séquence de parole est "comparée" à une référence caractéristique du locuteur. L'identité du locuteur dont la référence est la plus "proche" de la séquence de parole est donnée en sortie du système d'IAL.
Deux modes sont proposés en IAL : l'identification en ensemble fermé pour lequel on suppose que la séquence de parole est effectivement prononcée par un locuteur connu du système et l'identification en ensemble ouvert pour lequel le locuteur peut ne pas être connu. En mode "ensemble ouvert", le système d'IAL doit décider de la fiabilité de son jugement en acceptant ou rejetant l'identité qu'il a trouvée.

De par son principe - déterminer une identité parmi les identités potentielles – les performances des systèmes d'IAL se dégradent généralement au fur et à mesure que la population de locuteurs augmente.

Applications

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

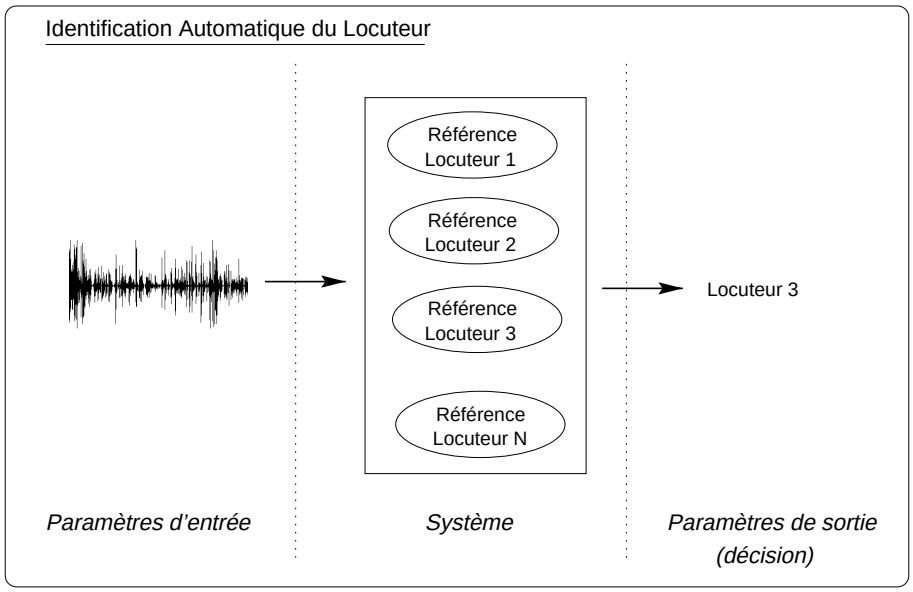

Figure 2.1: La tâche d'IAL. Principe de base de la tâche d'Identification Automatique du Locuteur.

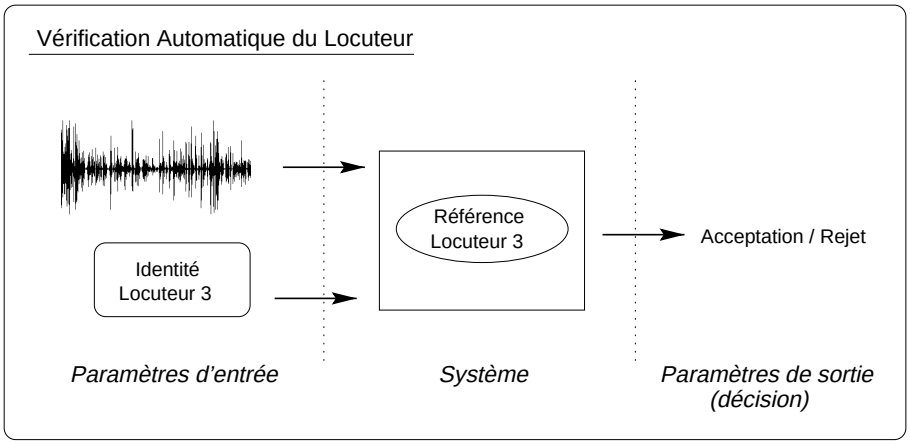

Figure 2.2: La tâche de VAL. Principe de base de la tâche de Vérification Automatique du Locuteur.

1.2.2 Vérification Automatique du Locuteur

La Vérification Automatique du Locuteur (VAL) est le processus décisionnel permettant de déterminer, au moyen d'un message vocal, la véracité de l'identité revendiquée par un individu (figure 2.2). L'identité ainsi que le message vocal constituent les deux entrées du système de VAL. L'identité, nécessairement connue du système, désigne automatiquement la référence caractéristique d'un locuteur. Une mesure de similarité est calculée entre cette référence et le message vocal puis comparée à un seuil de décision. Dans le cas où la mesure de similarité est supérieure au seuil, l'individu est accepté. Dans le cas contraire, l'individu est considéré comme un imposteur et rejeté.

Applications

Les applications de VAL sont multiples et principalement commerciales [Boves, 1998] :

- serrures vocales pour le contrôle d'accès à des locaux ;
- authentification pour l'accès à distance à des données sensibles ou à des services spécifiques à travers le réseau téléphonique (consultations ou transactions bancaires, consultations de bases de données à caractère confidentiel, consultations de boîtes vocales, télé-achat, etc) ;
- protection de matériel contre le vol (téléphones portables, voitures, etc.) ;
- incarcération à domicile nécessitant une authentification régulière du prévenu.

1.2.3 Détection de Locuteurs

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

Applications

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

1.2.4 Indexation par Locuteur et ses variantes

La tâche d'Indexation Automatique par Locuteur consiste à cibler les interventions des locuteurs dans un flux audio (figure 2.3). En d'autres termes, indexer un document audio en locuteurs revient à indiquer à quel moment un individu prend la parole et qui est cet individu. La seule entrée d'un système d'indexation est le document audio à indexer. Aucune information n'est donnée au système concernant le nombre de locuteurs présents dans le document ou leur identité. Contrairement aux systèmes d'IAL ou de VAL, les systèmes d'indexation ne détiennent pas de référence pour les locuteurs présents dans un document audio. Leur principe repose généralement sur une phase de segmentation "aveugle" en locuteurs suivie d'une phase de regroupement. Un système d'IAL permet finalement d'identifier les différents locuteurs présents dans le document. La sortie d'un système d'indexation ressemble généralement à la séquence suivante : le locuteur A est intervenu aux instants t_1, t_4, t_6, le locuteur B aux instants t_2, t_5, le locuteur C à l'instant t_3.

La tâche de suivi de locuteurs peut être considérée comme une version simplifiée de l'Indexation par Locuteur d'un flux audio (figure 2.4). Le principe reste le même : déterminer les interventions d'un ou plusieurs locuteurs, appelés locuteurs cibles, dans un flux audio. La simplification réside dans le fait que le système de suivi de locuteurs connaît nécessairement les locuteurs présents dans le document à indexer ou, du moins, ceux dont il doit détecter les interventions. Il possède une référence caractéristique pour chacun des

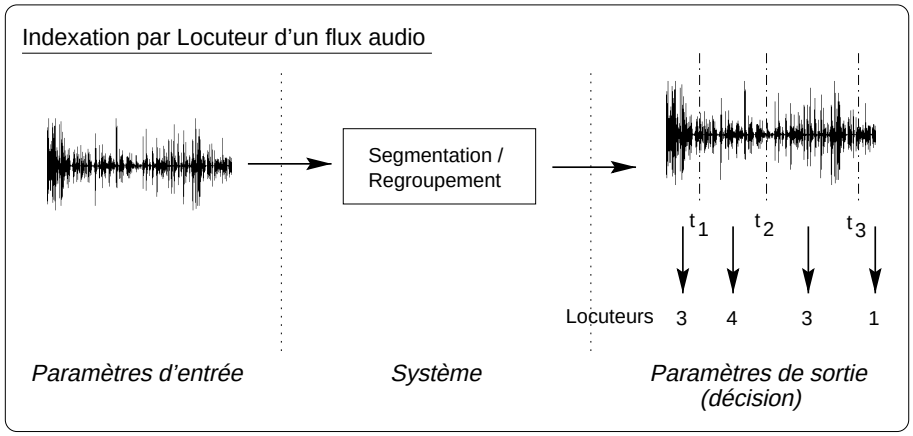

Figure 2.3: La tâche d'Indexation par Locuteur d'un flux audio. Principe de base.

locuteurs. Malgré cette simplification, le suivi de locuteurs reste une tâche très complexe. Trois grandes approches sont recensées dans la littérature :

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

Applications

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

1.2.5 Applications criminalistiques

Un volet que nous n'avons pas encore évoqué est l'utilisation de la RAL dans les domaines judiciaires ou criminalistiques [Hollien, 1990], [Kunzel, 1994], [Boe, 1998], [Champod et al., 1998]. Il s'agit par exemple de rechercher un individu parmi une

[1]Ce découpage selon une taille fixe de blocs est entiérement indépendant des événements acoustiques observés sur le signal de parole

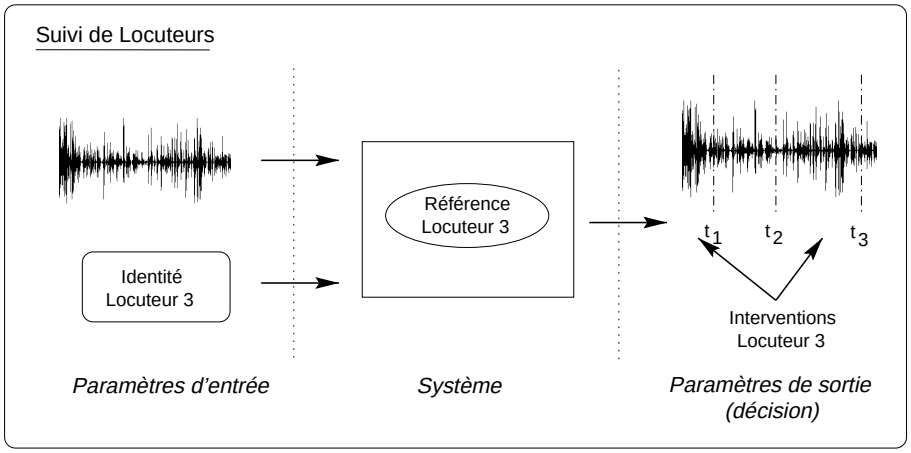

Figure 2.4: La tâche de suivi de locuteurs. Principe de base.

population de suspects potentiels (tâche d'IAL) ou encore de comparer un enregistrement vocal issu d'une écoute téléphonique à la voix d'un suspect potentiel (tâche de VAL).

Dans ce contexte, il est important de souligner que la voix est très souvent assimilée, **à tort**, à une *empreinte vocale* au même titre que les empreintes digitales ou génétiques et peut constituer une preuve dans une procédure pénale. Ce terme d'*empreinte vocale* est une aberration sachant que la voix ne possède pas de caractéristiques qui peuvent la rendre unique [Boe, 1998], [Boe et al., 1999].

1.3 Mise en place d'un système de RAL

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

1.4 Problèmes rencontrés en RAL

Le signal de parole est un signal très complexe où se mèlent informations linguistiques, informations caractéristiques du locuteur, informations relatives au matériel utilisé pour la transmission ou l'enregistrement du signal, etc. En outre, le signal de parole est très redondant. Cette caractéristique est d'ailleurs reconnue pour faciliter la communication entre deux personnes dans un environnement très bruyant ("cocktail party"). Par ces différents aspects, le signal de parole présente une très grande variabilité.

La capacité des systèmes de RAL à différencier plusieurs individus repose essentiellement sur la variabilité inter-locuteur i.e. la disposition du signal de parole à varier entre différents individus. Néanmoins, le signal de parole renferme d'autres types de variabilité qui rendent problématique la tâche de reconnaissance, telles que la variabilité intra-

[2]Dans ce cas, le système ne dispose pas de référence de locuteurs.

locuteur ou la variabilité due au matériel. Par ailleurs, les systèmes de RAL doivent faire face à d'autres difficultés liées davantage au domaine applicatif, comme l'utilisation des systèmes dans des conditions difficiles, les tentatives d'imposture, etc.

Variabilité due au locuteur

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

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

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

Variabilité due au matériel

Le signal de parole est porteur d'informations caractérisant le matériel utilisé lors de sa capture (ex : microphone, combiné téléphonique), de sa transmission (ex : lignes téléphoniques, air ambiant) et de son enregistrement (ex : microphones, convertisseurs). Ces informations apparaissent le plus souvent sous la forme de déformations/dégradations du signal de parole. Ces déformations sont différentes selon le type de matériel utilisé.

Si la bande téléphonique est reconnue pour dégrader les performances des systèmes de RAL, elle n'est pas la seule responsable. En effet, de nombreux travaux expérimentaux ont montré que des variations de matériel entre les phases d'apprentissage et de test sont à l'origine de graves dégradations des performances [Van Vuuren, 1996]. Par exemple, dans [Reynolds, 1996] et [Auckenthaler et al., 2000], il est démontré que des différences de types de combinés téléphoniques entre l'apprentissage et le test sont une des causes de ces dégradations.

Robustesse en environnements difficiles

Comme nous venons de l'évoquer, les environnements téléphoniques mettent à rude épreuve les systèmes de RAL. Néanmoins, d'autres environnements nécessitent de la part des systèmes de RAL une grande robustesse.

Le réseau GSM est considéré ici comme un environnement à part entière, en marge des environnements téléphoniques. Des travaux expérimentaux sur la comparaison du réseau téléphonique classique et d'un réseau GSM font état de différences significatives dans la qualité des signaux [Fissore et al., 1999]. En effet, les signaux transmis par réseau GSM montrent un niveau de bruit bien supérieur (les appels par téléphones mobiles sont souvent effectués dans des endroits plus bruyants que ceux d'un téléphone fixe : voiture, gare, rue), un niveau de voix plus élevé, souvent proche de la saturation entraînant des distorsions au sein du signal ainsi que de potentielles dégradations des signaux dues au codage de la parole. Jusqu'à présent, très peu de travaux ont porté sur la robustesse des systèmes de RAL à travers le réseau GSM [Besacier et al., 2000b], [Quatieri et al., 2000]. La raison est sans doute le manque de bases de données dédiées à cette problématique. Néanmoins, cette tendance est en train de s'inverser avec le développement toujours croissant de la téléphonie portable.

Finalement, les systèmes de RAL doivent renforcer leur robustesse face au bruit ambiant. En effet, d'une manière similaire à la variabilité intra-locuteur ou à la variabilité due aux changements de matériel, la variabilité du niveau de bruit entre apprentissage et test peut susciter une baisse de performances des systèmes de RAL.

Tentatives d'imposture - locuteurs non coopératifs

Selon l'application visée, un système de RAL peut faire l'objet d'attaques d'individus usurpant l'identité de quelqu'un d'autre. Ces attaques (ou tentatives d'imposture) peuvent, par exemple, avoir pour dessein des transactions frauduleuses sur le compte bancaire d'un client ou l'accès à des données confidentielles. Un système de RAL doit par conséquent être robuste face à de telles attaques [Homayounpour, 1995].

Dans un contexte judiciaire, le système de RAL peut être soumis à des locuteurs non-coopératifs i.e. des locuteurs qui ne désirent pas être reconnus par le système. Dans ce cas de figure, les locuteurs tentent fréquemment de transformer leur voix.

Contraintes imposées par le domaine applicatif

Le domaine applicatif et en particulier commercial impose des contraintes fortes quant à l'utilisation d'un système de RAL. Notamment, les sessions d'enrôlement doivent être peu contraignantes pour les clients d'une application. Dans ce sens, elles sont généralement très peu nombreuses et peu espacées dans le temps. Aussi, la quantité de signaux de parole collectés lors de ces sessions s'avère insuffisante pour une bonne estimation des modèles clients.

D'autre part, au cours de ces sessions d'enrôlement, les conditions d'utilisation du système, notamment pour un service accessible par le réseau téléphonique, sont peu variées (appel du domicile du client, du lieu de travail, d'une cabine publique...). Aussi, peu de variabilité due au matériel utilisé est introduite *in fine* dans les signaux d'apprentissage.

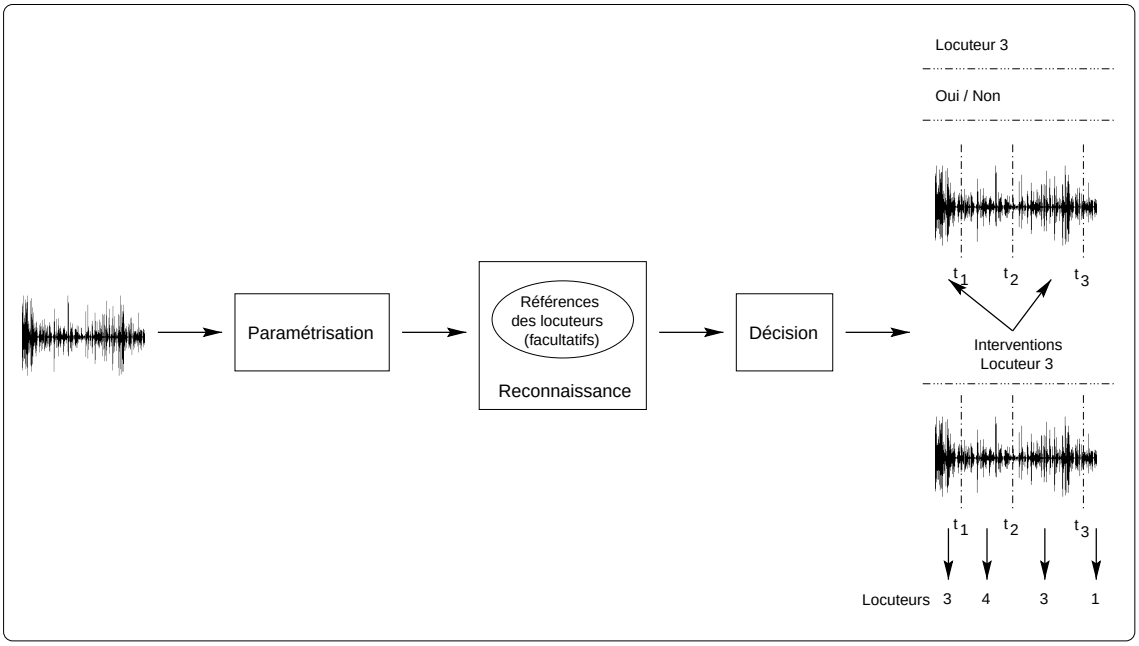

Figure 2.5: Structure d'un système de RAL. Illustration des trois grands processus œuvrant au sein d'un système de RAL.

2 Structure des systèmes de RAL et techniques associées

Un système de RAL, quelle que soit la tâche considérée, se résume à l'enchaînement de trois processus principaux que sont : la paramétrisation, la reconnaissance et la décision (illustration figure 2.5). Contrairement au processus de paramétrisation (généralement basé sur des techniques communes à d'autres domaines comme la RAP), les principes mis en œuvre pour la reconnaissance et la décision sont étroitement liés à la tâche visée. Le processus de reconnaissance est différent selon qu'il repose sur une modélisation des caractéristiques des locuteurs connus du système (modèles clients pour les tâches d'IAL, de VAL ou de suivi de locuteurs) ou non (Indexation par Locuteur d'un flux audio). Dans les sections suivantes, nous nous intéressons au premier cas de figure. Le lecteur pourra se reporter à [Delacourt, 2000] pour un état de l'art du processus de reconnaissance (Segmentation-Regroupement-Identification) employé pour la tâche d'Indexation par Locuteur d'un flux audio. D'une manière similaire, le processus de décision est présenté tâche par tâche.

2.1 Paramétrisation acoustique

Le processus de paramétrisation consiste à extraire du signal de parole les informations pertinentes en vue de la reconnaissance. Le signal de parole, de par sa complexité (multitudes d'informations et redondance), ne peut être exploité directement. Une représentation simplifiée du signal de parole est par conséquent nécessaire. Cette représentation repose généralement sur des vecteurs de paramètres acoustiques, calculés périodiquement sur le signal de parole.

La première étape de la paramétrisation acoustique consiste à décomposer le signal de parole, à cadence régulière (ex : toutes les 10 milli-secondes), en trames de signal (d'une longueur variant généralement de 20 à 31,5 milli-secondes). Un traitement particulier est ensuite appliqué à ces trames afin de produire les vecteurs de paramètres acoustiques.

La littérature propose un grand nombre de traitements selon la nature des informations à extraire du signal de parole. On considère généralement trois grandes classes de paramètres : les paramètres de l'analyse spectrale, les paramètres prosodiques et les paramètres dynamiques. Néanmoins, d'autres classifications sont envisageables. Par exemple, la partie I de cette thèse, dédiée aux traitements des informations dynamiques, préconise de séparer les traitements suivant qu'ils s'intéressent aux informations de nature statique ou dynamique véhiculées par le signal de parole.

2.1.1 Paramètres de l'analyse spectrale

L'analyse spectrale est l'analyse la plus employée en RAL. Les paramètres qui en découlent sont généralement représentatifs des caractéristiques physiques de l'appareil phonatoire (forme du conduit vocal) de chaque individu.

De multiples paramètres ont été étudiés dans la littérature (le lecteur se reportera aux travaux suivants [Reynolds, 1994], [Homayounpour et al., 1994] et [Charlet, 1997] pour une description rapide et une comparaison de différents paramètres). Nous citons ici les plus pertinents en RAL :

- coefficients issus d'une analyse par prédiction linéaire [Grenier, 1977] : LPCC (Linear Predictive Cepstral Coefficients) ou LPC (Linear Predictive Coefficients) ;
- coefficients spectraux issus d'une analyse en banc de filtres[3] : LFSC (Linear Frequency Spectral Coefficients) ou MFSC (Mel[4] Frequency Spectral Coefficients) ;
- coefficients cepstraux issus d'une analyse en banc de filtres : LFCC (Linear Frequency Cepstral Coefficients) ou MFCC (Mel Frequency Cepstral Coefficients).

2.1.2 Paramètres prosodiques

Les paramètres prosodiques illustrent en grande partie le style d'élocution d'un locuteur : vitesse d'élocution (débit), durée et fréquence des pauses, ... ainsi que les caractéristiques de la source glottale (fréquence fondamentale, énergie, taux de voisement,...).

Néanmoins, ces paramètres caractéristiques du locuteur, notamment la fréquence fondamentale et ses variations [Atal, 1976], ne sont pas suffisamment discriminants pour être utilisés seuls dans un système de RAL. Ils sont généralement associés aux paramètres de l'analyse spectrale pour améliorer les performances des systèmes de RAL.

2.1.3 Paramètres dynamiques

Comme nous le soulignerons dans la première partie de cette thèse, l'information dynamique véhiculée par le signal de parole est une source potentielle d'informations

[3] L'analyse en banc de filtres est un traitement particulier fournissant l'énergie d'un signal dans différentes bandes de fréquences.

[4] L'échelle Mel est une échelle des fréquences non-linéaire qui a pour particularité de mieux représenter la sélectivité de l'oreille humaine. Une meilleure résolution est donnée dans les basses fréquences.

pour la caractérisation du locuteur, qui reste encore mal exploitée par les systèmes de RAL.

Les paramètres dynamiques les plus répandus demeurent les coefficients dérivés des vecteurs de paramètres instantanés, appelés coefficients Delta (première dérivée) et Delta-Delta (seconde dérivée) [Furui, 1981], [Soong et al., 1988], [Bernasconi, 1990]. D'autres paramétrisations sont proposées dans la littérature pour exploiter les informations dynamiques du signal telles que l'utilisation des Composantes Principales Temps-Fréquence (TFPC : Time Frequency Principal Components) [Magrin Chagnolleau et al., 1999], la concaténation de trames successives de signal [Hattori, 1992], [Konig et al., 1998], [Fredouille et al., 1998], [Fredouille et al., 2000a]. Un panorama et un bref descriptif de ces différentes approches sont fournis au chapitre 4.

2.2 Reconnaissance - Modélisation et Mesure

Le processus de reconnaissance s'appuie généralement, pour les tâches d'IAL, de VAL et de suivi de locuteurs, sur une modélisation des caractéristiques de chaque locuteur connu du système (modèles de locuteurs ou modèles clients)[5]. Cette modélisation est réalisée à partir des données d'apprentissage collectées au cours des sessions d'enrôlement. Une mesure de similarité est ensuite calculée entre un modèle client et un signal de parole, puis transmise au processus de décision.

On peut distinguer quatre grandes approches pour la construction des modèles clients : les approches vectorielle, statistique, prédictive et connexionniste. Nous présentons ici brièvement les fondements de chacune de ces approches, les techniques qui leur sont associées ainsi que les mesures de similarité utilisées.

2.2.1 L'approche vectorielle

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

Programmation dynamique

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

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

[5]Il est à souligner qu'une petite minorité de techniques appliquées en RAL ne requièrent pas de phase de modélisation comme par exemple la comparaison brute des signaux de parole d'apprentissage et de test.

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 assignée au vecteur de test. La distance d'une séquence de vecteurs de test est obtenue par moyenne des distances minimales attribuées à chacun des vecteurs de test.

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

2.2.2 L'approche statistique

L'approche statistique consiste à représenter une séquence de vecteurs acoustiques issus de la paramétrisation par des statistiques à long terme. Les premiers travaux suggèrent d'utiliser les paramètres du spectre moyen à long terme comme seul modèle des locuteurs [Pruzansky, 1963]. Lors de la reconnaissance, le spectre moyen estimé sur les vecteurs de test est comparé, à l'aide d'une distance spectrale, au spectre moyen issu de l'apprentissage.

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

Méthodes Statistiques du Second Ordre

Le principe des Méthodes Statistiques du Second Ordre (MSSO) est de représenter une séquence de vecteurs acoustiques par une distribution gaussienne multi-dimensionnelle [Bimbot et al., 1995]. Le modèle d'un locuteur se résume alors par le triplet $\{\overline{x}, \mathcal{X}_0, M\}$ où $\overline{x}$ est un vecteur moyen, $\mathcal{X}_0$ est une matrice de covariance, tous deux estimés à partir de la séquence de M vecteurs acoustiques.

Les MSSO sont généralement associées à des mesures de similarité particulières en vue de la reconnaissance. Ces mesures ont pour particularité de faire intervenir le triplet $\{\overline{y}, \mathcal{Y}_0, N\}$. Ce dernier est estimé sur la séquence de vecteurs de test de manière analogue au triplet $\{\overline{x}, \mathcal{X}_0, M\}$. Les mesures reposent ainsi essentiellement sur une ressemblance entre les matrices X_0 et Y_0 [Gish et al., 1986], [Bimbot et al., 1995], [Magrin Chagnolleau et al., 1996].

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

$$\begin{aligned} \mu_{G_\beta}(X,Y) &= \frac{M.\mu_G(X,Y) + N.\mu_G(Y,X)}{M+N} \\ \mu_G(X,Y) &= \frac{1}{p} tr(\mathcal{Y}_0 {\mathcal{X}_0}^{-1}) - \frac{1}{p} log\left(\frac{det(\mathcal{Y}_0)}{det(\mathcal{X}_0)}\right) \\ &\quad + \frac{1}{p}(\overline{y} - \overline{x})^T \mathcal{X}_0^{-1}(\overline{y} - \overline{x}) - 1 \end{aligned} \tag{2.1}$$

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

Mélange de gaussiennes

Un moyen de pallier ce problème (variations locales moyennées par les MSSO) est de considérer les modèles à mélanges de gaussiennes multi-dimensionnelles (Gaussian Mixture Model : GMM) [Reynolds, 1992], [Reynolds, 1995], [Reynolds et al., 2000]. Dans ce contexte, une séquence de vecteurs acoustiques d'apprentissage est représentée par un mélange de gaussiennes i.e. une somme pondérée de M distributions gaussiennes multi-dimensionnelles, chacune caractérisée par un vecteur moyen et une matrice de covariance. Lors de l'apprentissage, les paramètres des modèles clients (vecteur moyen $\overline{x_i}$, matrice de covariance Σ_i, pondération p_i de chaque distribution gaussienne) sont généralement estimés à l'aide de l'algorithme EM (Expectation-Maximization) [Dempster et al., 1977] couplé à l'approche par Estimation du Maximum de Vraisemblance (EMV).
Lors de la reconnaissance, la mesure de similarité entre un modèle client et une séquence de vecteurs de test repose à nouveau sur l'approche EMV.

La vraisemblance pour qu'un vecteur de test, y_t, soit produit par le mélange de gaussienne $\mathcal{X}$ s'exprime par :

$$\begin{aligned} L(y_t|\mathcal{X}) &= \sum_{i=1}^{M} p_i.L_i(y_t) \\ L_i(y_t) &= \frac{1}{(2\pi)^{D/2}|\Sigma_i|^{1/2}} exp\left\{-\frac{1}{2}(y_t - \overline{x_i})^T(\Sigma_i)^{-1}(y_t - \overline{x_i})\right\} \end{aligned} \tag{2.2}$$

où p_i, $\overline{x_i}$ et $\sum_i$ représentent, respectivement, le poids, le vecteur moyen (de dimension D) et la matrice de covariance (de dimension $D*D$) de la $i^{ème}$ distribution gaussienne.

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

Modèles de Markov cachés

Empruntés à la RAP [Rabiner, 1989], les modèles de Markov cachés (Hidden Markov Models : HMM) permettent de caractériser les variations temporelles du signal de parole. Ils reposent sur une machine à états[6], i.e. une succession d'états associés à des probabilités de transition d'un état à l'autre. Une ou plusieurs distributions de probabilité associées à chaque état caractérisent les probabilités d'émission des vecteurs acoustiques par un état. Lors de la reconnaissance, la vraisemblance pour qu'une séquence de vecteurs de test soit issue de la chaîne de Markov est calculée.

NB : Les mélanges de gaussiennes peuvent être considérés comme un modèle de Markov caché à un seul état. De même, la quantification vectorielle décrite précédemment est souvent interprétée comme une dégénérescence des modèles de Markov cachés à un seul état pour lequel les probabilités d'émission sont remplacées par des mesures de distance [Furui, 1995], [Pierrot, 1998].

De par leur principe, les modèles de Markov cachés s'appliquent parfaitement au mode dépendant du texte, obtenant d'excellents résultats [Rosenberg et al., 1991], [De Veth et al., 1994]. En revanche, l'utilisation des modèles HMM en mode indépendant du texte n'améliore pas les performances obtenues par des modèles plus simples à base de GMM [Reynolds et al., 2000].

2.2.3 L'approche connexionniste

L'approche connexionniste, telle que nous l'entendons ici, repose sur la discrimination entre locuteurs. Elle consiste à fournir à un réseau de neurones un ensemble de signaux de parole issus d'une population de locuteurs clients afin que ce dernier apprenne comment discriminer un locuteur des autres. L'approche connexionniste se résume, par conséquent, à une tâche de classification. Un modèle client se présente sous la forme d'un ou plusieurs réseaux de neurones pour lequel la séquence de vecteurs d'apprentissage du client concerné ainsi que celles des autres clients du système sont fournies en entrée. Différents types de modèles de réseaux sont proposés dans la littérature : MLP[7] [Oglesby et al., 1990], RBF[8] [Oglesby et al., 1991], [Frederickson et al., 1994] et LVQ[9] [Bennani et al., 1990]. Lors de la reconnaissance, la vraisemblance pour qu'une séquence de vecteurs de test soit produite par un réseau de neurones est calculée.

Le principal inconvénient de l'approche connexionniste est la complexité d'apprentissage. En outre, elle pose le problème de l'ajout d'un nouveau client qui nécessite dans la majorité des mises en œuvre le ré-apprentissage de tous les modèles. En effet, une nouvelle phase de classification est nécessaire afin de prendre en compte le nouveau client au sein du processus de discrimination entre locuteurs.

2.2.4 L'approche prédictive

L'approche prédictive repose sur le principe qu'une trame de signal peut être prédite par la seule observation des trames précédentes. De par ce concept, cette approche est considérée dans la littérature comme une approche dynamique i.e. une approche tenant compte des informations dynamiques véhiculées par le signal de parole. Elle s'appuie

[6] encore appelée chaîne de Markov.
[7] Multi-Layer Perceptron.
[8] Radial Basis Functions.
[9] Learning Vector Quantisation.

principalement sur l'estimation d'une fonction de prédiction, propre à chaque locuteur et apprise sur les signaux d'apprentissage. Lors de la reconnaissance, une erreur de prédiction peut être calculée entre une trame prédite (par la fonction de prédiction) et la trame réellement observée dans la séquence de test. L'erreur de prédiction moyenne constitue alors la mesure de similarité entre le signal de test et le modèle de locuteur (fonction de prédiction). Une autre solution envisagée est d'estimer une fonction de prédiction sur la séquence de test et de la comparer, à l'aide d'une distance, à la fonction de prédiction estimée lors de l'apprentissage.

Deux grandes techniques sont rattachées à l'approche prédictive : les modèles ARV [Grenier, 1980], [Bimbot et al., 1992], [Montacié et al., 1992], [Griffin et al., 1994], [Magrin Chagnolleau et al., 1996] et les réseaux prédictifs [Hattori, 1992], [Artières et al., 1993], [Bennani et al., 1994], [Paoloni et al., 1996]. Ces deux techniques sont détaillées dans le chapitre 4.

2.3 Normalisation des mesures de similarité

Pour les tâches de VAL et de suivi de locuteurs, un processus de normalisation des mesures de similarité, fournies par le processus de reconnaissance s'avère nécessaire pour améliorer la robustesse de la décision finale. Ce processus de normalisation fait l'objet de la deuxième partie de cette thèse où il sera discuté en détail.

2.4 Décision et mesure des performances

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

2.4.1 Identification Automatique du Locuteur

En identification, un signal de test est comparé à toutes les références de locuteurs connus du système, résultant en un ensemble de mesures de similarité en entrée du processus de décision. Aussi, ce processus a pour tâche de rechercher la mesure de similarité maximale (ou minimale dans le cas de mesures de distance) et de désigner l'identité du locuteur correspondant.

Dans ce contexte, la mesure des performances d'un système d'IAL est généralement donnée en termes de taux d'identification correcte. Ce taux s'obtient par la formulation suivante :

$$Taux\ d'identification\ correcte = \frac{\#\ tests\ ayant\ amené\ à\ une\ identification\ correcte}{\#\ tests\ total} \tag{2.3}$$

2.4.2 Vérification Automatique du Locuteur

En vérification, le processus de décision, détaillé au chapitre 8, consiste à comparer la mesure de similarité entre le signal de test et le modèle client à un seuil de décision. Si la mesure est supérieure au seuil, le locuteur est considéré comme un client et accepté. Dans le cas contraire, le locuteur est considéré comme un imposteur et rejeté.

Plusieurs solutions sont proposées dans la littérature pour mesurer les performances d'un système de VAL. Toutes s'appuient sur deux erreurs caractéristiques des systèmes de VAL que sont :

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

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

$$\begin{aligned} p(FA) &= \frac{\#\ tests\ ayant\ amené\ à\ une\ fausse\ acceptation}{\#\ tests\ imposteurs} \\ p(FR) &= \frac{\#\ tests\ ayant\ amené\ à\ un\ faux\ rejet}{\#\ tests\ clients} \end{aligned} \tag{2.4}$$

Performances intrinsèques du système
Les performances intrinsèques d'un système de VAL sont estimées en comparant les mesures de similarité, issues de tests clients et imposteurs, à différentes valeurs de seuil. À chaque valeur de seuil est associé un couple $(p(FA), p(FR))$. L'ensemble des couples obtenus se représente sous la forme d'une courbe COR – Caractéristique Opérationnelle du Récepteur – (Receiver Operating Characteristic : ROC) [Oglesby, 1995] ou d'une courbe DET (Detection Error Tradeoff) [Martin et al., 1997] sur lesquelles les taux $p(FA)$ sont donnés en fonction des taux $p(FR)$.

Les performances intrinsèques d'un système sont très souvent résumées par un point particulier des courbes ROC et DET indiquant un taux identique d'erreurs de fausse acceptation et de faux rejet : taux d'égale erreur (Equal Error Rate : EER).

Performances d'exploitation du système
Dès lors que le seuil de décision est fixé *a priori*, les performances d'exploitation d'un système de VAL peuvent être évaluées au moyen du couple $(p(FA), p(FR))$ correspondant. Ce couple peut être exprimé sous la forme d'une moyenne des deux taux, appelée HTER [Bimbot et al., 1998] ou encore sous la forme d'une fonction de coût total du système faisant intervenir les coûts respectifs des erreurs de fausses acceptations et de faux rejets (voir chapitre 8 pour plus de détails sur les fonctions de coût).

2.4.3 Suivi de locuteurs

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

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

Modèle HMM
Lors du processus de décision, un algorithme de type Viterbi attribue chaque bloc à

un état particulier du modèle HMM (état parole, état non-parole ou état locuteur). La décision consiste à accepter tous les blocs attribués à l'état du locuteur cible.

D'une manière similaire aux systèmes de VAL, deux types d'erreurs peuvent être produites par un système de suivi de locuteurs : les erreurs de faux rejet ou de fausse acceptation. Les taux de faux rejet et de fausse acceptation s'expriment ici différemment par :

$$\begin{aligned} p(FA) &= \frac{\#\ trames\ attribuées\ au\ locuteur\ cible}{\#\ trames\ ne\ correspondant\ pas\ au\ locuteur\ cible} \\ p(FR) &= \frac{\#\ trames\ non\ attribuées\ au\ locuteur\ cible}{\#\ trames\ correspondant\ au\ locuteur\ cible} \end{aligned} \tag{2.5}$$

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

2.5 Évaluation des systèmes de RAL

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

Avant chaque nouvelle campagne d'évaluation, un corpus de développement est défini et distribué aux participants, permettant la mise au point des systèmes de RAL. Lors de l'évaluation à proprement parler, chaque laboratoire participant reçoit un nouveau corpus de données comprenant les signaux d'apprentissage[10] et de test pour chacune des tâches à réaliser. Les systèmes de RAL sont alors soumis, en aveugle[11], à diverses séries de tests dont les résultats doivent être retournés en un temps donné (généralement trois semaines).

Ces campagnes sont intéressantes à plusieurs niveaux. Elles permettent en premier lieu d'avoir accès à des sous-corpus de la base de données Switchboard, organisés spécialement pour les tâches de RAL. Ces sous-corpus constituent une base de plus de 1000 locuteurs (hommes et femmes) dans un contexte conversationnel et téléphonique.
D'autre part, il devient facile, sur une base de données commune, de comparer les performances des différents systèmes et d'apprécier les nouvelles tendances et techniques proposées. Par exemple, les techniques GMM ont fait l'unanimité pour la tâche de détection de locuteurs au cours des évaluations de cette année (campagne d'évaluation NIST 2000).

3 Les tendances

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

[10] Cet ensemble est évidemment vide pour la tâche d'Indexation par Locuteur d'un flux audio.
[11] Les résultats des tests ne sont évidemment pas connus des participants.

Adaptation d'un modèle générique

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

Apprentissage incrémental

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

Téléphonie mobile

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

Première partie

Informations dynamiques caractéristiques du locuteur

Cette première partie est consacrée au traitement des informations dynamiques caractéristiques du locuteur présentes dans le signal de parole. Nous présentons différentes approches de la littérature, spécialement conçues ou adaptées pour prendre en compte cette classe d'informations et discutons de leurs limites respectives.

Nous proposons une nouvelle approche basée sur la concaténation de trames successives de parole représentées par des vecteurs acoustiques instantanés. L'avantage majeur de cette méthode est de manipuler, à faible coût, une large fenêtre temporelle (de l'ordre de 100 milli-secondes). Son originalité vient d'une procédure de sélection qui permet, parmi la masse d'informations véhiculée par l'empan temporel considéré, d'extraire l'information dynamique spécifique du locuteur. Dans ce contexte, nous discutons le choix d'un algorithme de sélection adapté ainsi que d'un critère adéquat.

Finalement, des travaux expérimentaux illustrent l'utilisation en RAL de l'approche "dynamique" proposée. Ils montrent, notamment, la nature dynamique de l'information prise en compte par cette approche.

Chapitre 3

Informations dynamiques : présentation, intérêt et problématique

1 Caractérisation du locuteur

L'une des caractéristiques principales du signal de parole est sa complexité. Outre le message linguistique émis par un individu, il transmet l'identité de cet individu, son état émotionnel (colère, anxiété, étonnement...) ainsi que son état pathologique (fatigue, rhume...).

Dans cette thèse, nous nous intéressons évidemment aux informations caractéristiques du locuteur et notamment celles relatives à son identité. Néanmoins, différencier les informations propres au message linguistique de celles relatives à l'identité de la personne est une tâche très délicate sachant que ces sources d'informations sont intimement liées.

1.1 Variabilité inter-individuelle

La reconnaissance du locuteur est fondée sur la variabilité inter-individuelle (encore appelée variabilité inter-locuteur) observée dans le signal de parole. Deux facteurs sont reconnus pour être à l'origine de cette variabilité. D'une part, chaque individu possède un appareil phonatoire qui lui est propre [Wolf, 1972], [Fant, 1973], [Calliope, 1989]. Par exemple, des différences de taille et de forme des conduits vocal et nasal, de longueur et de fréquence de vibration des cordes vocales, de volume des poumons peuvent être relevées au sein d'une population de locuteurs. Une variante d'une ou plusieurs caractéristiques des différents organes impliqués dans le processus de production a pour effet l'émission de signaux de parole différents pour un même message linguistique. Le deuxième facteur à l'origine de la variabilité inter-individuelle fait intervenir non plus l'aspect physiologique de l'appareil phonatoire mais plutôt son utilisation. Chaque individu, durant la phase d'apprentissage de la parole et du langage, met en place un certain nombre de mécanismes menant à la production des sons, mécanismes qui lui sont propres. Ceux-ci sont par exemple guidés par l'origine sociale, culturelle ou géographique des individus. [Homayounpour, 1995] cite l'exemple de la langue française pour laquelle des variations au niveau prosodique sont relevées pour différentes régions (accentuation de la première syllabe des mots pour les gens du nord de la France contrairement aux gens du sud qui prolongent la dernière syllabe).

1.2 Informations caractéristiques du locuteur

Le signal de parole véhicule différentes informations caractéristiques du locuteur. Une classification de ces informations est proposée dans [Bonastre, 1994] :

- Les **informations acoustiques et articulatoires** sont relevées principalement sur les spectres à court terme de voyelles et consonnes formantiques [Fant, 1973], [Mella, 1989]. Elles caractérisent la forme des conduits vocal et nasal ainsi que les propriétés psychomotrices des locuteurs durant le processus de production [Sambur, 1975], [Traunmuller, 1984]. Le suivi des formants et l'étude des transitions formantiques des voyelles associées à différents contextes consonantiques permettent de mettre en évidence les phénomènes de co-articulation[1], spécifiques du locuteur [Calliope, 1989], [Duez, 1995], notamment dans la production des nasales [Su et al., 1974] et des liquides [Nolan, 1983].

[1]Par phénomène de co-articulation, il est entendu la manière dont la prononciation d'un phonème est affectée par la production des phonèmes voisins.

- Les **informations prosodiques** concernent principalement l'intonation, l'accentuation et la distribution des pauses. Elles sont souvent caractérisées par la moyenne et les variations de la fréquence fondamentale [Sambur, 1975], [Rosenberg, 1976], [Corsi, 1982], [Dubreucq et al., 1994].
- Les **informations phonologiques** sont représentatives de la manière dont un locuteur utilise les sons pour coder un message linguistique : le choix des allophones, les liaisons, la réalisation acoustique du e muet – schwa –, etc.
- Les **informations temporelles** sont principalement liées à la vitesse d'élocution, aux délais d'établissement du voisement pendant la tenue des occlusives sonores ou encore aux durées de tenue des occlusives sourdes et sonores en position intervocalique [Mella, 1989].
- Les **informations liées au choix des mots** ont trait à la couverture lexicale (au niveau des mots, mais aussi des structures de phrases, morphèmes, etc.) sur laquelle chaque individu s'appuie pour transmettre un message linguistique. Cette couverture, inhérente à la phase d'apprentissage de la langue, évoluant suivant les contextes, demeure propre à un individu.
- Les **informations psychophonétiques** renseignent sur les attitudes d'un individu que ce soit à long terme – traits de la personnalité – ou ponctuellement – état émotionnel [Scherer et al., 1998], état pathologique.

Parmi ces différentes classes d'informations, toutes ne sont pas exploitables dans le cadre d'un système automatique de reconnaissance du locuteur. Une grande part des informations prosodiques, phonologiques et temporelles nécessite un nombre important d'énoncés pour les phases d'apprentissage et de reconnaissance. Ce même problème se pose pour les informations liées au choix des mots, qui nécessitent de surcroît un système de reconnaissance automatique de la parole.

Les informations psychophonétiques, quant à elles, rendent compte des attitudes et non de l'identité d'un locuteur.

Par ailleurs, des études sur des attaques d'imposteurs (simulées par des imitateurs occasionnels ou professionnels) ont démontré que certains paramètres prosodiques tels que l'intensité ou le débit sont facilement imités [Lummis et al., 1972], [Rosenberg, 1976], [Homayounpour, 1995]. La difficulté d'imitation est en revanche accrue pour d'autres indices comme le contour de la fréquence fondamentale.

Enfin, certains paramètres sont sensibles aux variations pathologiques des locuteurs (rhume, inflammation des voies respiratoires...) notamment les caractéristiques prosodiques et les informations relatives à la cavité nasale.

1.3 Informations statiques vs. dynamiques

Dans le signal de parole, les 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 ou encore les paramètres caractérisant les parties stables des phonèmes (centres phonémiques) ;

- 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).

Par leur nature divergente, ces informations nécessitent des traitements différents pour leur intégration dans un système de RAL. La principale différence est inhérente au degré de complexité croissant dès lors que l'on s'intéresse aux informations dynamiques. Ce degré de complexité rend obligatoire la mise en œuvre de nouvelles méthodes, mieux appropriées à ce type d'informations [Furui, 1981], [Artières et al., 1991], [Montacié et al., 1992], [Hattori, 1992], [Bennani et al., 1994], [Fredouille et al., 1998], [Magrin Chagnolleau et al., 1999], [Besacier et al., 2000a], [Fredouille et al., 2000a] ...

2 Intérêt des informations dynamiques

Au regard des sections précédentes, les informations dynamiques sont une source d'informations potentielle pour caractériser le locuteur au même titre que les informations statiques. La complémentarité des informations statiques et dynamiques a été à plusieurs reprises mise en évidence [Furui, 1981], [Soong et al., 1988], [Xu et al., 1989]. Les informations dynamiques constituent par conséquent un apport non négligeable pour les systèmes de RAL.

D'autre part, les travaux de [Soong et al., 1988] montrent que les informations dynamiques représentées par des coefficients dérivés des spectres instantanés, communément appelés coefficients Delta/Delta-Delta (l'utilisation de ces coefficients sera détaillée dans le chapitre suivant) sont plus robustes que les informations statiques aux différences de canaux de transmission entre les données d'apprentissage et de test. Ces différences de canaux sont la cause, au sein d'un système de VAL par exemple, de graves perturbations en termes de performances (le lecteur se reportera au chapitre 8 pour plus de détails sur les perturbations causées par une différence de canaux de transmission entre données d'apprentissage et de test).

L'intérêt que suscitent les informations dynamiques est cependant tempéré par la complexité de traitement de ce type d'informations. Comme évoqué précédemment, la complexité est sensiblement accrue par rapport à l'exploitation simple des informations statiques.

Chapitre 4

Traitement des informations dynamiques : État de l'art

Ce chapitre est consacré aux traitements des informations dynamiques. Il présente et commente les différentes techniques utilisées dans la littérature pour le traitement des informations dynamiques. Finalement, une réflexion sur ces diverses techniques permet de mettre en évidence les limites de chacune d'elles.

---

1 Introduction

Le traitement des informations dynamiques est généralement localisé au niveau d'une fenêtre temporelle (également appelée empan temporel) composée de trames successives de signal. Cette fenêtre, glissant le long du signal de parole, concentre les informations dynamiques véhiculées par le signal. Lors de la paramétrisation acoustique, le nombre de paramètres à considérer est directement proportionnel au nombre de trames composant la fenêtre temporelle. Cette mise en œuvre particulière est à l'origine de l'accroissement de complexité bien connu du traitement des informations dynamiques par rapport au simple traitement des informations statiques.

Choisir la taille de la fenêtre temporelle est l'une des difficultés majeures inhérentes au traitement des informations dynamiques. L'empan doit être suffisamment étendu pour "capturer" l'ensemble des phénomènes qui lui sont associés. Dans le cas des informations articulatoires, l'empan doit assurer la prise en compte des phénomènes de co-articulation sur plus d'une centaine de milli-secondes. Néanmoins, la quantité de données disponibles ainsi que la complexité de traitement limitent la taille de la fenêtre. Plus la taille augmente, plus la complexité de traitement et la redondance des informations sont accrues. Aussi, la taille de la fenêtre temporelle doit être un compromis entre quantité de données à traiter et qualité de l'information dynamique capturée.

La littérature propose plusieurs techniques pour le traitement des informations dynamiques. Une première classification permet de distinguer deux grandes catégories de techniques :

- les techniques appliquées durant la phase de paramétrisation ;
- les techniques appliquées durant la phase de modélisation.

2 Notations

Soit un signal de parole s représenté par une suite de N vecteurs de coefficients acoustiques $\{y_t\}_{1\leq t\leq N}$. Chaque vecteur caractérise une trame du signal et a pour dimension p. Dans le traitement des informations dynamiques, une fenêtre temporelle, f_T, est alors définie par (voir figure 4.1) :

- T sa taille (fixe) en nombre de trames ;
- $\{y_k\}_{t\leq k<t+T}$ la séquence de vecteurs de coefficients issus de s qui la composent à l'instant t ;
- d le déplacement, en nombre de trames, de la fenêtre temporelle le long du signal ; un recouvrement entre deux fenêtres consécutives étant autorisé.

Ces notations sont utilisées dans la suite de ce chapitre ainsi que dans les chapitres suivants.

3 Informations dynamiques et paramétrisation

Un panorama non exhaustif des techniques utilisées durant la paramétrisation montre deux tendances pour le traitement des informations dynamiques :

- la première tendance consiste à extraire de manière explicite des coefficients dynamiques de la fenêtre temporelle. Elle peut se résumer par l'application d'une fonction g sur

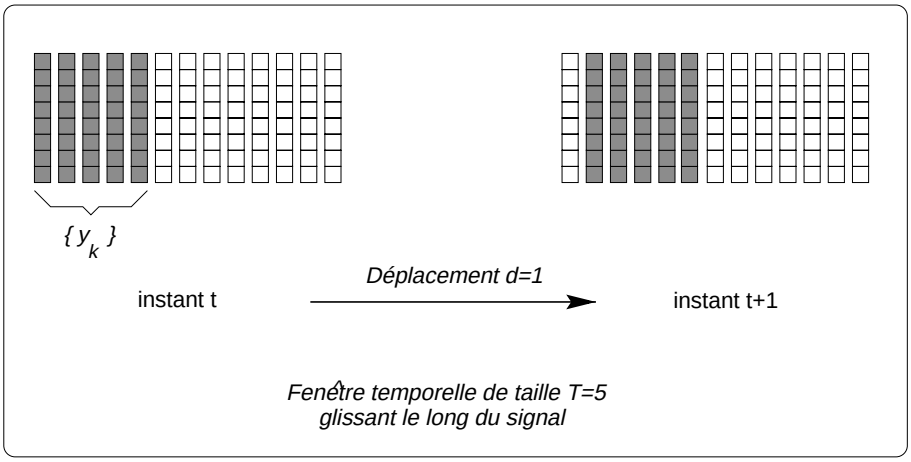

Figure 4.1: Fenêtre temporelle et notations. Illustration d'une fenêtre temporelle de taille T, composée de la séquence de vecteurs $\{y_k\}$ et se déplaçant le long du signal d'une trame (d=1).

la séquence de vecteurs $\{y_k\}_{t \leq k < t+T}$ composant la fenêtre temporelle f_T. Le résultat de cette fonction est un vecteur de coefficients dynamiques (tendance 1 sur la figure 4.2).
Cette approche se retrouve notamment dans le calcul des dérivées des coefficients instantanés ou le suivi de trajectoires temporelles détaillés dans les sections suivantes.

- la deuxième tendance repose sur la prise en compte de toute la fenêtre temporelle sans extraction explicite de coefficients dynamiques (tendance 2 sur la figure 4.2). La concaténation de trames successives, détaillée en section 3.3, s'apparente à cette tendance.

3.1 Dérivées des coefficients instantanés

L'approche la plus répandue dans la littérature a été proposée par [Furui, 1981]. Basée sur les dérivées première et seconde (par rapport au temps) des coefficients instantanés, cette paramétrisation permet de représenter la vitesse et l'accélération de ces derniers.

Pour simplifier le calcul, une approximation des dérivées première et seconde est généralement obtenue à l'aide de fonctions polynomiales comme le montre l'équation 4.1 pour le calcul des coefficients issus de la dérivée première (coefficients Delta). Cette même équation sera appliquée sur les coefficients Delta afin d'obtenir les coefficients issus de la dérivée seconde (coefficients Delta-Delta).

$$\frac{\delta c_m(t)}{\delta t} \approx \Delta c_m(t) = \frac{\sum_{k=-K}^{K} k.c_m(t+k)}{\sum_{k=-K}^{K} k^2} \tag{4.1}$$

où $c_m(t)$ représente le coefficient à dériver, $\Delta c_m(t)$ désigne le coefficient Delta et K est relatif à la taille de la fenêtre temporelle de longueur $2K+1$ trames sur laquelle les coefficients dérivés sont calculés (voir figure 4.3).

Le facteur K a fait l'objet de plusieurs études expérimentales afin de connaître sa valeur optimale. [Furui, 1981] estime qu'une longueur de fenêtre de 90 milli-secondes, soit

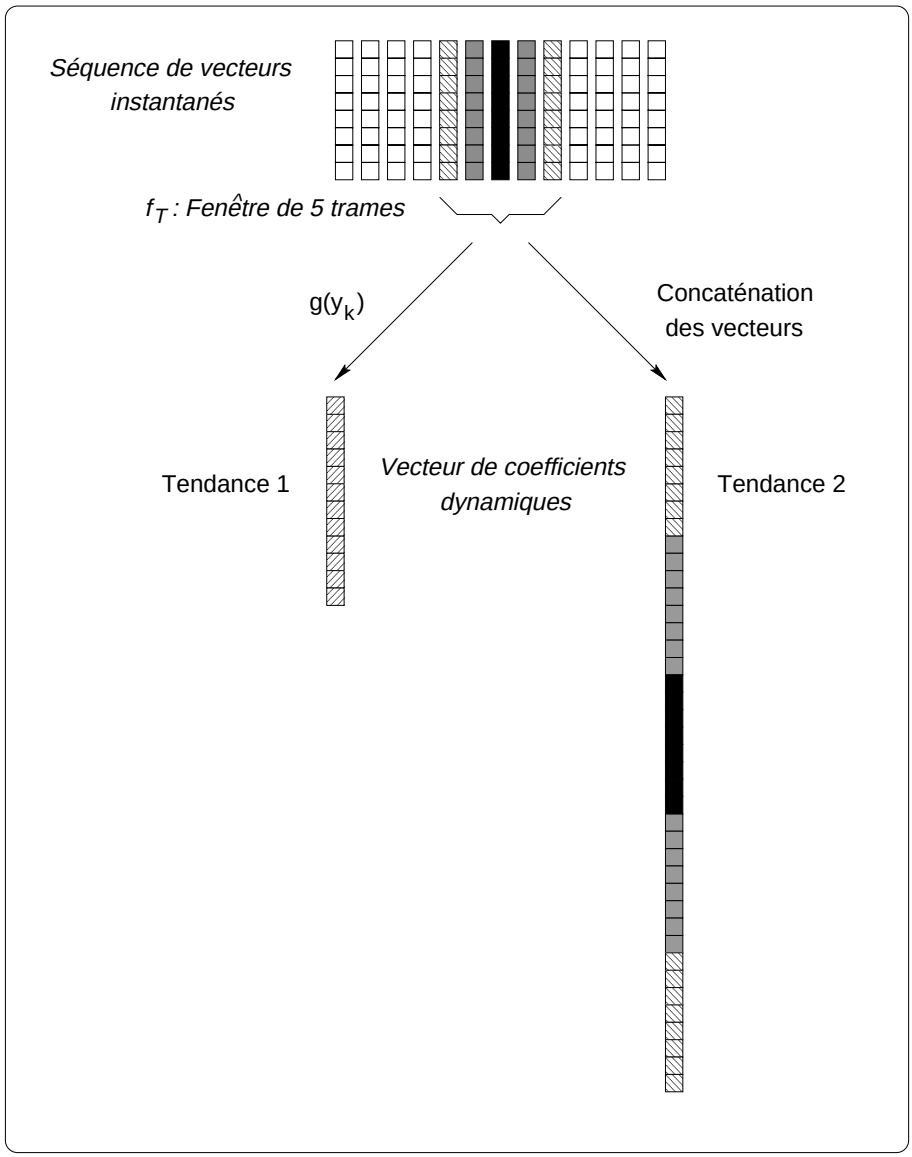

Figure 4.2: Informations dynamiques et paramétrisation. Illustration des deux tendances observées pour le traitement des informations dynamiques durant la phase de paramétrisation. La tendance 1 reflète l'extraction explicite d'informations dynamiques ($g(y_k)$) véhiculées par une fenêtre temporelle de 5 trames. La tendance 2 ne fait état d'aucune extraction. Les vecteurs de coefficients représentant chacune des trames de la fenêtre sont concaténés.

un facteur $K = 4$ pour des trames d'une durée de 10 milli-secondes, est un bon compromis entre performances et complexité de calcul. [Soong et al., 1988] montrent qu'une longueur de fenêtre entre 105 milli-secondes et 165 milli-secondes ($K = 3$ à $K = 5$ pour des trames de 15 milli-secondes) semble optimale en termes de performances pour la tâche d'identification du locuteur. Finalement, une fenêtre de 135 milli-secondes ($K = 4$ pour des trames de 15 milli-secondes) est considérée comme optimale dans [Bernasconi, 1990]...

Dans ces différentes études, l'efficacité des coefficients dynamiques (Delta et Delta-Delta) a été démontrée. Les auteurs parviennent cependant à des conclusions contradic-

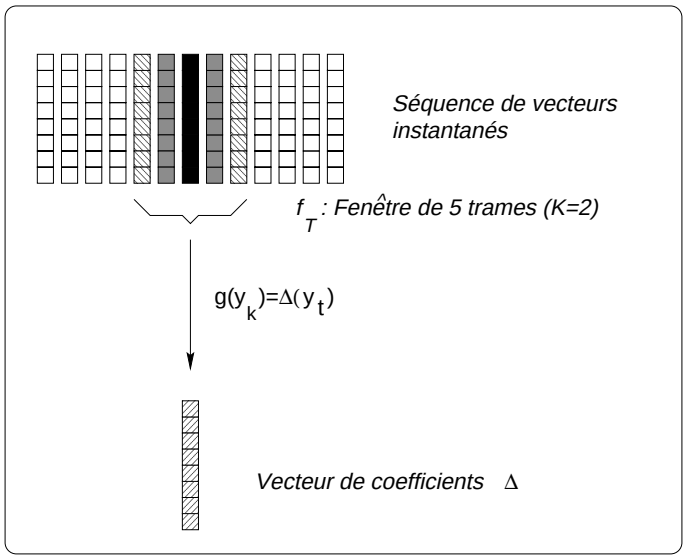

Figure 4.3: Coefficients Delta. Estimation d'un vecteur de coefficients Delta à partir d'une fenêtre de 5 trames.

toires quant à la corrélation entre coefficients dynamiques et statiques. En effet, dans [Soong et al., 1988] une faible corrélation est observée entre ces deux ensembles de coefficients amenant les auteurs à conclure sur une certaine complémentarité de ces derniers. Cette déduction est confirmée par une amélioration des performances d'identification du locuteur lors de la combinaison des deux ensembles, comparée à leur utilisation individuelle. À l'opposé, aucune amélioration des performances d'identification du locuteur n'est observée dans [Bernasconi, 1990] en combinant les coefficients statiques et dynamiques. Les auteurs interprètent ce résultat par l'existence d'une forte corrélation entre ces deux ensembles de coefficients.
En fait, cette différence s'explique en regardant les conditions expérimentales. Les tests d'identification dans [Bernasconi, 1990] et [Soong et al., 1988] ont été menés respectivement sur des signaux de parole de bonne qualité et des enregistrements téléphoniques. Les coefficients dynamiques sont connus pour être plus résistants aux différences de canaux de transmission (le lecteur se reportera au chapitre 9 section 1.4 pour des informations complémentaires concernant la robustesse des coefficients Delta) entre les données d'apprentissage et de test [Furui, 1981], [Soong et al., 1988]. Par conséquent, l'utilisation des coefficients statiques combinés aux coefficients dynamiques permet d'améliorer plus aisément les performances d'identification dans un environnement téléphonique.

3.2 Suivi de trajectoires temporelles ou de formants

Dans [Magrin Chagnolleau et al., 1999], le principe de base est d'appliquer un filtre temps-fréquence sur la séquence de vecteurs de coefficients $\{y_k\}_{t \leq k < t+T}$ composant la fenêtre temporelle. L'objectif de ce filtre temps-fréquence est d'extraire des composantes dynamiques (Time-Frequency Principal Components : TFPC), spécifiques du locuteur, résumant l'évolution spectrale de la séquence de vecteurs $\{y_k\}$.
Le filtre temps-fréquence est dépendant du locuteur et estimé par application d'une Analyse en Composantes Principales (ACP) sur les données d'apprentissage. Ce filtre est ensuite appliqué sur les données d'apprentissage et de test. Deux nouveaux ensembles de

vecteurs de coefficients résultent de ce procédé, sur lesquels des techniques de modélisation et de reconnaissance classiques peuvent être appliquées.

[Konig et al., 1998] proposent une approche très similaire pour extraire des composantes dynamiques spécifiques du locuteur. Dans ces travaux, l'ACP est remplacée par une analyse discriminante non linéaire, réalisée au moyen d'un perceptron multi-couches (MLP). Le perceptron a pour tâche de réduire l'espace des coefficients donnés en entrée (la séquence de vecteurs $\{y_k\}_{t \leq k < t+T}$) dans le but de maximiser les performances d'identification du locuteur. Cette transformation est ensuite appliquée sur les données d'apprentissage et de test, notamment pour la tâche de VAL.

3.3 Concaténation de vecteurs de paramètres instantanés

Contrairement aux techniques de paramétrisation citées précédemment, cette approche, proposée dans [Higgins et al., 1986], [Hattori, 1992] ne repose pas sur une extraction explicite des coefficients dynamiques. L'idée est de construire un vecteur de coefficients en concaténant les vecteurs de paramètres $\{y_k\}_{t \leq k < t+T}$ composant la fenêtre temporelle. Aucune transformation des coefficients n'étant appliquée, l'ensemble des informations, statiques comme dynamiques, est représenté dans ce vecteur étendu. Ce vecteur est utilisé lors de la modélisation pour caractériser, par exemple, les corrélations entre les différents coefficients qui le composent.

3.4 Remarque sur les vecteurs de paramètres instantanés

Une paramétrisation classique composée de vecteurs de paramètres instantanés peut être également envisagée pour caractériser l'information dynamique spécifique du locuteur. Dans ce cas, le traitement des informations dynamiques est réalisé durant la phase de modélisation. Ce principe est illustré au travers de différentes techniques présentées dans les sections suivantes.

4 Informations dynamiques et modélisation

Parmi les techniques de modélisation utilisées pour le traitement des informations dynamiques, deux grandes classes se distinguent : les modélisations dites prédictives et les modélisations statiques appliquées à des informations dynamiques. Les modélisations dites prédictives ont été spécialement conçues pour exploiter la notion de dynamique ou d'évolution temporelle au sein des modèles prédictifs. En revanche, les autres techniques de modélisation exploitent uniquement les aspects dynamiques de la paramétrisation acoustique comme décrit précédemment.

4.1 Modèles prédictifs

Les modèles prédictifs reposent sur le principe suivant : *la présence d'une trame dans un signal de parole est conditionnée par les trames précédemment observées* comme illustré par la figure 4.4. En d'autres termes, l'observation d'une séquence de trames de parole permet de prédire la ou les trames suivantes. Les modèles prédictifs se présentent donc sous la forme d'une fonction de prédiction, différente selon les approches envisagées et dépendante du locuteur.

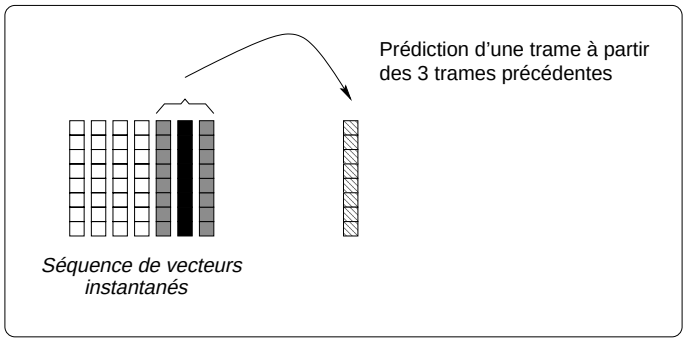

Figure 4.4: Principe de base des approches prédictives.

Lors du test, deux approches sont envisageables :

- Approche 1 : la fonction de prédiction, spécifique à un locuteur, est appliquée sur le signal de test. L'erreur de prédiction moyenne représente alors la "distance" entre les deux signaux de test et d'apprentissage.
- Approche 2 : une nouvelle fonction de prédiction est estimée sur le signal de test et comparée à celle issue de la phase d'apprentissage. Une distance entre les deux fonctions est alors calculée.

Les modèles ARV

Les premiers modèles prédictifs pour la RAL sont proposés par [Grenier, 1980]. S'inspirant de ses travaux sur la prédiction linéaire comme technique de paramétrisation du signal de parole, il propose de modéliser des séquences de vecteurs acoustiques, en l'occurrence des paramètres LPC, par des modèles Auto-Régressifs Vectoriels (modèles ARV). Ces modèles linéaires ont pour objectif de décrire les trajectoires suivies par les coefficients des vecteurs acoustiques et, par conséquent, de décrire les paramètres dynamiques potentiellement caractéristiques du locuteur. D'après [Grenier, 1980], *"le modèle calculé sur la voix d'un locuteur modélise les capacités articulatoires du locuteur, du moins en première approximation"*.

Etant donnée $\{y_t\}_{1 \leq t \leq N}$ une séquence de N vecteurs de paramètres de dimension p, l'évolution de cette séquence est caractérisée par le modèle auto-régressif vectoriel d'ordre q suivant :

$$\mathcal{Y}_t = \sum_{i=1}^{q} A_i . y_{t-i} + e_t \tag{4.2}$$

où q est le nombre de trames utilisées pour la prédiction, $A_{i\{i=1,\ldots,q\}}$ sont des matrices de dimension $p * p$ et e_t une fonction d'excitation représentée par un bruit blanc vectoriel centré correspondant à l'erreur de prédiction.

[Grenier, 1980] expérimente les modèles ARV dans le cadre de l'identification du locuteur. Durant la phase d'identification, les fonctions de prédiction dépendantes du locuteur sont appliquées sur les signaux de test (Approche 1). La minimisation de l'erreur de prédiction moyenne – calculée sur les différentes séquences de vecteurs acoustiques de test – désigne le locuteur à reconnaître. Sur une petite population de locuteurs, les taux d'identification obtenus sont de 100%.

Dans les années 90, de nombreux chercheurs se sont intéressés aux modèles ARV dans le cadre de la RAL. Les premiers travaux proposent d'utiliser l'approche 2 lors de la phase de test. Dans ce contexte, les modèles ARV sont appris sur les signaux de test et comparés, à l'aide d'une distance, aux modèles issus de l'apprentissage [Artières et al., 1991], [Bimbot et al., 1992], [Montacié et al., 1992], [Griffin et al., 1994], [Magrin Chagnolleau et al., 1996]. Des expériences menées en identification du locuteur sur la base de données TIMIT [Campbell et al., 1998] conduisent à de très bonnes performances : un taux d'identification de 98.4 % est obtenu avec des modèles ARV d'ordre 2 sur une population de 420 locuteurs et des signaux de parole de 15 secondes pour l'apprentissage et les tests [Montacié et al., 1992]. Des résultats comparables sont obtenus dans [Le Floch et al., 1996], [Magrin Chagnolleau et al., 1996]. Des expériences en vérification ont fourni des performances d'un même niveau [Montacié et al., 1993].

Dans un même temps, il est reproché aux distances utilisées de ne pas prendre en compte la discrimination entre locuteurs : les distances mesurent l'écart entre un signal de parole et un modèle ARV d'un locuteur donné mais ne visent pas à maximiser l'écart entre les modèles. Pour pallier ce problème, [Montacié et al., 1992b] proposent une nouvelle distance, dite discriminante, ainsi qu'un modèle ARV discriminant (modèle ARVD). Des expériences menées sur l'utilisation de cette distance semblent montrer un léger avantage en faveur de l'approche discriminante.

Les réseaux de neurones prédictifs

Deux techniques connexionnistes – les réseaux récurrents et les réseaux TDNN (Time Delay Neural Network) – sont proposées pour le traitement des informations dynamiques (le lecteur se reportera aux travaux de [Artières, 1995] pour plus d'informations sur ces différentes techniques). Dédiés en premier lieu à la RAP, ces réseaux dits prédictifs ont démontré leur capacité à capturer des invariants dans le temps mais également à caractériser les corrélations entre trames successives de signal [Waibel et al., 1988].

En RAL, une première utilisation des réseaux prédictifs comme classificateurs pour l'identification du genre et du locuteur est proposée dans [Bennani et al., 1991], [Artières et al., 1991]. Dans ce contexte, un seul réseau prédictif est appris sur l'ensemble des données d'apprentissage.

Par la suite, [Hattori, 1992] utilise les réseaux prédictifs en tant que modèles prédictifs. Un modèle prédictif composé d'un ou plusieurs réseaux de neurones[1] est appris pour chaque locuteur. Les expériences dans cette voie ont montré de très bonnes performances, en identification du locuteur, sur des signaux de test de courtes durées (3 secondes en moyenne) [Hattori, 1992], [Artières et al., 1993], [Bennani et al., 1994] ainsi qu'en vérification du locuteur [Paoloni et al., 1996].

[1] L'utilisation de plusieurs réseaux de neurones prédictifs permet de modéliser différentes portions du signal de parole d'un locuteur.

4.2 Modèles statiques appliqués à des coefficients dynamiques

L'extraction explicite des informations dynamiques durant la paramétrisation permet l'application de techniques simples pour la construction des modèles. Dans ce cadre, des techniques dites statiques, généralement appliquées sur des vecteurs acoustiques instantanés, peuvent être utilisées. Par exemple, les coefficients dérivés (détaillés en section 3.1) sont couramment concaténés aux vecteurs instantanés – dont ils sont issus – et modélisés par des méthodes statiques telles que la quantification vectorielle, les modèles de Markov cachés, les mixtures de gaussiennes... Ces techniques sont également utilisées dans le cas de vecteurs de coefficients dynamiques issus des techniques de suivi de trajectoires [Konig et al., 1998], [Magrin Chagnolleau et al., 1999].

Finalement, la concaténation de trames successives lors de la paramétrisation peut également conduire à l'utilisation de techniques statiques pour la modélisation. En effet, la séquence de vecteurs étendus issus de cette paramétrisation peut être assimilée à une simple séquence de vecteurs instantanés et caractérisée par un modèle statique. Des séquences de vecteurs étendus issus de fenêtres temporelles de 3 et 8 trames sont modélisées par quantification vectorielle respectivement dans [Hattori, 1992] et [Higgins et al., 1986].

4.3 D'autres approches

Outre la modélisation prédictive ou l'utilisation de techniques statiques, d'autres approches sont proposées pour modéliser les informations dynamiques.

Dans leurs travaux, [Afify et al., 1998] proposent une adaptation des mixtures de gaussiennes, sous la forme de mixtures de trajectoires stochastiques, pour caractériser non plus des trames instantanées mais des séquences de trames de signal de parole (Stochastic Trajectory Model : STM).

Dans la même optique, un modèle de transitions temporelles (Temporal Trajectory Model : TTM) est présenté dans [Li et al., 1995] pour caractériser les configurations articulatoires spécifiques du locuteur. Pour chaque séquence de T trames successives, une table de probabilités de transition est construite, modélisant la configuration articulatoire inhérente à la séquence de trames. Ces probabilités de transition sont estimées sur l'ensemble du signal de parole. Le modèle global des transitions temporelles est obtenu en moyennant l'ensemble des tables de probabilités de transition estimées individuellement pour chaque séquence de trames.

5 Limites des méthodes actuelles

5.1 Avant-propos

Une grande partie des méthodes présentées dans ce chapitre ont démontré leur potentialité au cours d'expériences diverses et variées de reconnaissance du locuteur, sur des populations plus ou moins larges. Le tableau 4.1 recense quelques résultats de ces expériences à titre indicatif, la diversité des bases de données et des conditions expérimentales ne permettant pas de comparer avec certitude les performances obtenues. Ces techniques liées soit à la phase de paramétrisation soit à la phase de modélisation

Tâche	Niveau Approche	Paramètres	Modèles	Base de données	% de réussite	Références
IAL	Modél.	3 trames successives	**MLP prédictifs**	TIMIT 24 loc.	100	[Hattori, 1992]
IAL	Param.	**3 trames sucessives**	QV[2]	TIMIT 24 loc.	85	[Hattori, 1992]
IAL	Modél.	20 LPCC[3]	**MARV**	TIMIT 630 loc.	99.9	[Montacié et al., 1993]
VAL	Modél.	20 LPCC	**ARV**	TIMIT 630 loc.	99.8[4]	[Montacié et al., 1993]
IAL	Modél.	16 coeff. cepstraux	**ARV**	15 loc. japonais	98	[Griffin et al., 1994]
VAL	Modél.	16 coeff. cepstraux	**ARV**	15 loc. japonais	99.7	[Griffin et al., 1994]
IAL	Modél.	24 coeff. banc filtres	**ARV**	TIMIT 63 loc.	96.8	[Magrin Chagnolleau et al., 1996]
IAL	Modél.	24 coeff. banc filtres	**ARV**	FTIMIT 63 loc.	78.7	[Magrin Chagnolleau et al., 1996]
IAL	Modél.	24 coeff. banc filtres	**ARV**	NTIMIT 63 loc.	48.8	[Magrin Chagnolleau et al., 1996]
IAL	Param.	**filtrage par TFPC**	GMM	Polycost 112 loc.	90.9	[Magrin Chagnolleau et al., 1999]
IAL	Modél.	12 coeff. cepstraux	**TTM**	72 loc. français	98.9	[Li et al., 1995]
VAL	Modél.	12 coeff. cepstraux	**TTM**	72 loc. français	99.5[5]	[Li et al., 1995]
IAL	Modél.	12 MFCC[6]	**STM**	TIMIT 168 loc.	98.3	[Afify et al., 1998]

Tableau 4.1: Performances des approches dynamiques. Taux de réussite de diverses techniques exploitant les informations dynamiques présentes dans le signal de parole pour les tâches d'identification (IAL) ou de vérification (VAL) du locuteur.

(la colonne *Niveau Approche* permet de différencier entre modélisation – Modél. – ou paramétrisation – Param.), interviennent dans des tests d'identification du locuteur (IAL) ou de vérification du locuteur (VAL). Dans les deux cas, les performances obtenues sont données en termes de taux de réussite.

Ces résultats semblent très satisfaisants pour la majorité de ces techniques[7]. Néanmoins, il est intéressant de discuter des limites de certaines approches concernant la prise en compte des informations dynamiques ou leur mise en œuvre.

[2]Quantification Vectorielle.
[3]Linear Predictive Cepstral Coefficients.
[4]Taux inverse de l'EER (Equal Error Rate).
[5]Moyenne des taux corrects d'acceptation et de rejet.
[6]Mel Frequency Cepstral Coefficients.
[7]Il est toutefois à noter que l'essentiel des bases de données impliquées ici, notamment TIMIT, sont composées de signaux de parole de très bonne qualité enregistrés dans un environnement de laboratoire.

5.2 Problématique des paramétrisations dynamiques

En section 3, les procédés d'extraction explicite des informations dynamiques sont assimilés à une fonction $g(\{y_k\}_{t \leq k < t+T})$ (figure 4.2). En considérant les tailles respectives du vecteur dynamique (résultant de l'application de g) et de la fenêtre temporelle dont ce vecteur est issu, une forte réduction de l'espace acoustique est observée. Par exemple dans [Furui, 1981], chaque trame de signal est caractérisée par un vecteur de 10 coefficients cepstraux instantanés. Les informations dynamiques présentes dans une fenêtre temporelle de 9 trames successives et relatives à la vitesse de transition de chacun des coefficients cepstraux sont représentées par un unique vecteur de 10 coefficients Delta, correspondant à un facteur de réduction de 9.

Par cette réduction, la fonction g peut être considérée comme une **fonction de compression** des informations dynamiques.

Lors de cette compression, aucune indication ne peut être donnée sur la qualité des informations dynamiques prises en compte. Il est possible que des informations dynamiques éventuellement utiles pour la caractérisation du locuteur soient perdues lors de ce procédé. Cette supposition peut d'ailleurs expliquer que l'utilisation des coefficients Delta et Delta-Delta nécessite la présence des coefficients statiques (issus des vecteurs instantanés) pour apporter une amélioration des performances d'un système de VAL [Soong et al., 1988].

5.3 Ordre des modèles ARV

Une des difficultés liées à la mise en œuvre des modèles ARV est l'estimation de l'ordre optimal, i.e. le nombre de trames successives nécessaires pour l'apprentissage des modèles ARV. Cet ordre est estimé empiriquement à 2 ou 3 par [Montacié et al., 1992a] et [Griffin et al., 1994]. Pour des ordres supérieurs, il semblerait que l'erreur de prédiction demeure stable alors que les taux de reconnaissance sont nettement dégradés. Ces faibles valeurs d'ordre s'interprètent, d'après [Montacié et al., 1993], par *la simplicité du mode de transition des coefficients qui est facilement approximé par un modèle ARV de faible ordre*. De ce fait, les modèles ARV d'ordre 2 sont généralement utilisés dans la littérature pour le traitement des informations dynamiques.

L'efficacité des modèles ARV d'ordre 2 est néanmoins remise en question par les travaux de [Magrin Chagnolleau et al., 1996]. Ces derniers montrent expérimentalement que les modèles ARV d'ordre 2 obtiennent des performances similaires en utilisant des signaux de parole normaux et des signaux de parole dont la structure temporelle a été détruite par un mélange aléatoire de trames. Ce mélange temporel des trames a pour but de détruire la dynamique du signal de parole et doit par conséquent perturber gravement les performances d'un système de reconnaissance dès lors que ce dernier est basé sur l'utilisation des informations dynamiques. Les expériences conduites par [Magrin Chagnolleau et al., 1996] démontrent que les modèles ARV sont plus enclins à extraire des informations statiques caractéristiques du locuteur que des informations dynamiques. En effet, les performances du modèle ARV d'ordre 2 ne sont pas dégradées après destruction de la structure temporelle du signal.

D'autre part, l'approche consistant à comparer des modèles ARV appris respectivement sur les données d'apprentissage et de test (Approche 2) [Montacié et al., 1992a],

[Bimbot et al., 1992] demande une quantité de données de test suffisante pour une estimation correcte des modèles.

5.4 Limitations des réseaux prédictifs

Les réseaux prédictifs requièrent une importante quantité de données d'apprentissage pour assurer des taux de reconnaissance satisfaisants. En effet, contrairement aux réseaux dits discriminants, la discrimination entre locuteurs n'est plus prise en compte lors de la modélisation. L'apprentissage des modèles ne fait intervenir que les données du locuteur concerné. Le manque de discrimination des réseaux prédictifs est en partie compensé par une augmentation des données d'apprentissage, améliorant la qualité intrinsèque de chaque modèle.

Par ailleurs, le choix d'une architecture (nombre de cellules cachées) optimale pour les réseaux prédictifs reste une phase problématique [Hattori, 1992], [Artières, 1995]. Seules des heuristiques sont proposées dans [Artières, 1995] pour y remédier.

Finalement, la complexité de calcul lors de l'apprentissage des modèles ne garantit pas la convergence vers un optimum global.

Ces divers problèmes expliquent que peu d'études aient porté sur l'utilisation des réseaux prédictifs pour la tâche de vérification. Suite à des expériences menées dans ce contexte, [Hattori, 1994] conclut : *However, the performance is still not enough for practical use and needs further research*[8].

5.5 Redondance au sein des vecteurs étendus

L'utilisation des vecteurs étendus pour caractériser les informations dynamiques pose le problème de la redondance d'informations. Il est bien connu que la parole est un signal très redondant : une information particulière peut être dupliquée sur plusieurs trames successives. Concaténer ces trames revient à répéter cette information plusieurs fois au sein d'un même vecteur étendu.

Ce phénomène de redondance peut être perturbant lors de la modélisation des vecteurs étendus par des techniques statiques, notamment avec l'utilisation d'approches statistiques. En effet, une information redondante mais peu utile pour la caractérisation du locuteur sera fortement mise en valeur par des modèles statistiques.

5.6 Problématique liée à la taille de la fenêtre temporelle

Dans la littérature, la taille de la fenêtre temporelle varie grossièrement entre 100 et 200 milli-secondes pour un traitement efficace des informations dynamiques.
Dans ce contexte, l'utilisation des approches prédictives est exclue en raison de leur difficulté de mise en œuvre, en termes de complexité de calcul et de quantité de données requise pour l'apprentissage des modèles.

[8] Néanmoins, les performances sont encore insuffisantes pour une utilisation pratique et nécessitent un travail de recherche complémentaire.

Par ailleurs, une large taille de fenêtre rend difficile l'utilisation des vecteurs étendus. Le nombre de coefficients de chaque vecteur étendu, proportionnel à la taille de la fenêtre, peut rendre inapplicables les techniques statiques. D'autre part, l'information dynamique utile peut s'avérer inopérante car "noyée" dans la masse importante d'informations présente dans les vecteurs étendus.

6 Conclusions sur les techniques dynamiques

Cet état de l'art révèle une grande variété d'approches pour le traitement des informations dynamiques. La majorité de ces approches ont démontré leur efficacité au cours d'expériences de validation dans des conditions particulières (tableau 4.1). Néanmoins, elles présentent, pour une grande part, des limites qu'il faut considérer :

- Dérivée des vecteurs instantanés, suivi de trajectoires ou de formants : un phénomène de compression des informations est observé (fonction g) n'apportant aucune garantie quant à l'optimalité du traitement des informations dynamiques.
- Modèles prédictifs : la mise en œuvre est inadaptée à l'utilisation d'une large fenêtre temporelle.
- Modèles ARV d'ordre 2 : les informations prises en compte par ce type de modèle ne semblent pas être de nature dynamique mais de nature statique.
- Utilisation des vecteurs étendus : une redondance non négligeable des données ainsi qu'une perte de l'information dynamique utile – noyée dans la masse d'informations présente – est prévisible avec une large fenêtre temporelle.

Chapitre 5

Sélection de l'information dynamique utile

Dans ce chapitre, nous proposons et discutons une nouvelle approche pour le traitement des informations dynamiques. Cette approche, inspirée des techniques de l'état de l'art et de leurs limites, est spécialement adaptée pour répondre au besoin d'une large fenêtre temporelle.

---

1 Motivations

La fenêtre temporelle est introduite dans le chapitre précédent comme un élément essentiel pour le traitement des informations dynamiques. La taille de la fenêtre est estimée empiriquement, selon les travaux, entre 100 et 200 milli-secondes [Furui, 1981], [Soong et al., 1988], [Bernasconi, 1990]. Les limites relatives aux approches prédictives et aux différentes techniques de paramétrisation désignent le couple concaténation/modélisation statique comme le plus apte à assumer une telle longueur de fenêtre, en termes de complexité de calcul. Ce constat est dû notamment à la facilité de mise en œuvre des méthodes statiques et à la faible quantité de données d'apprentissage nécessaire par comparaison avec les modèles prédictifs par exemple.

Ce chapitre propose une nouvelle approche "dynamique" basée sur la concaténation de trames successives et l'application de techniques statiques pour la modélisation des vecteurs étendus. Pour pallier le problème de redondance des données évoqué précédemment, cette approche intègre un processus de sélection de l'information dynamique utile présente dans chaque vecteur étendu.

2 Choix de la fenêtre temporelle

Une étude préliminaire est réalisée concernant la taille de la fenêtre temporelle. Basée sur l'analyse des corrélations entre trames plus ou moins éloignées sur l'échelle temporelle, cette étude a pour objectif de quantifier, en nombre moyen de trames, la propagation sur les trames $t+1$, $t+2$, ..., $t+k$ des effets d'un phénomène observé sur la trame t.

Dans ce cadre, des tests d'identification du locuteur sont menés sur une population réduite de la base de données TIMIT (63 locuteurs). Durant ces tests, le comportement, en termes de performances (taux d'identification), de paires de trames plus ou moins éloignées est étudié. Ces paires de trames, illustrées par la figure 5.1, sont issues d'une fenêtre temporelle fixée à 10 trames glissant le long du signal de parole. Aucun recouvrement n'a lieu entre deux déplacements. Lors de l'apprentissage, les vecteurs instantanés relatifs aux trames y_t et $\{y_{t+k}\}_{1\leq k\leq 9}$ sont concaténés et modélisés pour chacun des locuteurs. La même opération est répétée avec les signaux de test avant comparaison avec les modèles des locuteurs.

Le tableau 5.1 reporte les taux d'identification obtenus avec chacune des paires de trames. Le taux d'identification maximal est atteint avec la combinaison des trames 1 et 8 (59,2 % d'identification correcte).

Ce résultat montre qu'une fenêtre d'une longueur de l'ordre de 10 trames – soit 100 milli-secondes de signal de parole dans ce contexte expérimental – est nécessaire et suffisante pour exploiter les informations dynamiques caractérisées ici par les corrélations entre trames éloignées. Cette constatation est proche des valeurs (de 100 à 200 millisecondes) données dans la littérature.

NB : les taux d'identification obtenus dans ce contexte expérimental sont très faibles comparés aux résultats cités dans la littérature pour la base de données TIMIT. Ces taux s'expliquent par la faible quantité de données utilisée pour l'apprentissage des modèles de locuteurs. Dans ce contexte, des fichiers d'apprentissage de 6 secondes sont utilisés. Le

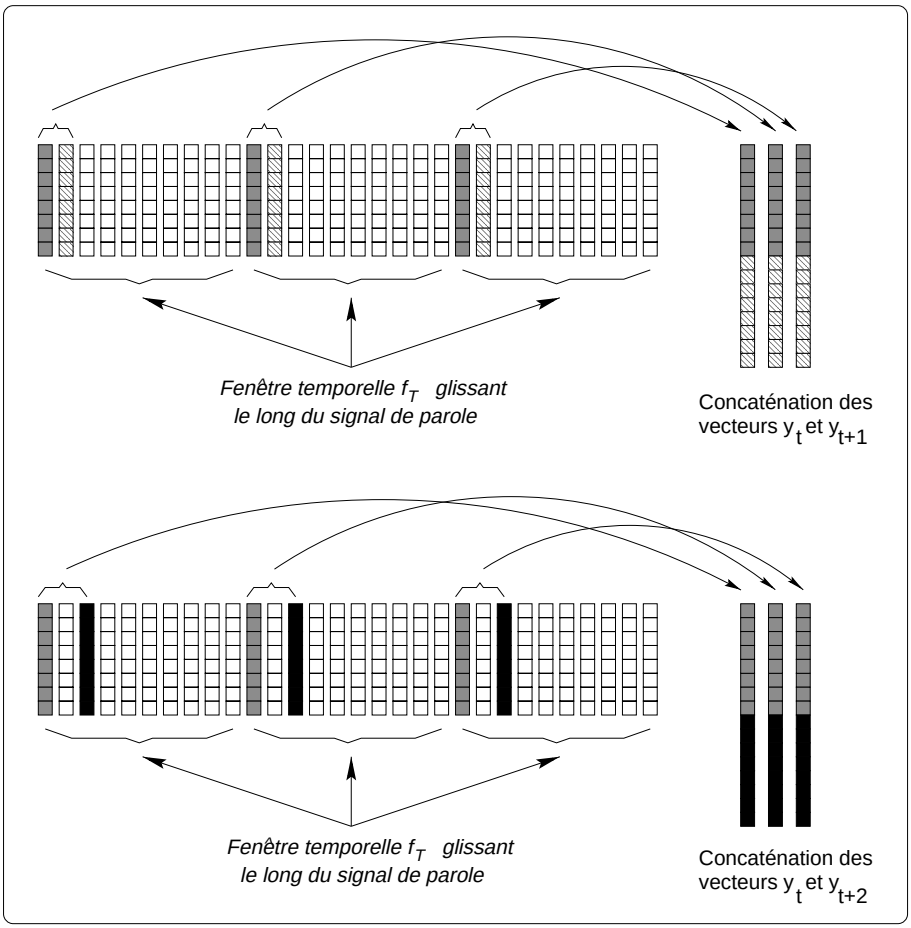

Figure 5.1: Choix de la fenêtre temporelle. Illustration de la concaténation de deux vecteurs utilisée pour le choix de la fenêtre temporelle. Ces vecteurs sont associés à des trames plus ou moins éloignées provenant d'une fenêtre temporelle glissant le long du signal de parole sans recouvrement possible entre deux déplacements. Les concaténations des trames t et $t+1$ ainsi que t et $t+2$ sont données en exemples.

déplacement de la fenêtre temporelle composée de 10 trames étant réalisé sans recouvrement possible, seuls 60 exemples de concaténations de trames sont disponibles par fichier pour l'apprentissage des modèles de locuteurs.

3 Concaténation et Modélisation

3.1 Formalisme des vecteurs étendus

Ce formalisme reprend les notations données dans le chapitre précédent (section 2).

Soit $\{y_t\}_{1 \leq t \leq N}$ une séquence de N vecteurs de coefficients représentant les trames d'un signal de parole s. Chaque vecteur y_t a pour dimension p.
Soit T la taille, en nombre de trames, de la fenêtre temporelle f_T choisie pour le traitement des informations dynamiques véhiculées par le signal s.
La phase de concaténation, illustrée par la figure 5.2, consiste à déplacer la fenêtre f_T le

Paire y_t et y_{t+k}	
Valeurs de k	Taux d'identification correcte en %
1	31,5
2	40
3	50
4	53,7
5	54,3
6	56,8
7	59,2
8	57,2
9	51,7

Tableau 5.1: Taille de la fenêtre temporelle. Taux d'identification correcte obtenus par concaténation de deux vecteurs y_t et y_{t+k} (suivant les valeurs de k) provenant d'une fenêtre temporelle glissant le long du signal de parole (Tâche d'identification sur TIMIT – 63 locuteurs – 1370 tests).

long du signal s et à concaténer la suite de vecteurs $\{y_k\}_{t\leq k<t+T}$ qui la composent.
Le résultat de la concaténation est une nouvelle séquence de vecteurs de coefficients, $\{y_{T\times i}\}_{1\leq i\leq M}$ (avec $M \leq N$), appelés vecteurs étendus. La dimension de chaque vecteur étendu est à présent $p.T$.

3.2 Modélisations statiques et statistiques

Après concaténation des trames successives, les vecteurs étendus résultant sont modélisés par une approche statique.
L'approche choisie dans ce chapitre est statistique. Elle repose sur une modélisation mono-gaussienne associée à une mesure de similarité, $\mu_{G\beta}$, issue des Méthodes Statistiques du Second Ordre (MSSO) [Gish et al., 1994], [Bimbot et al., 1995]. Utilisée en principe pour la modélisation des informations statiques véhiculées par le signal de parole, cette approche s'applique sur des vecteurs acoustiques instantanés. Il est néanmoins facile d'adapter ces méthodes aux vecteurs acoustiques étendus.

Selon les techniques MSSO, le modèle mono-gaussien est caractérisé par le triplet : $\{$*nombre de vecteurs*, *vecteur moyen*, *matrice de covariance*$\}$.

Soit x_t une séquence de N vecteurs instantanés de dimension p caractérisant un signal de parole prononcé par le locuteur $\mathcal{X}$, le triplet se définit par :

Nombre de vecteurs instantanés : N

Vecteur Moyen $\overline{x^{(t)}} = \frac{1}{N}\sum_{t=1}^{N} x_t$

Matrice de covariance $\mathcal{X}_{(t)} = \frac{1}{N}\sum_{t=1}^{N}(x_t - \overline{x^{(t)}}).(x_t - \overline{x^{(t)}})'$

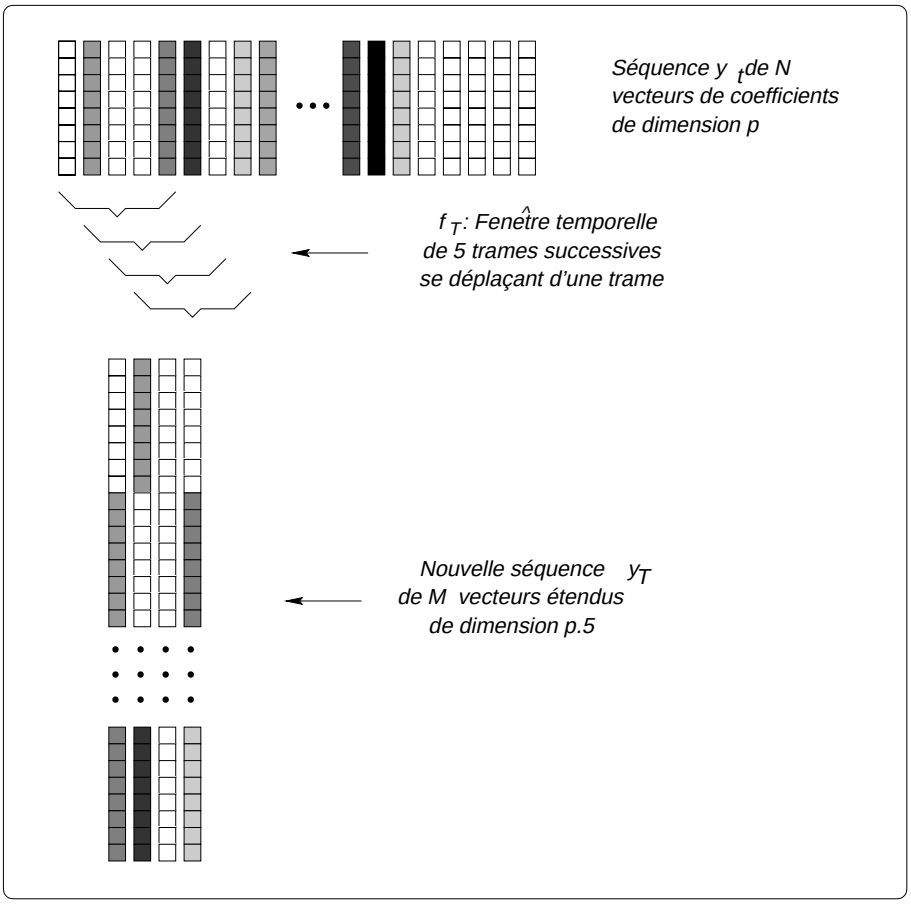

Figure 5.2: Exemple de concaténation de trames. Illustration de la concaténation de 5 trames successives provenant de la fenêtre temporelle f_T.

En considérant à présent la séquence $x_{T\times i}$ de M vecteurs étendus de dimension $p.T$, le triplet se définit simplement comme suit :

Nombre de vecteurs instantanés :	M
Vecteur Moyen	$\overline{x^{(T)}} = \frac{1}{M}\sum_{i=1}^{M} x_{T\times i}$
Matrice de covariance	$\mathcal{X}_{(T)} = \frac{1}{M}\sum_{i=1}^{M}(x_{T\times i} - \overline{x^{(T)}}).(x_{T\times i} - \overline{x^{(T)}})'$

La comparaison de ces deux triplets montre que la seule adaptation requise pour l'application d'un modèle mono-gaussien aux vecteurs étendus est la prise en compte d'un plus grand nombre de paramètres au sein des modèles, comme le montre la figure 5.3. La dimension du vecteur moyen est multipliée par le nombre de trames composant la fenêtre temporelle. Les dimensions de la matrice de covariance s'étendent de $p*p$ à $p.T*p.T$.

Durant la phase de test, cette constatation reste valable quelle que soit la mesure de similarité utilisée pour comparer le modèle du locuteur et le signal de test (le lecteur se reportera aux travaux de [Bimbot et al., 1995] pour une description détaillée des mesures de similarité applicables dans le cadre des MSSO).

L'adaptation de la méthode statistique retenue (MSSO) à un contexte dynamique est très simple et généralisable à d'autres méthodes. Dans ce travail, des modèles mono-gaussiens ainsi que des modèles à base de mélanges de gaussiennes sont employés.

4 Sélection du meilleur sous-ensemble de coefficients

4.1 Avant-propos

Concaténer les trames provenant d'une fenêtre temporelle permet de prendre en compte les informations dynamiques véhiculées par le signal de parole. À l'opposé des techniques d'extraction explicite des informations dynamiques, toutes les informations présentes dans la fenêtre temporelle sont considérées lors de la paramétrisation. Aucune compression semblable à la fonction g, introduite en section 5.2 du chapitre précédent, n'est réalisée.

Pourtant, toutes les informations issues des corrélations entre coefficients de la fenêtre temporelle ne sont pas pertinentes du point de vue de la caractérisation du locuteur. En particulier, la figure 5.3 montre une forte redondance des informations au sein de la matrice de covariance bloc-Toeplitz : les sous-matrices X_0, $\tilde{X}_0$, ..., sont quasiment identiques.

Par conséquent, il est nécessaire de sélectionner au sein de la fenêtre temporelle les informations les plus pertinentes pour la RAL.

Dans le modèle retenu, la sélection consiste, à partir de l'ensemble des composantes des vecteurs étendus, à déterminer le meilleur sous-ensemble de coefficients spécifiques du locuteur.

4.2 Procédure de sélection

En considérant un ensemble de n coefficients noté C, l'objectif d'une procédure de sélection est de retenir le meilleur sous-ensemble $P_{Best}(C)$ composé de p coefficients sélectionnés dans C ($p \leq n$). Trois grandes composantes interviennent dans une procédure de sélection :

1. l'algorithme de sélection $\mathcal{A}(C)$ permet de construire l'espace, noté $\mathcal{P}$, des sous-ensembles de coefficients $P_i(C)$ qui sont évalués tout au long de la procédure de sélection ;
2. la fonction d'évaluation $\mathcal{E}$ permet de mesurer la qualité de chaque sous-ensemble $P_i(C)$. Cette mesure de qualité est donnée par l'expression $\mathcal{E}(P_i(C))$;
3. le critère de sélection $\mathcal{S}$ détermine le meilleur sous-ensemble de coefficients $P_{Best}(C)$.

4.3 Algorithmes de sélection

L'algorithme de sélection exhaustif revient à évaluer la qualité de tous les sous-ensembles de coefficients possibles $P_i(C)$ – soit un total de 2^n avec n le nombre initial de coefficients – et à choisir le sous-ensemble $P_{Best}(C)$ optimal. Il est évident que cette tâche devient inconcevable, en termes de temps de calcul, lorsque le facteur n augmente.

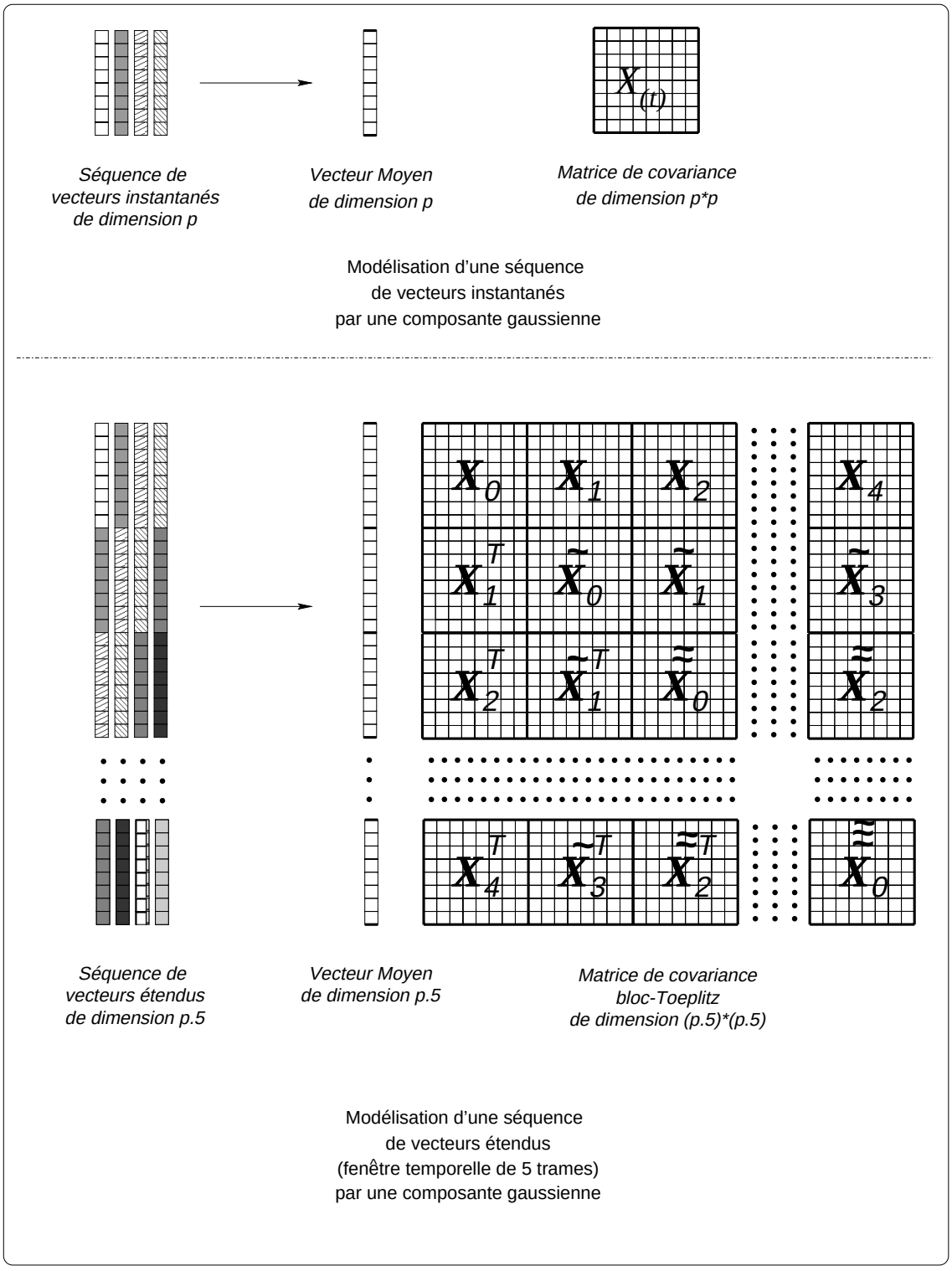

Figure 5.3: Approche statique appliquée aux informations dynamiques. Comparaison des modélisations par une mono-gaussienne – caractérisée par un vecteur moyen et une matrice de covariance – de vecteurs statiques (vecteurs instantanés) et vecteurs dynamiques (vecteurs étendus).

Les algorithmes de sélection proposés dans la littérature reposent sur le même principe d'évaluation de sous-ensembles de coefficients. Cependant, ils contournent les problèmes combinatoires sous-jacents en réduisant considérablement l'espace $\mathcal{P}$ des sous-ensembles évalués. Parmi ces techniques, [Doak, 1992] identifie trois catégories d'algorithmes (le lecteur se reportera à l'annexe A pour une description plus détaillée de ces différents algorithmes) :

- **Les algorithmes exponentiels** se caractérisent, comme leur nom l'indique, par une complexité exponentielle (de type $O(2^n)$). Les techniques exhaustives sont un exemple d'algorithmes exponentiels consistant à évaluer tous les sous-ensembles de coefficients $P_i(C)$ possibles. Ces techniques sont applicables sur un ensemble de coefficients de petite taille.
- **Les algorithmes séquentiels** se distinguent par leur complexité polynomiale (de type $O(n^2)$). Leur principe repose généralement sur l'ajout ou le retrait séquentiel de coefficients au sein d'un sous-ensemble dans le but de maximiser un critère particulier. Grâce à une complexité de calcul raisonnable, ce type d'algorithmes est fréquemment utilisé dans la littérature.
- **Les algorithmes aléatoires** reposent sur une construction et une évolution aléatoire de l'espace $\mathcal{P}$ jusqu'à satisfaction d'un critère d'arrêt. Les algorithmes génétiques ainsi que les algorithmes basés sur le recuit simulé sont des algorithmes aléatoires.

Algorithmes adaptés au contexte dynamique

Dans ce chapitre, l'algorithme de sélection intervient dans la recherche du meilleur sous-ensemble de coefficients spécifiques du locuteur, noté $P_{Best}(C_{Dyn})$. Selon la discussion sur les différents algorithmes de sélection proposée en annexe A, les algorithmes séquentiels, tels que la méthode knock-out ou ascendante, ne sont pas les mieux adaptés dans ce cadre. En effet, la condition d'optimalité de ce type d'algorithmes – l'indépendance entre les coefficients – n'est pas respectée par l'ensemble des coefficients C_{Dyn}. Néanmoins, l'utilisation des algorithmes génétiques, en tant que solution de remplacement, n'est pas compatible avec une taille variable de l'ensemble C_{Dyn}. Dans le cas d'un ensemble de grande taille, la complexité de calcul peut devenir difficile à gérer. Pour les mêmes raisons, les algorithmes basés sur le recuit simulé sont difficilement applicables. De plus, aucune garantie n'est donnée concernant l'optimalité de ce type d'algorithmes (convergence vers un optimum local).

Seuls les algorithmes séquentiels et en particulier la méthode ascendante tendent à un compromis entre complexité de calcul et degré d'optimalité de l'algorithme de sélection. L'approche ascendante, illustrée par la figure 5.4, est utilisée par conséquent dans la suite de ce chapitre.

4.4 Critère de sélection et fonction d'évaluation

Le rôle du critère de sélection $\mathcal{S}$ est de désigner, dans l'espace $\mathcal{P}$ défini par l'algorithme de sélection, le meilleur sous-ensemble de coefficients $P_{Best}(C)$.

Dans le cas d'un algorithme itératif, tel que l'approche ascendante, le critère de sélection intervient à chaque nouvelle itération comme indiqué par la figure 5.4.

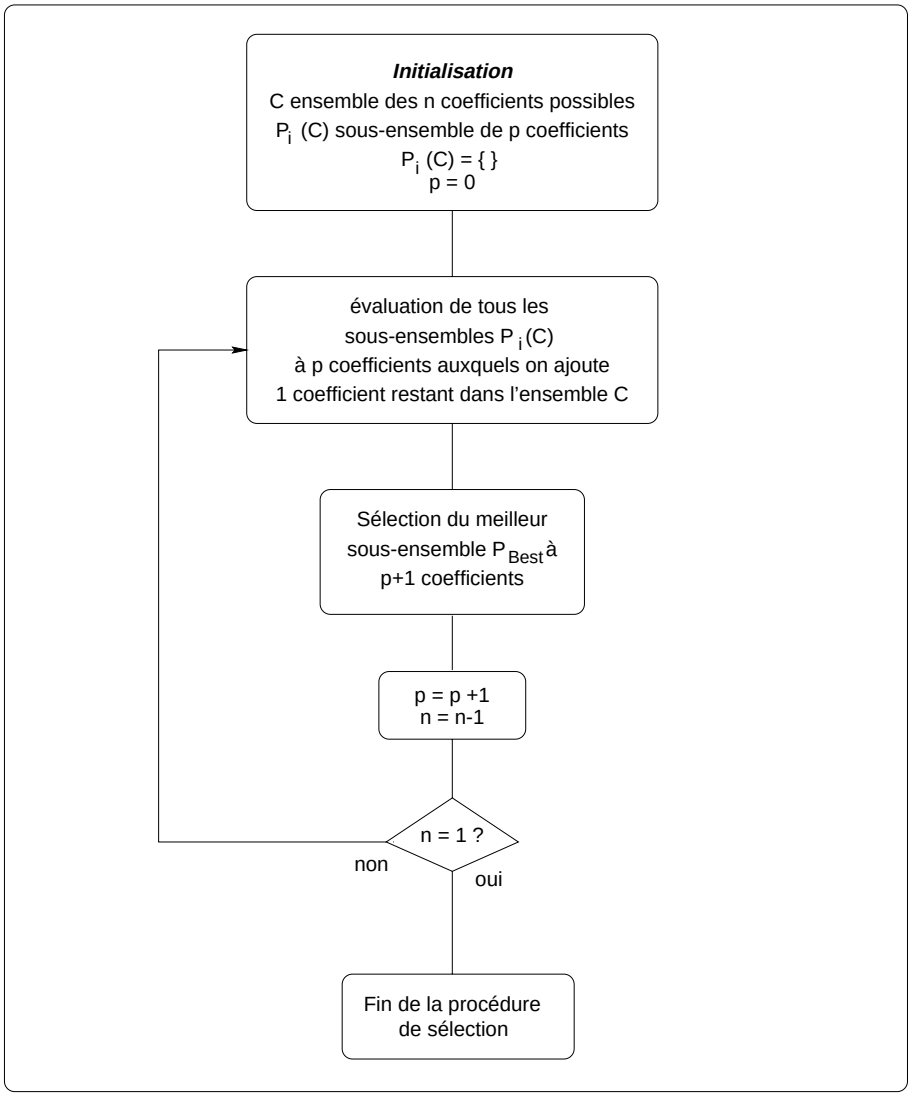

Figure 5.4: Technique de sélection ascendante. Descriptif algorithmique de la méthode de sélection ascendante.

Le critère de sélection conditionne le choix de la fonction d'évaluation $\mathcal{E}$. Cette dernière attribue à chaque sous-ensemble de coefficients de l'espace $\mathcal{P}$ une mesure de qualité en adéquation avec le critère de sélection choisi. La sélection du meilleur sous-ensemble se résume par la formulation suivante :

$$P_{Best}(C) = \mathcal{S}(P_i(C)) = arg\max_i \mathcal{E}(P_i(C)) \tag{5.1}$$

Critère du F-ratio

Ce critère statistique, basé sur une analyse de la variance, permet de mesurer la séparation entre n classes d'informations. Appliqué à la RAL, le F-ratio a pour objectif de mesurer la capacité discriminante d'un paramètre i.e. la capacité d'un paramètre à séparer les locuteurs [Pruzansky et al., 1964], [Wolf, 1972].

Etant donné un paramètre k et les observations de ce paramètre pour chaque locuteur, le critère du F-ratio se définit par le rapport suivant :

$$[F-ratio]_k = \frac{variance\ des\ moyennes\ des\ locuteurs}{moyenne\ des\ variances\ des\ locuteurs} \tag{5.2}$$

D'une manière plus formelle, ce rapport se réécrit :

$$\begin{aligned} [F-ratio]_k &= \frac{\frac{1}{q-1}\sum_{i=1}^{q}(S_k^i - U_k)^2}{\frac{1}{(s-1)q}\sum_{i=1}^{q}\sum_{j=1}^{s}(W_k^{ij} - S_k^i)^2} \\ S_k^i &= \frac{1}{s}\sum_{j=1}^{s} W_k^{ij} \\ U_k &= \frac{1}{q}\sum_{i=1}^{q} S_k^i \end{aligned} \tag{5.3}$$

où W_k^{ij} représente la valeur du paramètre k pour la séquence de parole j du locuteur i. S_k^i est une valeur moyenne du paramètre k du locuteur i estimée sur s séquences de parole. U_k est la valeur moyenne du paramètre k estimée sur les s séquences de parole prononcées par les q locuteurs considérés lors de l'analyse.

Le F-ratio est simple d'utilisation et très efficace pour la tâche de classification. Néanmoins, son application est conditionnée par la qualité de l'estimation des moyennes et des variances pour chaque locuteur. En effet, le F-ratio est bien adapté lorsque de grandes quantités de données sont disponibles pour chaque locuteur.

Critère de confusion : CritConf

Malgré cette limite majeure de mise en œuvre du F-ratio, l'idée de base sur laquelle ce critère repose – mesure de séparation entre n distributions – n'en reste pas moins très intéressante pour la RAL. Afin de pallier cette limite, un critère similaire, nommé CritConf, est proposé ici. Contrairement au F-ratio, ce critère se définit à partir de distributions globales calculées sur l'ensemble des locuteurs et nécessite par conséquent une quantité plus réduite de données par locuteur.

Pour un paramètre k, deux distributions sont estimées :

- la distribution intra-locuteur D_{intra} résume les valeurs du paramètre k relevées sur plusieurs observations d'un même locuteur et ce pour tous les locuteurs.

- la distribution inter-locuteur D_{inter} représente les valeurs du paramètre k relevées sur tous les locuteurs pour une observation donnée et ce pour toutes les observations.

Le critère CritConf mesure le taux de confusion entre les distributions D_{intra} et D_{inter}. Sa minimisation est recherchée durant la procédure de sélection, afin de déterminer le meilleur sous-ensemble $P_{Best}(C)$.

En pratique, étant données deux variables aléatoires x et y appartenant respectivement aux distributions D_{intra} et D_{inter}, le critère CritConf se traduit par la probabilité pour que x soit inférieure à y formalisée par $p(x < y)$ ou encore $p((x - y) < 0)$. Deux étapes sont nécessaires au calcul de cette probabilité :

1. on considère la distribution W issue de la différence des distributions D_{intra} et D_{inter} ; soit w une variable aléatoire appartenant à W, la probabilité $p((x - y) < 0)$ devient $p(w < 0)$.
2. on recherche la nouvelle probabilité $p(w < 0)$.

Pour obtenir une expression "symbolique" du critère, une hypothèse forte est émise : les distributions intra- et inter-locuteurs sont supposées gaussiennes. Cette hypothèse facilite le calcul de la probabilité par application des propriétés suivantes :

- la combinaison de deux distributions gaussiennes est une distribution gaussienne ;
- soit X et Y deux distributions gaussiennes, supposées indépendantes, caractérisées respectivement par leur moyenne μ_X et μ_Y et par leur écart à la moyenne σ_X et σ_Y, la distribution W de la combinaison de X et Y est caractérisée par sa moyenne $\mu_W = \mu_X - \mu_Y$ et son écart à la moyenne $\sigma_W = \sqrt{\sigma_X^2 + \sigma_Y^2}$.

Dans le cadre de la sélection du meilleur sous-ensemble de coefficients dynamiques, le paramètre k à étudier est la mesure de vraisemblance entre un signal de test et un modèle de locuteur. Les distributions D_{intra} et D_{inter} représentent respectivement les distributions des vraisemblances intra-locuteurs (le signal de test et le modèle appartiennent au même locuteur) et inter-locuteurs (le signal de test et le modèle n'appartiennent pas au même locuteur).

Taux d'identification : CritId

En RAL, maximiser le taux de reconnaissance (ou inversement minimiser le taux d'erreur) est un critère (orienté application) classiquement utilisé dans les procédures de sélection [Sambur, 1975], [Charlet, 1997]. De par sa nature, l'utilisation de ce critère semble la plus intuitive – *In practical terms, if a group of features, G, yields a smaller rate of error than another group of features, then the set G is necessarily a better set of features for recognizing speaker*[1] [Sambur, 1975]. Le second attrait de ce critère est évidemment sa facilité de mise en œuvre. En effet, la fonction d'évaluation se résume ici aux seuls traitements de reconnaissance du locuteur, i.e. à une série de tests de reconnaissance. Ici, le temps d'exécution d'une série de tests de reconnaissance devient évidemment un facteur dominant dans le coût d'une procédure de sélection sachant que pour la technique ascendante, par exemple, la fonction d'évaluation est répétée $\frac{n(n-1)}{2}$ avec n le nombre initial de coefficients.

[1] En pratique, si un ensemble de coefficients, G, produit un taux d'erreur plus faible qu'un autre ensemble de coefficients, alors l'ensemble G est nécessairement un meilleur ensemble de coefficients pour la reconnaissance du locuteur.

Critère d'émergence : CritEmer

Le principe du critère CritId est d'étudier la réponse du système de reconnaissance lors de l'utilisation d'un paramètre k. Le défaut majeur de ce critère est d'être dépendant du nombre de tests. et de ne pas tenir compte de la longueur des tests. Par ces deux points, le taux d'identification ne peut pas être considéré comme le résultat d'une fonction continue.

Une autre stratégie consiste donc à maximiser l'écart moyen entre la réponse du système de reconnaissance dans le cas d'un modèle de locuteur correct X_i et celle dans le cas de modèles compétiteurs X_j avec $j \neq i$.

Ces réponses sont données, ici, par les mesures de vraisemblance intra- ($L_s^i = L(s|X_i)$) et inter-locuteurs ($\overline{L_s^j} = L(s|X_j)$ avec $j \neq i$) respectivement pour le signal s. Le critère CritEmer vise à maximiser le taux d'émergence défini par l'équation :

$$Taux\ d'Emergence = \frac{1}{S}\sum_{s=1}^{S}\frac{L_s}{f(\overline{L_s})} \tag{5.4}$$

avec S le nombre de signaux de test utilisés et f une fonction agglomérant les mesures de vraisemblance issues des différents modèles compétiteurs. En considérant un nombre de locuteurs J, cette fonction f peut être de la forme :

$$f(\overline{L_s}) = \frac{1}{J}\sum_{j=1,j\neq i}^{J}\overline{L_s^j} \tag{5.5}$$

ou de la forme :

$$f(\overline{L_s}) = \max_{j,j\neq i}\overline{L_s^j} \tag{5.6}$$

5 Mise en œuvre de l'approche "dynamique"

La figure 5.5 résume les différentes étapes de l'approche "dynamique". Elle fait état d'une séparation nette entre sélection et exploitation de l'ensemble retenu $P_{Best}(C_{Dyn})$. Cette séparation implique l'usage de deux jeux de données : un jeu de *sélection* et un jeu d'*exploitation*.

L'approche "dynamique" se décompose en cinq étapes :

Etape 0 (étape préliminaire) : Paramétrisation.

L'étape de paramétrisation génère les vecteurs acoustiques instantanés de dimension p représentatifs du signal de parole. Elle est identique en apprentissage et en test.

Etape 1 : Concaténation.

La prise en compte des informations dynamiques au sein de vecteurs acoustiques successifs est réalisée en construisant un ensemble de vecteurs étendus durant les phases d'apprentissage et de test.

Etape 2 : Sélection.

Une procédure de sélection basée sur l'algorithme de sélection ascendante permet de rechercher le meilleur sous-ensemble de coefficients $P_{Best}(C_{Dyn})$ parmi l'ensemble C_{Dyn} composé des $p.T$ coefficients possibles relatifs à la fenêtre temporelle utilisée. Cette phase de sélection est appliquée uniquement sur le jeu de données de sélection.

Etape 3 : Exploitation de $P_{Best}(C_{Dyn})$.

Le résultat de la procédure de sélection, $P_{Best}(C_{Dyn})$, est appliqué sur le jeu de données d'exploitation résultant en une séquence de vecteurs étendus réduits, représentative des données d'exploitation.

Etape 4 : Modélisation.

Un modèle par locuteur est estimé à partir des vecteurs étendus réduits représentant les signaux d'apprentissage du jeu d'exploitation.

Etape 5 : Comparaison / Décision.

Une mesure de similarité est calculée entre les modèles de locuteurs et les vecteurs étendus réduits caractérisant les signaux de test du jeu d'exploitation.

6 Discussion sur l'approche "dynamique"

Par les mécanismes simples qu'elle met en jeu, l'approche "dynamique" proposée permet d'accroître significativement la longueur de fenêtres temporelles acceptées, comparée à la capacité des méthodes usuelles. Les 2 à 3 trames autorisées par les modèles ARV sont ici facilement dépassées. Néanmoins, cette longueur maximale est accrue mais non infinie. En particulier, l'application de modélisations statistiques liées à une faible quantité de données d'apprentissage limite la taille de la fenêtre puisqu'un nombre minimal d'échantillons est nécessaire pour une bonne estimation des paramètres des modèles. De plus, ces techniques sont également sujettes à des problèmes combinatoires dès lors qu'une forte augmentation est observée dans le nombre de paramètres des modèles. Ce nombre de paramètres est ici directement lié à la taille de la fenêtre.

La section 2 indique qu'une fenêtre temporelle de l'ordre d'une dizaine de trames de signal est nécessaire pour le traitement des informations dynamiques. Cette taille de fenêtre devient critique pour les méthodes statistiques du second ordre pour lesquelles la dimension des vecteurs moyens et des matrices de covariance est accrue d'un facteur équivalent à la taille de l'empan.

NB : Des vecteurs acoustiques de dimension 24 demandent par exemple l'estimation d'un vecteur moyen de dimension 240 et d'une matrice de covariance de dimension 240*240 pour chaque locuteur.

Une solution à ce problème est de considérer des espaces paramétriques plus réduits.

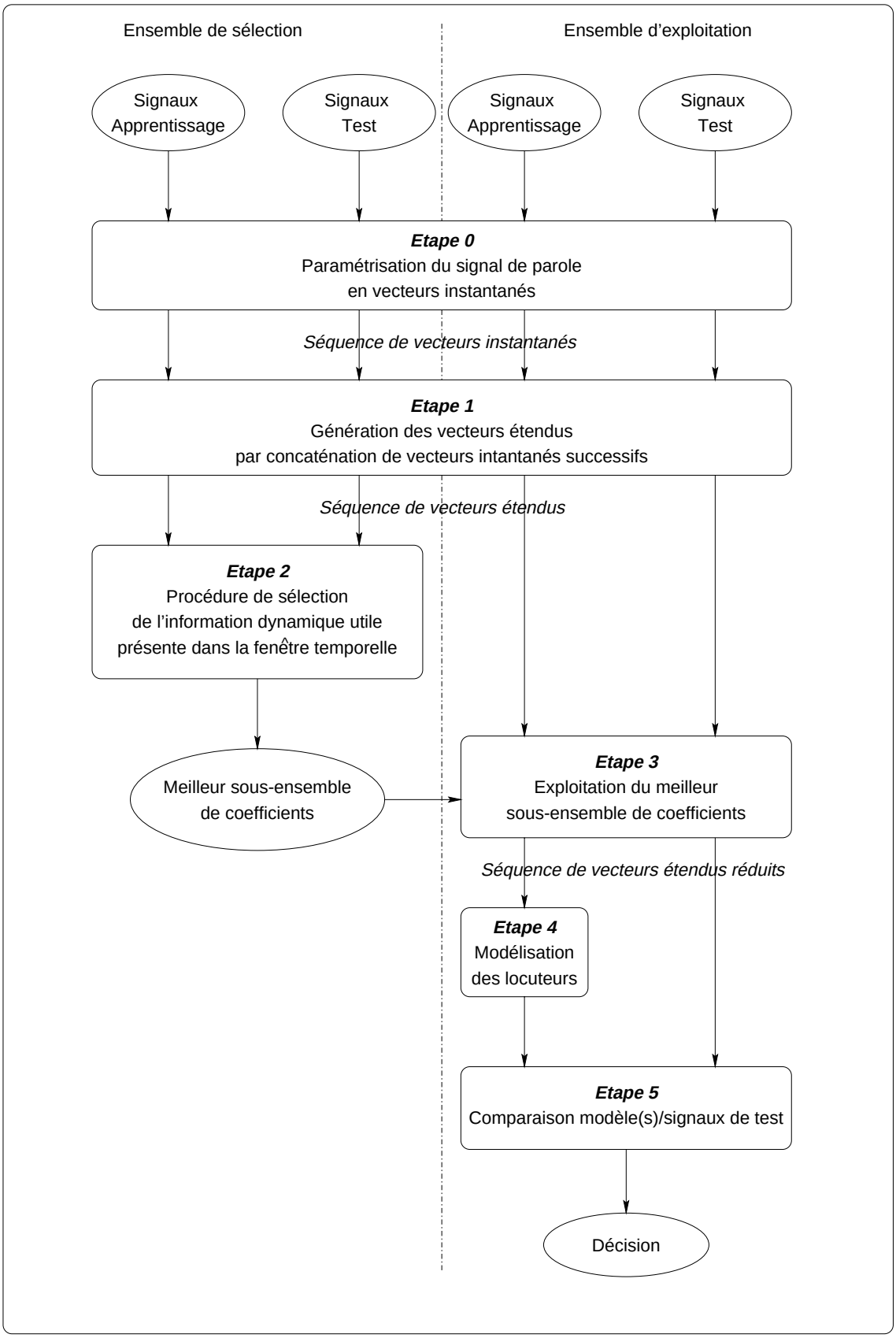

Figure 5.5: Vers une nouvelle approche "dynamique". Illustration des différentes étapes mises en œuvre par l'approche "dynamique".

6.1 L'approche multi-bande

L'approche multi-bande a été utilisée en RAP [Duchnowsky, 1993], [Bourlard et al., 1996], [Hermansky et al., 1996] et en RAL [Besacier et al., 1997], [Auckenthaler et al., 1997], [Besacier et al., 1998a], [Besacier et al., 1998b], [Besacier et al., 2000a]. Cette approche consiste à découper le domaine fréquentiel en sous-bandes traitées indépendamment les unes des autres, comme illustrée sur la figure 5.6. Un reconnaisseur est associé à une sous-bande particulière, fournissant un score de décision pour la tâche de reconnaissance visée. Les scores attribués à chaque sous-bande sont ensuite recombinés pour fournir un score final.

Dans ce cadre, les phases d'apprentissage et de reconnaissance sont réalisées sous-bande par sous-bande, dans un espace paramétrique plus réduit, comparé à l'utilisation d'une bande totale représentant le domaine fréquentiel dans sa globalité.

La motivation principale de l'approche multi-bande réside dans le fait que la qualité des informations fréquentielles peut être très dépendante de la bande de fréquences considérée. Par exemple, les travaux de [Besacier et al., 2000a] montrent que certaines sous-bandes semblent plus pertinentes que d'autres pour la RAL par exemple. Par ailleurs, certains bruits ne peuvent affectés qu'un nombre restreint de bandes de fréquences. Dans cette optique, un traitement particulier par sous-bande semble plus approprié qu'un traitement appliqué sur la bande fréquentielle totale car il permet de mieux tirer parti des avantages et défauts de chacune des sous-bandes.

6.2 Vers une approche dynamique multi-bande

En considérant les sous-bandes comme des sous-domaines fréquentiels particuliers, l'approche dynamique proposée dans ce chapitre demeure entièrement applicable. La seule différence réside dans la réduction de l'espace paramétrique initial.

Dans l'exemple que nous évoquions précédemment, nous considérions un domaine fréquentiel représenté par 24 coefficients spectraux. Dans ce contexte, la phase de modélisation de notre approche dynamique consistait en l'estimation d'une matrice de covariance de dimension 240 par 240 et d'un vecteur moyen de dimension 240. En considérant à présent une architecture multi-bande dynamique composée de 6 sous-bandes spectrales, la phase de modélisation consiste alors à estimer pour chacune des sous-bandes une matrice de covariance de dimension 40 par 40 et un vecteur moyen de dimenson 40. Dans ce cadre, l'utilisation d'une architecture multi-bande ne permet pas de réduire les temps de calcul mais la complexité des traitements.

Le chapitre suivant a pour objectif d'évaluer l'approche dynamique dans un tel contexte d'architectures multi-bandes.

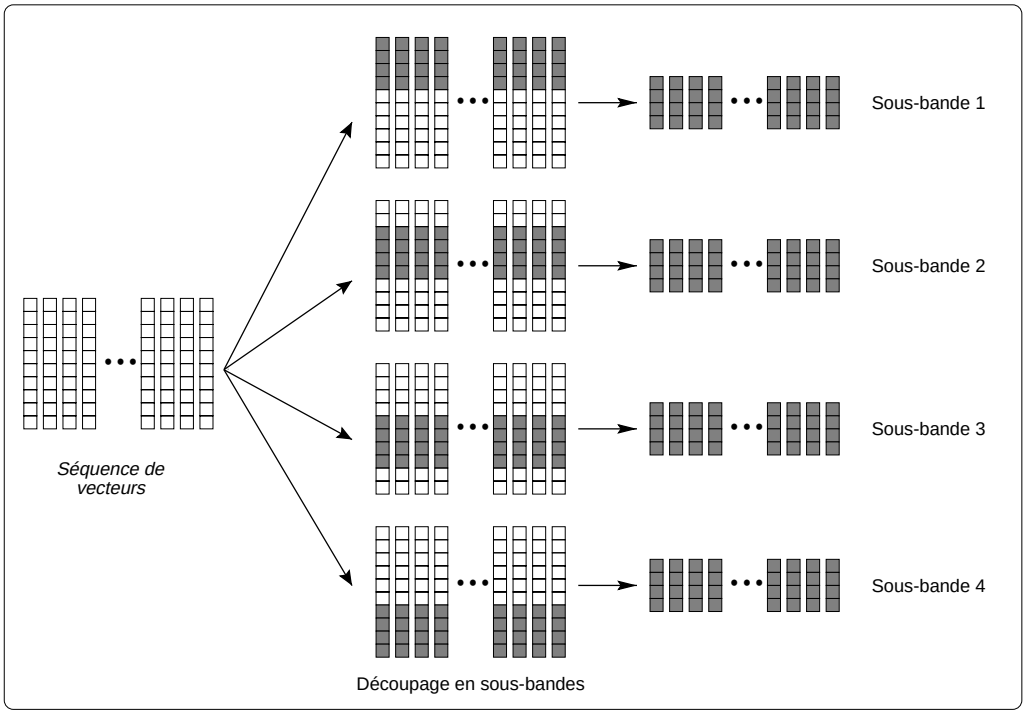

Figure 5.6: Approche multi-bande. Découpage du domaine fréquentiel en sous-bandes (avec recouvrement entre les sous-bandes), traitées individuellement par le système de RAL.

Chapitre 6

Évaluations de l'approche "dynamique"

L'approche "dynamique" est à présent évaluée dans un contexte expérimental, mettant en jeu deux bases de données différentes – TIMIT et Switchboard. Outre l'apport de la sélection, les expériences menées dans ce chapitre ont pour objectifs principaux de montrer la capacité de l'approche "dynamique" à prendre en compte l'information dynamique présente dans le signal de parole.

---

1 Introduction

Pour évaluer le potentiel de l'approche "dynamique", plusieurs séries d'expériences sont proposées dans ce chapitre. Le premier objectif, considéré comme essentiel, est de démontrer la prise en compte effective d'informations dynamiques par l'approche proposée. D'autres objectifs sont également visés comme :

- l'estimation de l'apport, en termes de performances, de la procédure de sélection ;
- la comparaison des différents critères de sélection proposés : CritId, CritConf, CritEmer d'une part en termes de performances et d'autre part en termes de corrélation entre sous-ensembles de coefficients sélectionnés.

La première série d'expériences est conduite sur une base de données de très bonne qualité – TIMIT – afin d'évaluer le comportement de l'approche "dynamique". Ce comportement est ensuite validé sur une base de données significativement plus difficile – Switchboard.

Tous les résultats présentés dans la suite de ce chapitre sont issus de tests d'identification du locuteur en ensemble fermé.

2 Comportement de l'approche "dynamique" sur TIMIT

2.1 Conditions expérimentales

2.1.1 Base de données : TIMIT

Les expériences présentées ici sont conduites sur la base de données TIMIT. Cette base de données comprend de la parole lue prononcée par 630 locuteurs américains (hommes et femmes), enregistrés en milieu calme.

Deux jeux de données sont constitués – jeu de sélection et jeu d'exploitation – pour les phases de sélection et d'exploitation du meilleur sous-ensemble de coefficients $P_{Best}(C_{Dyn})$. Les jeux de sélection et d'exploitation partagent la même population de locuteurs pour l'apprentissage, soit les 630 locuteurs de TIMIT. En revanche, les populations de test sont différentes ; le jeu de sélection comprend les 63 premiers hommes de TIMIT et le jeu d'exploitation les 567 locuteurs restant.

Les modèles de locuteurs sont appris sur 6 secondes de parole. Pour la phase de sélection, des signaux de parole de 6 secondes sont utilisés pour les tests d'identification. Le jeu de sélection dispose ainsi de 135 tests d'identification (en moyenne 2 à 3 tests d'identification par locuteur). Pour la phase d'exploitation, des durées de signaux plus courtes sont utilisées – 3 secondes – pour les tests d'identification, résultant en un total de 2639 tests (en moyenne 5 tests par locuteur).

2.1.2 Descriptif du système d'identification

Le système d'identification employé dans les expériences sur TIMIT repose sur l'utilisation conjointe d'une architecture multi-bande (pour réduire l'espace paramétrique) et de l'approche "dynamique" .

L'architecture multi-bande repose sur six sous-bandes – SB1, SB2, ..., SB6 – représentant les bandes fréquentielles suivantes : 0-430Hz, 430-1100Hz, 1100-2000Hz, 2000-3300Hz, 3300-5250Hz, 5250-8000Hz.
Chaque sous-bande est composée de 4 coefficients spectraux issus d'une analyse en banc de filtres. Lors de cette analyse, un banc de filtres constitué de 24 filtres triangulaires, dont les fréquences centrales sont réparties sur une échelle de Mel, est appliqué sur le spectre du signal de parole. Ce spectre est issu de l'application d'une transformée de Fourier de Winograd sur chaque trame du signal de parole (trames de 31,5 milli-secondes extraites toutes les 10 milli-secondes du signal de parole).

Une fenêtre temporelle de 100 milli-secondes – séquence de 10 trames successives – est adoptée pour l'exploitation des informations dynamiques. Cette fenêtre se déplace le long du signal de parole à hauteur d'une trame par déplacement.

Dans ce contexte dynamique, les sous-bandes SB1, SB2, ..., SB6 sont représentées par 40 coefficients (4 coefficients pris sur 10 trames). Appliquée sur chaque sous-bande, la procédure de sélection désigne parmi ces coefficients le meilleur sous-ensemble $P_{Best}(C_{Dyn})$.

L'application des techniques MSSO entraîne l'utilisation de modèles de locuteurs mono-gaussiens ainsi que l'application d'une mesure symétrique, μ_{G_β} proposée dans [Bimbot et al., 1995] (voir chapitre d'Introduction pour la formulation de cette mesure), pour estimer la distance entre un signal de test et un modèle de locuteur lors de la phase de test.

2.2 Informations statiques vs. informations dynamiques

Cette première expérience porte sur l'utilisation des informations dynamiques. L'approche "dynamique" est testée sur le jeu de données d'exploitation sans aucune procédure de sélection du meilleur sous-ensemble. L'utilisation des informations dynamiques est comparée à l'utilisation classique des informations statiques dans des conditions expérimentales identiques.

Le tableau 6.1 présente les résultats obtenus avec les informations statiques (table a) et les informations dynamiques (table b). Ces résultats sont donnés en termes de taux d'identification et taux de confusion pour chacune des sous-bandes de l'architecture multi-bande. Les taux d'identificatiin et de confusion sont introduits au chapitre 5 section 4.4.

La comparaison des tables a et b montre que l'apport des informations dynamiques, en termes de performances, est variable selon les sous-bandes. Comparée à l'utilisation des informations statiques, une dégradation des taux d'identification est observée sur les sous-bandes SB1, SB2, SB3 avec l'utilisation des informations dynamiques (application de l'approche "dynamique" sans sélection) alors que les sous-bandes SB4 et SB6 présentent un gain significatif en termes de taux d'identification. Les performances de la sous-bande SB5 restent, pour leur part, plutôt constantes.

Cette différence de comportement peut s'expliquer par la définition fréquentielle de chaque sous-bande. En effet, chacune d'elles représente un sous-ensemble du domaine

	Approche Statique	
SB	Taux (en %) d'Identification	Taux (en %) de Confusion
SB1	21,4	20,6
SB2	8,3	18,4
SB3	4,9	27,1
SB4	10,8	15,4
SB5	24,3	11,9
SB6	22,2	10,4

Table a. Utilisation des informations statiques du signal de parole.

	Approche "Dynamique" Sans Sélection	
SB	Taux (en %) d'Identification	Taux (en %) de Confusion
SB1	19	20,3
SB2	6,4	22,7
SB3	4,1	27,1
SB4	11,4	16,1
SB5	24,6	11,1
SB6	25	10,4

Table b. Utilisation des informations dynamiques par application de l'approche "dynamique" sans sélection.

Tableau 6.1: Informations statiques vs. dynamiques. Comparaison des performances du système d'identification multi-bande lors de l'utilisation des informations statiques ou de l'exploitation des informations dynamiques par l'approche "dynamique" sans sélection (Tâche d'identification sur TIMIT – jeu de données d'exploitation – 2639 tests – 3s de signal par test).

fréquentiel total. Nous pouvons supposer que certaines bandes de fréquences sont plus sensibles à la forte redondance d'informations ou à la quantité massive d'informations peu pertinentes inhérentes à l'approche "dynamique". Dans le cas des sous-bandes SB1, SB2 et SB3, une perte d'information utile est ressentie entraînant une dégradation des performances. À l'opposé, les sous-bandes SB4, SB5 et SB6 semblent bénéficier de la redondance d'informations pour améliorer leurs performances. Les expériences suivantes sur la sélection de l'information utile devraient confirmer ou infirmer cette hypothèse.

2.3 Approche "dynamique" et sélection de l'information utile

L'approche "dynamique" est à présent associée à la procédure de sélection de l'information utile. La procédure de sélection s'appuie sur le jeu de données de sélection issu de TIMIT pour déterminer le meilleur sous-ensemble de coefficients $P_{Best}(C_{Dyn})$ pour chacune des sous-bandes. Ces ensembles sont ensuite validés sur le jeu de données d'exploitation en utilisant à nouveau l'approche "dynamique".

Le tableau 6.2 fournit les résultats obtenus respectivement en utilisant les critères CritId (table a) et CritConf (table b) comme critère de sélection (les sous-ensembles de coefficients sélectionnés suivant les critères CritId et CritConf sont décrits en annexe B). Pour chaque critère, ces résultats sont donnés en termes de taux d'identification uniquement, afin de juger de l'apport de la procédure et du critère de sélection considéré en termes de performances du système d'identification.

La comparaison de ces résultats avec les résultats des informations statiques et des informations dynamiques sans sélection (voir le tableau récapitulatif 6.3) montre une nette amélioration des performances pour la majorité des sous-bandes et ce quel que soit le critère de sélection considéré. Suivant les sous-bandes, un gain relatif de 6 à 20% de taux d'identification est relevé pour le critère CritId contre 8 à 31% pour le critère CritConf. Ces taux d'identification sont comparés à ceux obtenus avec l'utilisation des informations statiques.

Ces premières observations montrent tout l'intérêt de la sélection de l'information dynamique utile au sein de la fenêtre temporelle.

Elles confirment, de surcroît, l'hypothèse émise dans la section précédente concernant la perte de performances observée sur les sous-bandes SB1, SB2 et SB3. Cette perte peut être effectivement attribuée à la forte redondance ou à la masse d'informations peu pertinentes, présentes au sein de la fenêtre temporelle, dès lors qu'aucune sélection de l'information utile n'est réalisée. En effet, avec une sélection de l'information utile, ces sous-bandes obtiennent des gains de performances significatifs.

Par ailleurs, il est intéressant de remarquer que le critère CritConf conduit à des taux d'identification bien meilleurs que le critère CritId sur les sous-bandes intermédiaires – SB2, SB3[1] et SB4 – et à des taux plus faibles sur les autres.
Les travaux de [Besacier, 1998] sur les systèmes multi-bandes ont montré que l'information utile à la caractérisation du locuteur est surtout présente dans les basses fréquences ($f \leq 500Hz$) et dans les hautes fréquences ($f \geq 2500Hz$). Les sous-bandes SB2, SB3 et SB4

[1] Pour la sous-bande SB3, les taux sont identiques pour les deux critères.

	Approche "Dynamique" Sélection suivant CritId
SB	Taux d'Identification (en %)
SB1	23,9
SB2	8,8
SB3	5,3
SB4	12,6
SB5	27,8
SB6	26,6

Table a. Sélection de l'information suivant le critère CritId.

	Approche "Dynamique" Sélection suivant CritConf
SB	Taux d'Identification (en %)
SB1	15,8
SB2	9,7
SB3	5,3
SB4	14,2
SB5	27
SB6	25

Table b. Sélection de l'information suivant le critère CritConf.

Tableau 6.2: Sélection de l'information utile. Application de l'approche "dynamique" et de la procédure de sélection à la tâche d'identification suivant différents critères de sélection : le critère d'identification CritId (a), le critère de confusion CritConf (b). Les résultats sont donnés en termes de taux d'identification (Tâche d'identification sur TIMIT – jeu de données d'exploitation – 2639 tests – 3s de signal par test).

	Taux (en %) d'identification			
SB	Statique	Dynamique	Dynamique avec sélection	
			CritId	CritConf
SB1	21,4	19	23,9	15,8
SB2	8,3	6,4	8,8	9,7
SB3	4,9	4,1	5,3	5,3
SB4	10,8	11,4	12,6	14,2
SB5	24,3	24,6	27,8	27
SB6	22,2	25	26,6	25

Tableau 6.3: Comparaison des résultats. Récapitulatif des résultats obtenus avec les informations statiques, les informations dynamiques sans sélection et avec sélection (critères CritId et CritConf) (Tâche d'identification sur TIMIT – jeu de données d'exploitation – 2639 tests – 3s de signal par test).

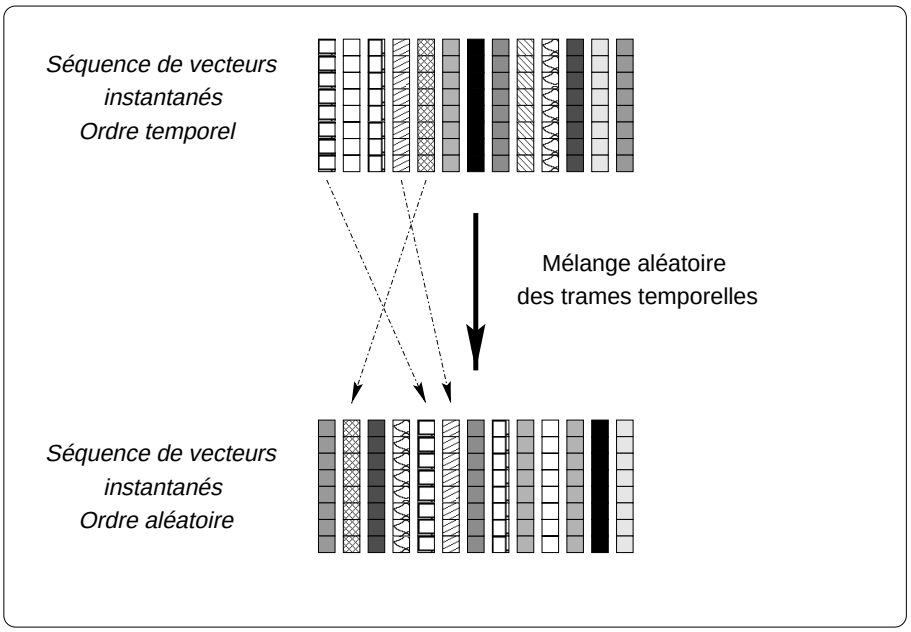

Figure 6.1: Mélange aléatoire des trames. Application sur une séquence de vecteurs intantanés du mélange aléatoire des trames temporelles d'un signal de parole.

couvrent une bande de fréquences de 430 à 3300Hz. Elles ne comportent par conséquent que peu d'informations spécifiques du locuteur. Cette déduction est d'ailleurs confirmée par leurs faibles performances comparées aux sous-bandes SB1, SB5 et SB6. Nous pouvons, par conséquent, supposer que le critère CritConf se révèle plus performant que CritId dans les cas où la quantité d'information utile à sélectionner est très faible au sein de la fenêtre temporelle.

2.4 Mélange aléatoire des trames temporelles

Un moyen simple de vérifier qu'une méthode dite "dynamique" prend effectivement en compte les informations dynamiques du signal de parole est d'appliquer un mélange aléatoire sur les trames temporelles du signal [Magrin Chagnolleau et al., 1996]. Ce mélange des trames, illustré par la figure 6.1, permet de "détruire" la structure temporelle du signal de parole et sa dynamique. L'association de ce mélange et d'une approche dite "dynamique" doit avoir pour conséquence une dégradation nette des performances puisque la succession temporelle des trames n'est plus respectée. En revanche, une approche statique soumise aux mêmes conditions expérimentales ne subit aucune perte de performances. En effet, le traitement statique est, par définition, appliqué aux trames indépendamment les unes des autres.

Les expériences précédentes sont répétées après application d'un mélange aléatoire des trames temporelles. Le tableau 6.4 présente les résultats obtenus avec l'utilisation des informations dynamiques par l'approche "dynamique" avec sélection. Pour comparaison, les résultats obtenus précédemment avec un ordre naturel des trames temporelles sont également reportés.

Comme attendu, l'application du mélange aléatoire n'a aucune incidence sur les

	Approche "Dynamique" avec Sélection – Mélange aléatoire des trames	
SB	Taux d'identification (en %) obtenus avec un ordre naturel des trames temporelles	Taux d'identification (en %) obtenus avec un ordre aléatoire des trames temporelles
SB1	23,9	21,5
SB2	8,8	7,8
SB3	5,3	4,7
SB4	12,6	11,2
SB5	27,8	24,4
SB6	26,6	22,3

Tableau 6.4: Destruction de la structure temporelle. Illustration, en termes de taux d'identification, de l'effet produit par un mélange aléatoire des trames temporelles du signal de parole lors de la prise en compte d'informations dynamiques par l'approche "dynamique" (Tâche d'identification sur TIMIT – 2639 tests – 3s de signal par test).

performances des informations statiques. En effet, les performances (non présentées ici) obtenues avec et sans mélange des trames sont similaires (différence inférieure à $10e-3\%$). À l'opposé, une dégradation sensible des performances de l'approche "dynamique" est observée dans le cas du mélange aléatoire.

Le comportement de l'approche "dynamique" face au mélange aléatoire des trames temporelles démontre la prise en compte effective d'informations dynamiques par cette technique.

2.5 Résumé des résultats sur TIMIT

Les expériences et résultats que nous venons de présenter sur la base de données TIMIT ont montré :

- un comportement opposé (gain ou dégradation de performances) selon les sous-bandes lors de l'application de l'approche "dynamique" sans sélection de l'information utile ;
- une amélioration significative des performances sur la majorité des sous-bandes lors de l'application de l'approche "dynamique" avec sélection de l'information utile par comparaison avec l'approche statique ;
- des performances meilleurs pour le critère CritConf dès lors que la quantité d'information utile à sélectionner est très faible au sein de la fenêtre temporelle ;
- la prise en compte réelle d'informations dynamiques, véhiculées par le signal de parole, par l'approche "dynamique".

Compte tenu des caractéristiques de la base de données TIMIT – parole lue enregistrée en milieu calme – , des expériences similaires doivent être menées sur une base de données plus "difficile", en l'occurrence la base de données Switchboard. Ces expériences ont pour objectif d'évaluer le comportement de l'approche "dynamique" dans un contexte conversationnel en milieu téléphonique.

3 Comportement de l'approche "dynamique" sur Switchboard

3.1 Conditions expérimentales

3.1.1 Base de données

Les expériences présentées ici sont conduites sur la base de données Switchboard. Deux jeux de données – jeux de sélection et d'exploitation – sont constitués afin de séparer les phases de sélection et d'exploitation du meilleur sous-ensemble de coefficients $P_{Best}(C_{Dyn})$.

Ces jeux de données proviennent de deux populations de locuteurs différentes. Chaque population est composée de 100 locuteurs dont 50 masculins et 50 féminins.

Les modèles de locuteurs sont appris sur des signaux de parole de 2 minutes, enregistrés lors de deux sessions d'enrôlement.

Les tests pour le jeu d'exploitation comme pour le jeu de sélection sont constitués de signaux de parole de 6 secondes. Le jeu de sélection dispose en moyenne de 3,7 tests par locuteur, soit un total de 180 tests d'identification pour les hommes et de 190 tests pour les femmes. Le jeu d'exploitation contient de 207 à 2258 tests d'identification pour les hommes et de 205 à 1668 tests pour les femmes.

3.1.2 Descriptif du système d'identification

Le système d'identification employé dans ces expériences repose toujours sur une architecture multi-bande et l'approche "dynamique". Néanmoins, compte tenu de la nature des données fournies par Switchboard, une nouvelle architecture est proposée.

L'architecture multi-bande repose sur trois sous-bandes – SB1, SB2, SB3. Les sous-bandes SB1, SB2 et SB3 sont composées chacune de 8 coefficients cepstraux dérivés d'une analyse en banc de filtres et représentent respectivement les bandes fréquentielles suivantes : 300-1600Hz, 1100-3100Hz, 2500-4000Hz.

Lors de l'analyse en banc de filtres, le signal de parole est caractérisé, toutes les 10 milli-secondes par 24 coefficients banc de filtres issus d'une FFT. Celle-ci est calculée sur une fenêtre de Hamming de 31,5 milli-secondes. Pour chaque bande fréquentielle, le sous-ensemble de coefficients banc de filtres considéré est ensuite transformé en 8 coefficients cepstraux. Finalement, une normalisation basée sur le retrait de la moyenne cepstrale (CMS[2]) est appliquée sur chacun des coefficients cepstraux afin d'atténuer les distorsions dues aux canaux de transmission (cf. chapitre 9 pour une description plus détaillée de la CMS).

Une fenêtre temporelle de 90 milli-secondes – séquence de 9 trames successives – est adoptée pour l'exploitation des informations dynamiques. Cette fenêtre se déplace le long du signal de parole à hauteur d'une trame par déplacement.

[2]Cepstral Mean Subtraction.

Dans ce contexte dynamique, les sous-bandes SB1, SB2 et SB3 sont représentées par 72 coefficients (8 coefficients pris sur 9 trames). Appliquée sur chaque sous-bande, la procédure de sélection désigne parmi ces coefficients le meilleur sous-ensemble $P_{Best}(C_{Dyn})$.

En vue d'une comparaison avec une technique dynamique "état de l'art", la bande fréquentielle totale (300-3400 Hz) est également utilisée. Cette sous-bande particulière est composée de 16 coefficients cepstraux issus de l'analyse en banc de filtres décrite précédemment. L'utilisation d'une fenêtre de 9 trames conduit à une bande totale "dynamique" représentée par 144 coefficients.

Il est à noter qu'une séparation des données homme/femme est réalisée lors de la procédure de sélection. En effet, nous supposons à présent que les sous-ensembles de coefficients sélectionnés par la procédure de sélection peuvent être différents selon le genre des locuteurs.

L'application des techniques MSSO suit les mêmes directives que les expériences précédentes : modèles de locuteurs mono-gaussiens et application d'une mesure symétrique pour estimer la distance entre un signal de test et un modèle de locuteur lors de la phase de test.

3.2 Informations statiques vs. informations dynamiques

L'utilisation des informations dynamiques, par application de l'approche "dynamique" sans sélection, est comparée à l'utilisation classique des informations statiques dans des conditions expérimentales identiques.

Le tableau 6.5 présente les résultats obtenus avec les informations statiques (table a) et dynamiques (table b). Ces résultats sont donnés en termes de taux d'identification, taux de confusion et taux d'émergence pour les sous-bandes SB1, SB2 et SB3 et par genre de locuteur.

La comparaison de ces deux tableaux montre que les informations dynamiques permettent une amélioration significative des résultats, quelle que soit la sous-bande considérée. En termes de taux d'identification, un gain relatif de 10 à 20 % est apporté par l'exploitation des informations dynamiques.

Le résultat de cette comparaison diffère des observations émises sur la base de données TIMIT (section 2.2) pour laquelle des dégradations de performances sont constatées pour certaines sous-bandes lors de l'utilisation des informations dynamiques sans sélection. Cette différence de comportement peut s'expliquer par la qualité des bases de données utilisées. Dans le cas de Switchboard, la redondance d'information amenée par l'approche "dynamique" permet de compenser la dégradation des signaux de parole, causée par le contexte téléphonique, ainsi que la perte d'information due au contexte conversationnel.

3.3 Mélange aléatoire des trames temporelles

Le mélange aléatoire des trames temporelles, introduit en section 2.4 pour la base de données TIMIT, est appliqué ici sur le jeu de données d'exploitation issu de Switchboard.

	Approche Statique					
	Taux (en %) d'Identification		Taux (en %) de Confusion		Taux d'Emergence	
SB	Hommes	Femmes	Hommes	Femmes	Hommes	Femmes
SB1	30,7	26,4	30,2	35,2	0,0280	0,0358
SB2	30,1	27,8	31,9	34,1	0,0211	0,0219
SB3	33,3	26,0	31,2	36,3	0,0226	0,0265

Table a. Utilisation des informations statiques du signal de parole.

	Approche "Dynamique" Sans Sélection					
	Taux (en %) d'Identification		Taux (en %) de Confusion		Taux d'Emergence	
SB	Hommes	Femmes	Hommes	Femmes	Hommes	Femmes
SB1	38,8	32,1	27,1	33,4	0,0105	0,0129
SB2	36,0	35,0	28,8	33,0	0,0740	0,0085
SB3	37,7	31,0	29,1	35,6	0,0092	0,0117

Table b. Utilisation des informations dynamiques par application de l'approche "dynamique" sans sélection.

Tableau 6.5: Informations statiques vs. dynamiques. Comparaison des performances du système d'identification multi-bande lors de l'utilisation des informations statiques ou de l'exploitation des informations dynamiques par l'approche "dynamique" sans sélection (Tâche d'identification sur Switchboard – jeu de données d'exploitation – 2258 tests pour les hommes, 1668 tests pour les femmes – 6s de signal par test).

	Approche Statique – Mélange aléatoire des trames			
	Taux d'identification (en %) obtenus avec un ordre naturel des trames temporelles		Taux d'identification (en %) obtenus avec un ordre aléatoire des trames temporelles	
SB	Homme	Femme	Homme	Femme
SB1	30,7	26,4	31,0	26,1
SB2	30,1	27,8	30,2	27,8
SB3	33,3	26,0	32,9	26,2

Table a. Mélange aléatoire des trames temporelles associé à une prise en compte des informations statiques.

	Approche "Dynamique" Sans Sélection – Mélange aléatoire des trames			
	Taux d'identification (en %) obtenus avec un ordre naturel des trames temporelles		Taux d'identification (en %) obtenus avec un ordre aléatoire des trames temporelles	
SB	Homme	Femme	Homme	Femme
SB1	38,8	32,1	2,3	3,8
SB2	36,0	35,0	1,1	3,1
SB3	37,7	31,0	2,8	2,0

Table b. Mélange aléatoire des trames temporelles associé à une prise en compte des informations dynamiques.

Tableau 6.6: Destruction de la structure temporelle. Illustration, en termes de taux d'identification, de l'effet produit par un mélange aléatoire des trames temporelles du signal de parole lors de la prise en compte soit d'informations statiques, soit d'informations dynamiques par l'approche "dynamique" (Tâche d'identification sur Switchboard – 2258 tests pour les hommes, 1668 tests pour les femmes – 6s de signal par test).

Le tableau 6.6 fournit les résultats obtenus avec l'utilisation des informations statiques (table a) et l'exploitation des informations dynamiques par l'approche "dynamique" (table b). Pour faciliter la comparaison, les résultats produits précédemment avec un ordre naturel des trames temporelles sont également reportés dans ce tableau.

Comme précédemment, les performances des informations statiques demeurent similaires avec un ordre naturel et aléatoire des trames du signal. En revanche, les performances de l'approche "dynamique" sont très fortement dégradées par le mélange aléatoire.

Le comportement de l'approche "dynamique" face à la destruction de la structure temporelle du signal confirme la prise en compte effective d'informations dynamiques par cette technique.

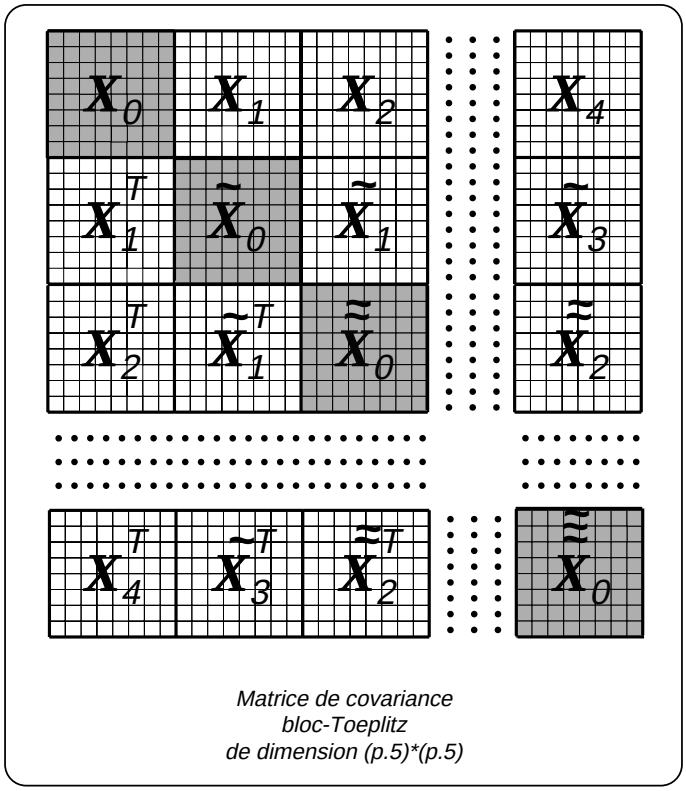

Figure 6.2: Matrice de covariance et mélange aléatoire des trames temporelles sur TIMIT. Visualisation des informations pertinentes au sein de la matrice bloc-toeplitz (parties grisées) après application du mélange temporelle des trames sur la bases de données TIMIT (Paramétrisation spectrale – matrice de covariance pleine).

Comparaison des résultats obtenus sur TIMIT et SwitchBoard après application du mélange aléatoire des trames temporelles

L'analyse des résultats du mélange aléatoire sur les bases de données TIMIT et Switchboard soulèvent deux questions principales :

1. Concernant TIMIT, pourquoi les résultats de l'approche dynamique après mélange aléatoire des trames sont comparables à ceux obtenus avec l'utilisation des informations statiques ?
2. Concernant Switchboard, pourquoi les résultats après mélange aléatoire des trames sont-ils si faibles ?

Le type de paramètres (coefficients spectraux pour TIMIT contre cepstraux pour Switchboard) ainsi que le type des matrices de covariance (matrice pleine pour TIMIT contre diagonale pour Switchboard), différents suivant la base de données utilisée, apportent une réponse à ces questions.

Considérons la matrice de covariance bloc-toeplitz utilisée dans le cadre de l'approche dynamique (se reporter à la figure 5.3). Après application du mélange aléatoire des trames temporelles, les seules informations pertinentes[3] présentes dans la matrice bloc-toeplitz

[3]Ces informations sont relatives à l'auto-covariance de chaque coefficient qui n'est en rien perturbée par le mélange des trames contrairement aux covariances entre coefficients.

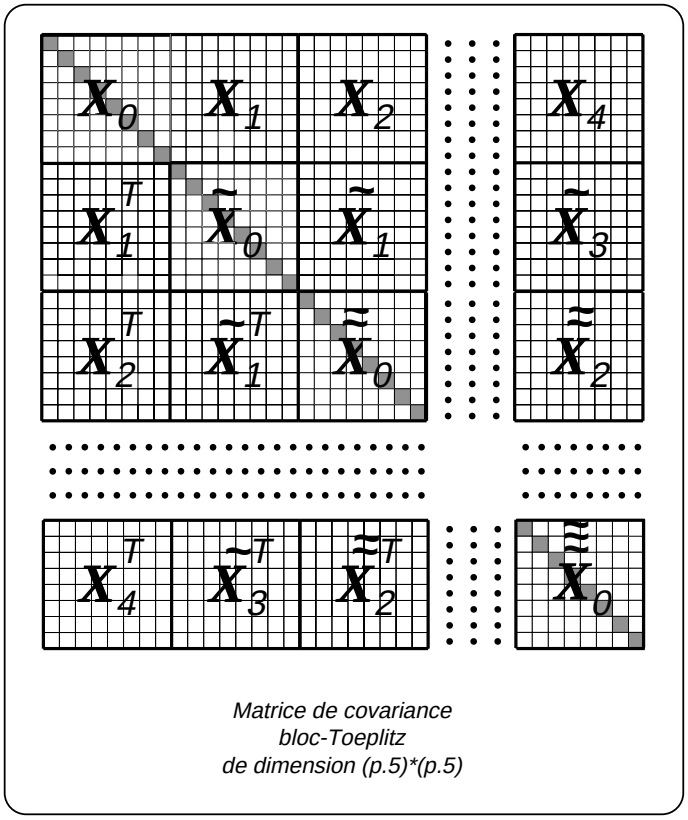

Figure 6.3: Matrice de covariance et mélange aléatoire des trames temporelles sur Switchboard. Visualisation des informations pertinentes au sein de la matrice bloc-toeplitz (parties grisées) après application du mélange temporelle des trames sur la bases de données Switchboard (Paramétrisation cepstrale – matrice de covariance diagonale) .

sont :

- concernant TIMIT, les matrices de covariance pleines localisées dans la diagonale de la matrice bloc-toeplitz (voir figure 6.2) ;
- concernant Switchboard, la diagonale de la matrice bloc-toeplitz (voir figure 6.3).

De cette répartition des informations pertinentes au sein de la matrice boc-toeplitz, nous pouvons en déduire que :

- les informations pertinentes portées par les matrices de covariance pleines localisées dans la diagonale de la matrice bloc-toeplitz dans le cas de TIMIT sont suffisamment robustes pour conserver un niveau de performances similaire à celui obtenu avec l'utilisation des informations statiques ;
- les seules informations pertinentes portées par la diagonale de la matrice bloc-toeplitz dans le cas de Switchboard sont en nombre insuffisant pour maintenir un niveau de performances raisonnable (similaire à l'utilisation des informations statiques). En effet, leur poids est moindre face à la quantité d'informations inconsistentes (en raison du mélange aléatoire) portées par la matrice bloc-toeplitz.

	Delta et Delta-Delta vs. Approche "dynamique" sans sélection			
	Taux d'identification (en %) Approche "Dynamique"		Taux d'identification (en %) Delta	
SB	Hommes	Femmes	Hommes	Femmes
BT	69,6	60,5	67,1	60

Tableau 6.7: Utilisation des coefficients Delta et Delta-Delta. Comparaison, en termes de taux d'identification, de l'utilisation des coefficients Delta et Delta-Delta pour l'exploitation des informations dynamiques véhiculées par le signal de parole et de l'approche "dynamique" (Tâche d'identification sur Switchboard – 207 tests pour les hommes, 205 tests pour les femmes – 6s de signal par test).

3.4 Approche "dynamique" vs. l'utilisation des coefficients Delta et Delta-Delta

Dans l'expérience présentée en section 3.2, l'approche "dynamique" s'est révélée plus performante face à une approche statique classique (utilisation des vecteurs instantanés uniquement).

Il est, à présent, intéressant de comparer l'approche "dynamique" à une technique "état de l'art", prenant en compte les informations dynamiques du signal. Dans cette optique, l'utilisation des coefficients Delta et Delta-Delta, bien connus pour leur efficacité en RAL et leur simplicité de mise en œuvre, est choisie.

Pour cette expérience, les dérivées première et seconde de chaque vecteur instantané sont calculées résultant respectivement en des vecteurs de coefficients Delta et Delta-Delta. Les vecteurs instantanés, Delta et Delta-Delta sont concaténés pour fournir des vecteurs dynamiques, utilisés en apprentissage et en test.

Le tableau 6.7 reporte les résultats, en termes de taux d'identification, obtenus d'une part en utilisant les coefficients Delta et Delta-Delta et d'autre part par application de l'approche "dynamique" sans sélection. Seuls les résultats de la bande totale sont présentés dans le tableau.

Ce tableau indique que l'approche "dynamique" parvient à des taux d'identification équivalents à ceux des coefficients Delta et Delta-Delta voire meilleurs dans le cas des hommes.

Cette comparaison montre que l'approche "dynamique", malgré sa simplicité de mise en œuvre face aux modèles ARV ou neuro-mimétiques, permet d'obtenir des résultats comparables aux méthodes "état de l'art" basées sur le traitement des informations dynamiques, représentées ici par l'utilisation des coefficients Delta et Delta-Delta.

3.5 Approche "dynamique" et sélection de l'information utile

La sélection de l'information dynamique utile est à présent testée en utilisant les deux jeux de données de Switchboard. Le meilleur sous-ensemble de coefficients $P_{Best}(C_{Dyn})$,

issu de la fenêtre temporelle, est déterminé sur le jeu de données de sélection puis utilisé sur le jeu d'exploitation.

Les résultats obtenus en utilisant les critères d'identification (table a), de confusion (table b) et d'émergence (table c) – détaillés en section 4.4 – comme critère de sélection sont présentés dans le tableau 6.8. Pour chaque critère, les résultats sont donnés en termes de taux d'identification, de taux de confusion et de taux d'émergence (exprimé pour ce dernier dans le domaine logarithmique par application des équations 5.4 et 5.6).

En première observation, il est à remarquer que ce tableau montre une variation relativement faible dans les résultats obtenus pour chacun des critères de sélection. Malgré la spécificité de mise en œuvre de chaque critère et la sélection de sous-ensembles $P_{Best}(C_{Dyn})$ différents (les sous-ensembles de coefficients sélectionnés par chacun des critères sur le jeu de développement sont reportés en annexe B), des performances équivalentes sont obtenues.

Ce tableau de résultats est comparé au tableau 6.5 afin d'étudier les effets de la sélection de l'information utile. Quel que soit le critère de sélection considéré, les performances résultant de l'exploitation des informations dynamiques sans sélection sont plus ou moins dégradées après sélection de l'information utile. Néanmoins, le gain significatif de performances observé entre l'utilisation des informations statiques et dynamiques est toujours présent après sélection.

La procédure de sélection de l'information utile semble inopérante dans ce contexte. Le fait de déterminer le meilleur sous-ensemble de coefficients sur le jeu de données de sélection et d'exploiter ce dernier sur le jeu de données d'exploitation semble avoir pour conséquence de pénaliser les performances de ce dernier jeu sur Switchboard.

Cette constatation est de nouveau en contradiction avec l'application de cette approche sur la base de données TIMIT. Dans ce dernier cas, une amélioration significative des performances est obtenue après sélection de l'information utile.

La différence de qualité entre les bases de données TIMIT et Switchboard peut encore être à l'origine de cette altération de comportement de l'approche "dynamique" avec sélection. En effet, la dégradation du signal de parole due au contexte téléphonique de Switchboard peut avoir pour conséquence de réduire l'efficacité du meilleur sous-ensemble de coefficients, défini sur le jeu de données de sélection, dès lors qu'il est utilisé sur le jeu de données d'exploitation. D'autre part, la sélection de l'information utile a pour effet de réduire la redondance des informations. Or, nous avons pu constater avec l'utilisation de l'approche "dynamique" sans sélection sur Switchboard que cette redondance semble nécessaire pour compenser les pertes d'informations dues au contexte conversationnel en milieu téléphonique.

Enfin la nature même de la base de données Switchboard (multi-session, conversation réelle) joue certainement un rôle non négligeable dans le manque de pertinence de la sélection. En effet, des différences significatives entre les signaux des jeux de sélection et d'exploitation peuvent être à l'origine de ce dysfonctionnement.

	Approche "Dynamique" – Sélection suivant CritId					
	Taux (en %) d'Identification		Taux (en %) de Confusion		Taux d'Emergence	
SB	Hommes	Femmes	Hommes	Femmes	Hommes	Femmes
SB1	37,2	31,4	28,1	34,5	0,0124	0,0140
SB2	34,5	33,1	29,8	34,5	0,0090	0,0091
SB3	36,5	27,5	29,8	36,7	0,0116	0,0157

Table a. Sélection de l'information suivant le critère CritId.

	Approche "Dynamique" – Sélection suivant CritConf					
	Taux (en %) d'Identification		Taux (en %) de Confusion		Taux d'Emergence	
SB	Hommes	Femmes	Hommes	Femmes	Hommes	Femmes
SB1	37,4	30,4	27,4	33,4	0,0127	0,0145
SB2	34,5	32,0	30,5	33,4	0,0099	0,0111
SB3	33,7	28,4	30,9	35,9	0,0119	0,0159

Table b. Sélection de l'information suivant le critère CritConf.

	Approche "Dynamique" – Sélection suivant CritEmer					
	Taux (en %) d'Identification		Taux (en %) de Confusion		Taux d'Emergence	
SB	Hommes	Femmes	Hommes	Femmes	Hommes	Femmes
SB1	36,8	32,6	27,4	31,2	0,0109	0,0118
SB2	35,3	31,7	29,5	34,1	0,0083	0,0096
SB3	35,6	29,9	29,8	37,1	0,0090	0,0115

Table c. Sélection de l'information suivant le critère CritEmer.

Tableau 6.8: Sélection de l'information utile. Application de l'approche "dynamique" et de la procédure de sélection à la tâche d'identification suivant différents critères de sélection : le critère d'identification CritId (a), le critère de confusion CritConf (b) ou le critère d'émergence CritEmer (c). Les résultats sont donnés en termes de taux d'identification, taux de confusion et taux d'émergence (Tâche d'identification – jeu de données de validation – 2258 tests pour les hommes, 1668 tests pour les femmes – 6s de signal par test).

Chapitre 7

Conclusion sur les informations dynamiques

Cette première partie a été dédiée au traitement des informations dynamiques caractéristiques du locuteur. Nous avons souligné, dans un premier temps, l'intérêt de ce type d'informations dans le cadre d'un système de RAL. Nous avons ensuite présenté un état de l'art détaillé des différentes techniques, appliquées dans la littérature pour le traitement des informations dynamiques. L'étude de ces techniques a mis en évidence un certain nombre de limites à considérer, notamment lors de l'utilisation d'une large fenêtre temporelle.

Une première étude expérimentale ayant confirmé la nécessité d'une large fenêtre temporelle (100 milli-secondes de signal de parole) pour le traitement des informations dynamiques, nous avons proposé une nouvelle approche mieux adaptée à ce contexte. Le principe de base de cette approche est de sélectionner au sein de la fenêtre temporelle l'information dynamique utile, spécifique du locuteur. À cette fin, nous avons utilisé un algorithme de sélection séquentiel classique, l'approche ascendante, et proposé différents critères de sélection : CritId, CritConf et CritEmer.

La validation expérimentale de l'approche "dynamique", sur les bases de données TIMIT et Switchboard, a mis en évidence différents points.

Informations dynamiques

Les expériences basées sur le mélange aléatoire des trames du signal de parole ont démontré la prise en compte effective d'informations dynamiques par l'approche "dynamique" proposée.

Informations statiques vs. dynamiques

Les expériences comparatives sur l'utilisation des informations statiques et l'exploitation des informations dynamiques par l'approche "dynamique" sans sélection ont confirmé l'intérêt d'intégrer des informations dynamiques au sein des systèmes de RAL, notamment dans un contexte conversationnel en milieu téléphonique (base de données Switchboard). Dans ce contexte, l'utilisation des informations dynamiques, ainsi que la forte redondance d'informations au sein de la fenêtre temporelle a permis un gain significatif des performances.

En revanche, il est difficile de se prononcer sur l'intérêt de l'approche "dynamique" sans sélection dans un contexte de parole lue en milieu calme (base de données TIMIT). Dans ce contexte, l'utilisation des informations dynamiques peut induire un gain de performances pour certaines sous-bandes ou à l'inverse une dégradation pour d'autres. Dans ce dernier cas, la redondance d'informations ou la présence d'informations peu pertinentes semblent à l'origine des dégradations.

Sélection de l'information dynamique utile

La contribution de la sélection de l'information utile au sein de la fenêtre temporelle s'est avérée variable selon les bases de données.
Les expériences menées sur TIMIT ont montré tout l'intérêt de cette procédure. Les sous-bandes dégradées lors de l'application de l'approche "dynamique" sans sélection sont parvenues à une amélioration notable de leurs performances après sélection.

À l'opposé, les expériences conduites sur Switchboard se sont révélées peu concluantes quant à l'utilité de cette procédure de sélection et ce quel que soit le critère de sélection considéré.

Trois facteurs peuvent expliquer ce comportement :

- Nous avons vu précédemment que la forte redondance d'informations, inhérente à l'approche "dynamique" sans sélection, était bénéfique aux données de Switchboard pour pallier la mauvaise qualité des signaux de parole. Sélectionner l'information utile revient à réduire considérablement la redondance nécessaire en milieu bruité.
- La propriété multi-session des signaux de parole de la base Switchboard peut rendre inopérante la sélection de l'information utile sur un jeu de données séparé.
- L'optimalité de l'algorithme de sélection peut aussi être remise en cause. Nous avions précisé, lors du choix de l'approche ascendante (chapitre 5 section 4.3), que la condition d'optimalité de ce type d'algorithmes – l'indépendance entre les coefficients – n'est pas respectée par l'ensemble des coefficients C_{Dyn} que nous considérons. Cette corrélation entre coefficients est peut-être davantage accentuée sur la base de données Switchboard dégradant d'autant les performances de l'algorithme de sélection.

État de l'art vs. approche "dynamique"

L'approche "dynamique" proposée a été comparée à l'utilisation conjointe de paramètres instantanés classiques et de leurs dérivées première et seconde au sein du système d'identification du locuteur. Dans des conditions expérimentales identiques, l'approche "dynamique", malgré sa simplicité de mise en œuvre, parvient à des résultats comparables aux techniques "état de l'art", représentées ici par l'utilisation des coefficients Delta et Delta-Delta.

Deuxième partie

World+MAP, une nouvelle technique de normalisation

Cette deuxième partie s'intéresse au processus de décision impliqué en Vérification Automatique du Locuteur et, plus particulièrement, au couple estimation des vraisemblances et seuil de décision. Ces deux composantes du processus de décision sont tout d'abord présentées au chapitre 8. La faible qualité d'estimation des vraisemblances ainsi que le choix du seuil de décision sont, ensuite, longuement discutés : le choix d'un seuil global fixé a priori, préconisé dans la littérature, est inconcevable en raison de la grande variabilité des vraisemblances, manifeste en présence de données disparates entre apprentissage et test. Cette discussion met en évidence la nécessité d'une normalisation des vraisemblances.

La littérature propose trois grandes classes de normalisation pour pallier cette problématique : normalisation de l'espace des paramètres acoustiques, normalisation de l'espace des mesures de similarité et normalisation de l'espace des seuils. Chacune de ces normalisations et les techniques associées sont présentées au chapitre 9.

Finalement, le chapitre 10 est consacré à la proposition d'une approche originale de normalisation, appelée World+MAP. Basée sur l'utilisation d'un modèle du monde dans un cadre bayesien, cette approche permet de réduire la variabilité des vraisemblances et de "projeter" ces dernières dans un espace probabiliste. L'approche World+MAP est validée au cours de travaux expérimentaux menés sur un sous-ensemble de la base de données Switchboard.

Chapitre 8

Processus de décision en VAL

1 Généralités

La Vérification Automatique du Locuteur (VAL) est le processus décisionnel permettant d'accepter ou de rejeter l'identité d'une personne par l'analyse de sa voix.

Une personne présentée au système de VAL décline son identité et atteste cette dernière en prononçant un message. Ce message vocal est utilisé par le système lors du processus de vérification de l'identité. Lorsque le message vocal et l'identité revendiquée correspondent à une même personne, l'accès est dit "client". Dans le cas inverse, l'accès est dit "imposteur".

Dans le cas de transactions bancaires par téléphone, sécurisées par un système de VAL, les accès imposteurs doivent être détectés et rejetés par le système afin de protéger les intérêts financiers de chaque client. Inversement, les accès clients doivent être acceptés pour satisfaire et fidéliser la clientèle. Toute la problématique d'un système de VAL repose sur la manière d'accepter les accès clients et de refuser les accès imposteurs.

1.1 Présentation schématique d'un système de VAL

Comme évoqué précédemment et décrit brièvement en chapitre d'Introduction (chapitre 1), l'identité revendiquée par une personne (client ou imposteur) ainsi que le message vocal attestant cette identité constituent les deux entrées du système de VAL.

Lors d'un accès, une identité correspond obligatoirement à un client connu du système. Chaque client est représenté au sein du système par une référence caractéristique, appelé modèle client. Ces modèles clients sont estimés sur des signaux de parole d'apprentissage collectés lors de sessions d'enrôlement.

Lors du processus de vérification, le signal de parole prononcé par un locuteur est comparé au modèle client correspondant à l'identité revendiquée. Ce processus s'appuie, en premier lieu, sur l'estimation d'une mesure de similarité entre le signal de parole et le modèle client. Cette mesure de similarité est, ensuite, comparée à une valeur de seuil pour décider d'accepter ou de rejeter l'identité proclamée. Cette présentation schématique du processus de vérification est illustrée par la figure 8.1.

1.2 Formalisme

Soit $\mathcal{X}$ le modèle client correspondant à l'identité fournie par un locuteur Y. Soit $\{y_t\}_{1 \leq t \leq N}$ une séquence de vecteurs acoustiques caractérisant le signal de parole émis par le dit locuteur. Cette séquence sera notée plus simplement y_N dans la suite de cette thèse. La règle de décision du système de VAL se définit par :

$$\begin{aligned} p(\mathcal{X}|y_N) &\geq \Theta \quad \Rightarrow acceptation \\ p(\mathcal{X}|y_N) &< \Theta \quad \Rightarrow rejet \end{aligned} \tag{8.1}$$

où $p(\mathcal{X}|y_N)$ est la probabilité du modèle client $\mathcal{X}$ sachant le signal de parole émis y_N et Θ représente le seuil de décision.

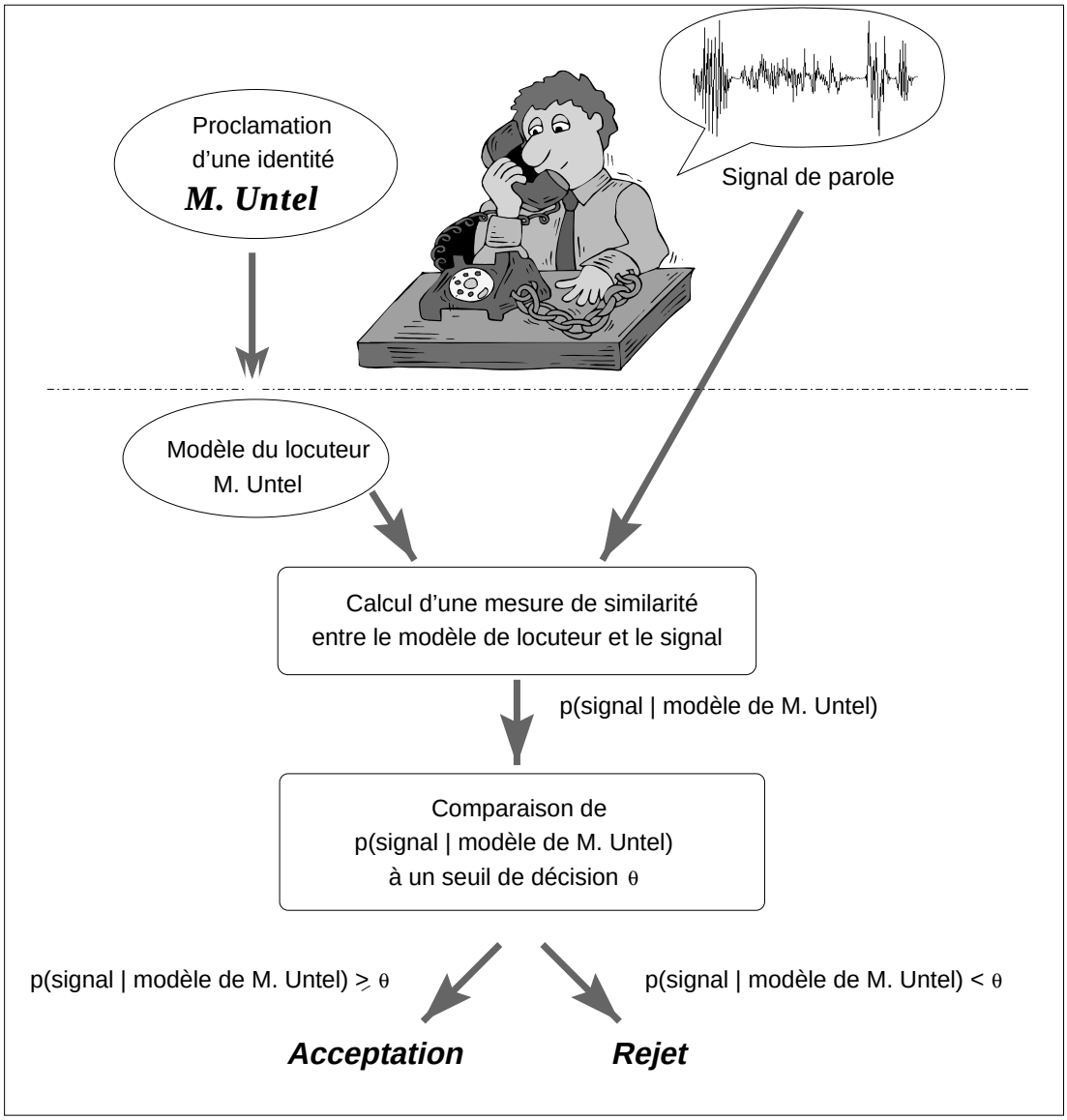

Figure 8.1: Processus de décision. Présentation schématique d'un système de VAL en environnement téléphonique.

En appliquant la règle de Bayes, la probabilité $p(\mathcal{X}|y_N)$ se réécrit sous la forme :

$$p(\mathcal{X}|y_N) = \frac{p(y_N|\mathcal{X}).P(\mathcal{X})}{P(y_N)} \tag{8.2}$$

avec $p(y_N|\mathcal{X})$ la probabilité que le signal de parole (représenté par la séquence de vecteurs y_N) soit issu du modèle client $\mathcal{X}$, $P(\mathcal{X})$ la probabilité *a priori* pour qu'un signal de parole quelconque soit issu du modèle $\mathcal{X}$[1] et $P(y_N)$ la probabilité *a priori* du signal de parole.

La règle de décision devient alors :

$$\begin{aligned} \frac{p(y_N|\mathcal{X}).P(\mathcal{X})}{P(y_N)} &\geq \Theta \quad \Rightarrow acceptation \\ \frac{p(y_N|\mathcal{X}).P(\mathcal{X})}{P(y_N)} &< \Theta \quad \Rightarrow rejet \end{aligned} \tag{8.3}$$

2 Estimation des vraisemblances

Dans la littérature, les méthodes statistiques, comme par exemple les mixtures de gaussiennes (GMM [Reynolds, 1995]), sont classiquement employées pour la construction des modèles clients. Dans ce contexte statistique, la probabilité pour qu'un signal de parole soit émis par un modèle client donné est approchée généralement par l'estimation du maximum de vraisemblance (EMV). Une approximation de la probabilité $p(y_N|\mathcal{X})$ est alors donnée par la vraisemblance $L_{\mathcal{X}}(y_N)$ pour que le signal y_N soit produit par le modèle statistique $\mathcal{X}$. Pour des raisons calculatoires, cette vraisemblance est très souvent donnée sous sa forme logarithmique, notée $LL_{\mathcal{X}}(y_N)$.

Au regard de la règle de décision énoncée précédemment (équation 8.3), l'estimation de la vraisemblance $L_{\mathcal{X}}(y_N)$ est une étape importante pour le processus de décision, en vue de la comparaison avec le seuil de décision. La qualité de cette estimation est intrinsèquement liée à la qualité de l'estimateur – ici l'approche EMV.

2.1 Estimation du Maximum de Vraisemblance : EMV

L'approche EMV peut être utilisée à deux niveaux au sein d'un système de VAL basé sur des méthodes statistiques. Couplée à l'algorithme EM[2], elle est très souvent utilisée pour estimer les paramètres des modèles clients. Lors de la phase de vérification, elle est appliquée pour l'estimation des vraisemblances entre un signal de parole et un modèle.

L'approche EMV est connue pour ses propriétés asymptotiques. En présence d'une très grande quantité de données d'apprentissage, cette approche parvient à une estimation très fiable des paramètres d'un modèle. De même, s'appuyant sur un modèle très bien appris i.e. ayant une très large couverture statistique, l'approche EMV fournit de très bonnes estimations de la vraisemblance entre de nouvelles observations et le dit modèle. Dans ce cas de figure, des valeurs de vraisemblances tendant vers l'infini (infini également dans le domaine logarithmique) correspondent à des observations émises par le modèle alors que des valeurs tendant vers 0 (l'infini négatif dans le domaine logarithmique) sont

[2]Expectation-Maximization.

relatives à des observations provenant d'un autre modèle.

En VAL, la qualité de l'estimation des vraisemblances $L_{\mathcal{X}}(y_N)$ par EMV est, par conséquent, très liée à la qualité des modèles clients.

2.2 Qualité des modèles

En VAL, la perfection d'un modèle client signifie une couverture complète de la variabilité de la voix du client ainsi que la prise en compte de toutes les conditions d'utilisation du système. Or, il est bien connu que la voix d'un locuteur évolue avec le temps, comme souligné à la section 1.4 du chapitre 2. Les conditions d'utilisation du système peuvent également être très variables : différents matériels (combinés téléphoniques, microphones pour la prise de son), différents canaux de transmission (lignes téléphoniques, air ambiant) et/ou différents fonds sonores (niveaux de bruits).

La construction de modèles clients "parfaits" requiert, par conséquent, une quantité infinie de données d'apprentissage pour chaque client[3]. Ce concept n'est pas concevable pour un système de VAL opérationnel. Dans le cadre d'une application commerciale, le système de VAL doit être aussi transparent que possible pour le client et les sessions d'enrôlement les moins contraignantes possibles [Bimbot et al., 1999]. Dans cette optique, il est difficile de demander à chaque client plusieurs sessions d'enrôlement mettant en jeu diverses conditions d'utilisation du système (appel du domicile, du lieu de travail, à différentes heures de la journée[4]...). Aussi, la quantité de données d'apprentissage collectées durant ces sessions est généralement très faible, entraînant une mauvaise estimation des modèles clients.

Dans de telles conditions, la qualité de l'estimation des vraisemblances par l'approche EMV est considérablement réduite. Les conséquences, au sein des systèmes de VAL, sont les suivantes :

1. Pour un locuteur donné (et son modèle associé) et différents signaux de parole produits par ce même locuteur (accès client), une variance significative des valeurs de vraisemblances (vraisemblances intra-locuteurs) peut être observée, suivant le degré de variabilité entre les signaux d'apprentissage et de test (variations dues au locuteur et au matériel) [Li et al., 1988]. Lors d'accès imposteurs, la variance des valeurs de vraisemblances (vraisemblances inter-locuteurs) observées est nettement accentuée.
2. La présence de variations entre les signaux d'apprentissage (utilisés pour la construction des modèles) et de test (signaux de parole y_N), amenées notamment par un changement de matériel peut causer de graves dégradations de performances du système de VAL [Van Vuuren, 1996], [Heck et al., 1997], [Reynolds, 1997].

3 Seuil de décision

3.1 Définition du seuil de décision

La dernière phase d'un système de VAL consiste à comparer la vraisemblance $L_{\mathcal{X}}(y_N)$ à une valeur de seuil. Le résultat de cette comparaison permet au système d'accepter ou

[3] Cette condition n'est évidemment pas suffisante. Il serait aussi nécessaire d'y associer une modélisation "parfaite".

[4] La voix d'un individu peut montrer des variations au cours d'une même journée.

de rejeter l'identité revendiquée. Le seuil est, par conséquent, un facteur prépondérant dans le processus de décision. Comme précisé dans le chapitre d'Introduction (chapitre 1), il conduit à deux types d'erreurs – erreur de fausse acceptation et erreur de faux rejet – résumées par les taux de fausse acceptation et de faux rejet.

Une mesure pondérée de ces taux d'erreurs est souvent proposée dans la littérature afin de rendre compte d'un coût total, inhérent au système [Bimbot et al., 1997], [Doddington, 1998]. Cette mesure, appelée fonction de coût totale du système, est généralement présentée sous la forme suivante :

$$C = C_{FR}.p(FR).P(\mathcal{X}) + C_{FA}.p(FA).P(\overline{\mathcal{X}}) \tag{8.4}$$

où C_{FR} et C_{FA} sont les coûts respectifs d'une erreur de faux rejet et de fausse acceptation, $P(\mathcal{X})$ et $P(\overline{\mathcal{X}})$ sont les probabilités *a priori* que le signal de parole appartienne au client ou à un autre locuteur et finalement $p(FR)$ et $p(FA)$ sont les taux de faux rejet et de fausse acceptation engendrés par le système.

3.2 Choix du seuil de décision

Le choix du seuil de décision est une phase critique dans le processus de décision. En effet, les taux d'erreurs $p(FA)$ et $p(FR)$ dépendent directement de la valeur de ce seuil. Une déviation entre la valeur théorique d'un seuil et la valeur choisie pour une application donnée se traduit en VAL par une déviation comparable des taux $p(FA)$ et $p(FR)$. Cette déviation des taux d'erreurs peut se révéler préjudiciable pour l'application visée.

Dans la littérature, le seuil de décision est généralement fixé de manière à minimiser une fonction de risque du système de VAL [Bimbot et al., 1997], [Lindberg et al., 1998], [Doddington, 1998], [Genoud, 1999]. Cette fonction de risque s'apparente à la fonction de coût que nous venons de présenter. Néanmoins, elle est définie *a priori* pour répondre aux contraintes du domaine applicatif dans lequel le système de VAL intervient. En effet, selon l'application considérée, une erreur de faux rejet peut avoir un coût négligeable face à une fausse acceptation (service de transactions bancaires par exemple) et inversement. Dans ce cadre, le seuil de décision est assimilé à un point de fonctionnement du système, défini par les contraintes de l'application.

En pratique, le seuil *a priori* est déterminé – selon la fonction de risque à minimiser – à partir des distributions des vraisemblances clients et imposteurs, estimées sur un jeu de données de réglage.

Par ailleurs, le choix du seuil doit prendre en compte d'autres considérations :

- Le seuil de décision doit être indépendant du locuteur i.e. unique pour tous les clients considérés par le système de VAL. En effet, l'indépendance au locuteur conduit à une seule recherche du seuil optimal et facilite l'ajout de nouveaux modèles clients au sein du système (nouveaux clients d'une application).
- Dans la même optique, le seuil doit rester optimal face à d'éventuels changements dans les conditions d'utilisation du système de VAL.

Pour résumer, le seuil de décision doit respecter les contraintes fixées par l'application considérée, ne doit pas dépendre des clients de cette application et doit être stable quelles

que soient les conditions d'utilisation du système de VAL.

4 Comparaison des vraisemblances et du seuil de décision

Le choix d'un seuil global fixé *a priori* impose au système de VAL de produire une distribution de vraisemblances de faible variance quels que soient les conditions d'utilisation du système ou les locuteurs considérés.

Or, la qualité de l'estimation des vraisemblances par l'approche EMV, discutée en section 2, ne permet pas de respecter une telle contrainte au sein des systèmes de VAL. En effet, les vraisemblances intra-locuteurs (accès clients) et inter-locuteurs (accès imposteurs) produites par les systèmes de VAL actuels présentent, au contraire, une forte variabilité, entraînant une grande variance au sein des distributions.

"[...] the variability of matching scores for a single speaker clearly shows that no absolute threshold on raw scores[5] *can be chosen which will give reliable decisions even if the threshold is speaker dependent*[6]*. [...]"* [Li et al., 1988].

Il apparaît que la problématique du processus de décision est liée à la qualité d'estimation des vraisemblances et aux contraintes imposées pour le choix du seuil de décision.

La solution majoritairement adoptée dans la littérature est de pallier ce problème en proposant des techniques de normalisation. Ces techniques visent à minimiser la plage de variation des vraisemblances afin de favoriser l'utilisation d'un seuil global fixé *a priori*.

[5] L'expression "raw scores" fait référence à l'utilisation de vraisemblances.

[6] "la variabilité des scores pour un locuteur donné montre clairement qu'aucun seuil absolu, même dépendant du locuteur, choisi sur les scores bruts, ne donnera de décisions efficaces."

Chapitre 9

Techniques de normalisation : État de l'art

Ce chapitre propose un état de l'art des différentes techniques de normalisation proposées dans la littérature. Ces techniques s'attachent essentiellement à réduire la variabilité des vraisemblances due au matériel ou au locuteur.

La variabilité des vraisemblances intra- et inter-locuteurs a suscité de nombreux travaux de recherche ces dix dernières années. Dans la majorité des cas, ces travaux ont mené à l'élaboration de techniques de normalisation permettant de réduire cette variabilité. La diversité des facteurs à l'origine de cette variabilité (évoquée au chapitre précédent) a initié différentes voies d'investigation permettant de répertorier trois grandes classes de normalisations, détaillées dans les sections suivantes :

- Normalisation de l'espace des paramètres acoustiques ;
- Normalisation de l'espace des mesures de similarité ;
- Normalisation de l'espace des seuils.

1 Espace des paramètres acoustiques

La première grande classe de normalisations s'intéresse, comme son nom l'indique, à l'espace des paramètres acoustiques (parameter-domain normalization [Furui, 1997]). Les techniques de normalisation concernées sont appliquées lors de la phase de paramétrisation. Elles ont pour objectif de réduire les distorsions présentes dans tout signal de parole et dues aux canaux de transmission. Appliquées sur les signaux d'apprentissage et de test, ces techniques ont pour effet de minimiser les variations – dues aux canaux de transmission – entre ces deux entités.

Ces techniques sont, en principe, indépendantes de la tâche visée. Aussi, leur usage est fréquent en RAL mais également en Reconnaissance Automatique de la Parole (RAP) [Hermansky et al., 1994], [Mokbel et al., 1994], [Matrouf, 1997], [Mokbel et al., 1998].

Plusieurs techniques, connues pour leur efficacité en VAL, sont présentées ici.

1.1 Retrait de la moyenne cepstrale

La technique la plus classique est la normalisation de la moyenne cepstrale (Cepstral Mean Normalization : CMN). Cette normalisation est fondée généralement sur le retrait de la moyenne cepstrale, encore appelée soustraction du cepstre moyen (Cepstral Mean Subtraction). Elle s'applique lors d'une paramétrisation du signal de parole en vecteurs de coefficients cepstraux (illustration en figure 9.1).
Elle s'appuie sur l'hypothèse suivante : *en considérant que les distorsions du signal sont suffisamment stables sur un intervalle de temps long alors la moyenne des coefficients cepstraux sur ce même intervalle de temps est une estimation raisonnable des distorsions cepstrales* [Atal, 1974], [Furui, 1981], [Rosenberg et al., 1994].

Soit $\{y_t\}_{1 \leq t \leq N}$ une séquence de vecteurs de coefficients cepstraux de dimension p représentant un signal de parole paramétrisé et $\overline{y_t}$ le vecteur cepstral moyen correspondant, défini par :

$$\overline{y_t} = \frac{1}{N} \sum_{t=1}^{N} y_t \tag{9.1}$$

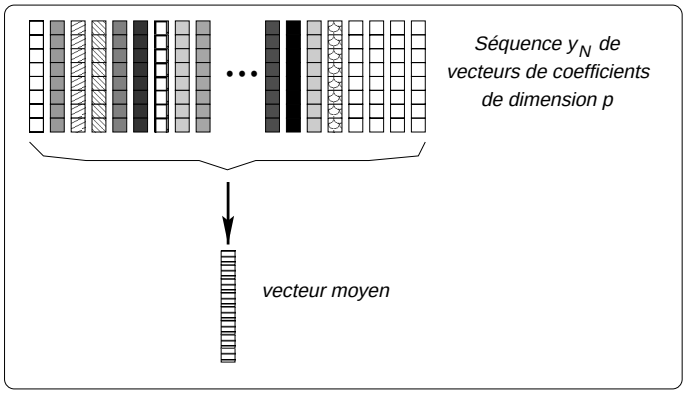

Figure 9.1: Retrait de la moyenne cepstrale. Illustration de l'estimation de la moyenne cepstrale sur l'ensemble du signal.

L'application du retrait de la moyenne sur un vecteur y_t se décrit par :

$$y'_t = y_t - \overline{y_t} \tag{9.2}$$

où y'_t représente le vecteur de coefficients cepstraux normalisé par retrait de la moyenne cepstrale.

Il est important de noter que l'efficacité de cette normalisation est très dépendante de la longueur du signal de parole sur lequel est estimée la moyenne cepstrale. En effet, le retrait de la moyenne cepstrale peut avoir pour effets secondaires la suppression d'informations pertinentes (informations spécifiques du locuteur par exemple) véhiculées par le signal. La CMS n'est donc pas appropriée pour des signaux de parole de très courtes durées, notamment en VAL [Furui, 1997].

Par ailleurs, la CMS n'est pas applicable pour les tâches de suivi de locuteurs par exemple. Dans ce contexte, la séquence de parole considérée est une concaténation de signaux de parole produits par plusieurs locuteurs et potentiellement transmis sur des canaux de transmission différents. L'estimation de la moyenne cepstrale sur l'ensemble de la séquence de parole perd alors tout son intérêt.

1.2 Retrait de la moyenne mobile

Cette variante du retrait de la moyenne cepstrale relâche l'hypothèse émise précédemment et suppose que les distorsions cepstrales varient sur un intervalle de temps donné. La moyenne cepstrale calculée sur l'ensemble du signal de parole n'est donc plus applicable. Elle est remplacée par une moyenne cepstrale estimée sur une fenêtre temporelle glissant le long du signal de parole [Rosenberg et al., 1994] (figure 9.2).

Soit $\{y_k\}_{t \leq k < t+T}$ la séquence de vecteurs de coefficients cepstraux issus de la fenêtre temporelle de taille T et $\overline{y_T}$ le vecteur cepstral moyen correspondant défini par :

$$\overline{y_T} = \frac{1}{T} \sum_{k=1}^{T} y_{t+k} \tag{9.3}$$

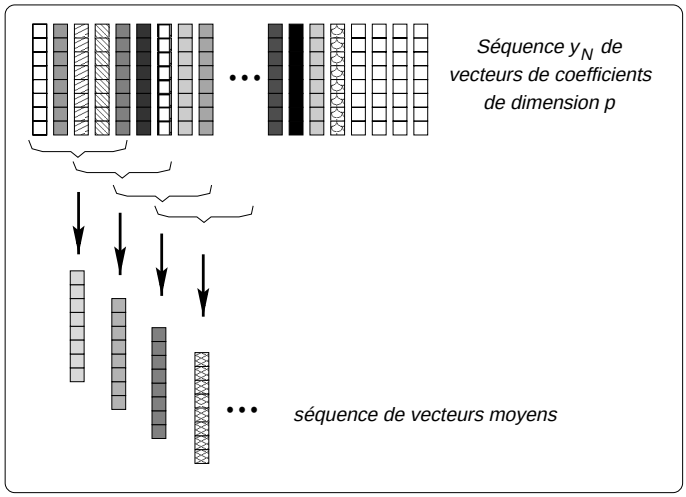

Figure 9.2: Retrait de la moyenne mobile. Illustration de l'estimation de la moyenne cepstrale mobile sur une fenêtre composée de 5 trames glissant le long du signal.

L'application du retrait de la moyenne mobile sur le vecteur y_k issu de la séquence $\{y_k\}_{t \leq k < t+T}$ s'exprime par :

$$y'_k = y_k - \overline{y_T} \tag{9.4}$$

1.3 Filtrage RASTA

Il est souvent observé que les bruits additifs et convolutifs varient lentement par rapport au signal de parole. Le filtrage RASTA (RelAtive SpecTrAl processing) repose sur ces observations. Il permet de filtrer les variations lentes du signal et d'éliminer ainsi une grande part des distorsions liées aux canaux de transmission (le lecteur se reportera à la référence [Hermansky et al., 1994] pour plus de détails sur les aspects théoriques de cette technique).

1.4 Coefficients dynamiques : Delta et Delta-Delta

L'utilisation de coefficients cepstraux dynamiques de type Delta et Delta-Delta (le détail du calcul des coefficients Delta et Delta-Delta est donné au chapitre 4) au sein du système de VAL est souvent considérée comme une forme de normalisation [Reynolds, 1996], [Furui, 1997]. En effet, les coefficients dynamiques sont bien connus pour leur robustesse face aux distorsions linéaires du signal de parole [Furui, 1981], [Soong et al., 1988]. Cette propriété s'explique notamment par le fait que ces coefficients dynamiques sont relativement indépendants des variations lentes additives dans le domaine cepstral [Matrouf, 1997].

1.5 Remarque sur la linéarité des distorsions du signal

Il est communément admis que les techniques évoquées précédemment sont efficaces dans le traitement des distorsions linéaires du signal de parole. Néanmoins, [Reynolds, 1996] montre, sur la base d'expériences, que ces techniques ne sont pas suffisantes pour pallier toutes les distorsions présentes dans le signal. Il démontre, par exemple,

que les distorsions du signal pour un combiné téléphonique donné sont dépendantes du niveau d'énergie du signal et préconise par conséquent un traitement particulier suivant les niveaux d'énergie observés. Néanmoins, la solution proposée dans ce contexte, qui se résume à un retrait de la moyenne cepstrale dépendante du niveau d'énergie, s'avère peu probante en termes d'amélioration des taux de reconnaissance du locuteur.

2 Espace des mesures de similarité

La deuxième grande classe de normalisations met en jeu l'espace des mesures de similarité (similarity-domain normalization [Gravier et al., 1998]/likelihood normalization [Furui, 1997]). Ces techniques sont appliquées en aval de la phase d'estimation des vraisemblances. Elles visent à réduire la variabilité des vraisemblances imputable à des changements de matériel ou des conditions d'enregistrement entre les tests.

Une des particularités de ce type de normalisation est de dépendre fortement de la tâche visée ainsi que des techniques de modélisation retenues. À l'opposé, les techniques de normalisation de l'espace des paramètres acoustiques, notamment, sont généralement indépendantes de l'application.

2.1 Test d'hypothèses et rapport de vraisemblances

Le processus de décision, détaillé au chapitre précédent, est souvent assimilé dans la littérature à un test d'hypothèses [Liu et al., 1996], [Gravier et al., 1998], [Genoud, 1999], [Fredouille et al., 1999] dans lequel :

- l'hypothèse H_0 représente l'événement : *le signal de parole, y_N, est produit par le modèle client $\mathcal{X}$.*
- l'hypothèse H_1 représente l'événement : *y_N est produit par un autre modèle.*

Sous un test d'hypothèses, la règle de décision s'écrit sous la forme :

$$\begin{array}{rcl} p(H_0) & \geq & p(H_1) \Rightarrow Acceptation \\ p(H_0) & < & p(H_1) \Rightarrow Rejet \end{array} \tag{9.5}$$

ou encore :

$$\begin{array}{rl} \dfrac{p(H_0)}{p(H_1)} & \geq \Theta \Rightarrow Acceptation \\ \dfrac{p(H_0)}{p(H_1)} & < \Theta \Rightarrow Rejet \end{array} \tag{9.6}$$

En VAL, la probabilité $p(H_0)$ (resp. $p(H_1)$) est représentée par la vraisemblance, $L_{\mathcal{X}}(y_N)$ (resp. $L_{\overline{\mathcal{X}}}(y_N)$), pour que le signal de parole représenté par y_N soit émis par le modèle de locuteur $\mathcal{X}$ (resp. par le modèle du non-locuteur $\overline{\mathcal{X}}$). Le non-locuteur $\overline{X}$ représente tous les locuteurs potentiels autres que le locuteur client X.

L'équation 9.6 devient alors :

$$\begin{aligned} \frac{L_{\mathcal{X}}(y_N)}{L_{\overline{\mathcal{X}}}(y_N)} &\geq \Theta \quad \Rightarrow Acceptation \\ \frac{L_{\mathcal{X}}(y_N)}{L_{\overline{\mathcal{X}}}(y_N)} &< \Theta \quad \Rightarrow Rejet \end{aligned} \tag{9.7}$$

ou en considérant le domaine logarithmique :

$$\begin{aligned} LL_{\mathcal{X}}(y_N) - LL_{\overline{\mathcal{X}}}(y_N) &\geq \Theta' \quad \Rightarrow Acceptation \\ LL_{\mathcal{X}}(y_N) - LL_{\overline{\mathcal{X}}}(y_N) &< \Theta' \quad \Rightarrow Rejet \end{aligned} \tag{9.8}$$

Ce rapport de vraisemblances, opposant le modèle du locuteur et celui du non-locuteur, est proposé dans la littérature comme une normalisation de l'espace des mesures de similarité [Liu et al., 1996], [Furui, 1997], [Gravier et al., 1998].

Cette technique est essentiellement motivée par le principe suivant :
Soit un signal de parole, y_N, présentant des caractéristiques nouvelles (présence de bruit par exemple) comparé aux signaux d'apprentissage des modèles locuteur et non-locuteur. Le signal de parole peut se diviser en deux composantes : y_a (signal présentant des caractéristiques comparables à celles des signaux d'apprentissage) et y_b (signal représentant les nouvelles caractéristiques). Dans ce contexte, la vraisemblance $L_{\mathcal{X}}(y_N)$ (resp. $L_{\overline{\mathcal{X}}}(y_N)$) est remplacée par le produit des vraisemblances $L_{\mathcal{X}}(y_a)$ et $L_{\mathcal{X}}(y_b)$ (resp. $L_{\overline{\mathcal{X}}}(y_a)$ et $L_{\overline{\mathcal{X}}}(y_b)$). L'introduction de ces produits au sein de l'equation 9.7 conduit à :

$$\begin{aligned} \frac{L_{\mathcal{X}}(y_a).L_{\mathcal{X}}(y_b)}{L_{\overline{\mathcal{X}}}(y_a).L_{\overline{\mathcal{X}}}(y_b)} &\geq \Theta \quad \Rightarrow Acceptation \\ \frac{L_{\mathcal{X}}(y_a).L_{\mathcal{X}}(y_b)}{L_{\overline{\mathcal{X}}}(y_a).L_{\overline{\mathcal{X}}}(y_b)} &< \Theta \quad \Rightarrow Rejet \end{aligned} \tag{9.9}$$

Les nouvelles caractéristiques de y_b étant inconnues des modèles locuteur et non-locuteur, il est raisonnable de supposer que l'estimation des vraisemblances $L_{\mathcal{X}}(y_b)$ et $L_{\overline{\mathcal{X}}}(y_b)$ conduira à des valeurs de vraisemblances très semblables pour les deux modèles [Bimbot et al., 1998]. Les vraisemblances $L_{\mathcal{X}}(y_b)$ et $L_{\overline{\mathcal{X}}}(y_b)$ s'annulent, par conséquent, dans l'équation 9.9.

Dans cette optique, le calcul du rapport de vraisemblances est une approche théorique simple pour éliminer tout biais induit par le signal de parole y_N (variations entre données d'apprentissage et de test) dans l'estimation des vraisemblances.

Cependant, la construction d'un modèle de non-locuteur n'est pas concevable dans un cadre autre que théorique. En effet, un modèle représentant tous les locuteurs potentiels hormis le locuteur X demande une quantité de données d'apprentissage infinie ainsi qu'un bon estimateur de modèle capable de prendre en compte toutes les caractéristiques présentes dans les données. Une approximation du modèle $\overline{\mathcal{X}}$ est donc nécessaire pour la mise en pratique du rapport de vraisemblances.

2.1.1 Cohorte de locuteurs

Les travaux de [Higgins et al., 1991], repris par [Rosenberg et al., 1992] suggèrent d'utiliser un groupe de locuteurs, appelé cohorte, soigneusement sélectionnés pour l'approximation du modèle $\overline{\mathcal{X}}$. Cette sélection est dépendante du locuteur ; chaque locuteur de la population est associé à une cohorte de locuteurs qui lui est propre.

En général, la sélection d'une cohorte par locuteur est réalisée grâce à une mesure de similarité entre modèles : pour un locuteur donné X et son modèle associé $\mathcal{X}$, la cohorte correspondante, notée $C(X)$, est constituée de locuteurs dont le modèle est le plus proche du modèle $\mathcal{X}$. Cette notion de proximité entre deux modèles de locuteurs peut se résumer, dans ce cas, par une distance croisée entre modèles et échantillons de signal (voir [Reynolds, 1995] pour le détail de cette distance).

Plusieurs facteurs sont à considérer lors de l'utilisation d'une cohorte de locuteurs :

1. Quelle taille choisir pour la cohorte (nombre de locuteurs) ?
2. Faut-il construire un seul modèle à partir de l'ensemble des données d'apprentissage des locuteurs de la cohorte ou apprendre un modèle par locuteur de la cohorte ?
3. Les locuteurs de la cohorte doivent-ils être issus de la même population que le locuteur X ou d'une population différente présentant des conditions d'enregistrement comparables ?
4. La sélection de la cohorte par mesure de similarité entre modèles est-elle justifiée ?

Des expériences menées par [Rosenberg et al., 1996] tentent d'apporter une réponse à ces questions. Après analyse des résultats, il semble préférable d'utiliser un seul modèle $\overline{\mathcal{X}}$ plutôt que des modèles individuels pour chaque locuteur de la cohorte.

Par ailleurs, la comparaison entre une cohorte de 5 locuteurs, sélectionnés[1] dans la même population que les locuteurs clients, et une cohorte de 40 locuteurs, tirés aléatoirement dans une population distincte, ne montre pas d'écarts significatifs en termes de performances. Ce résultat montre qu'une position absolue vis à vis de la taille de la cohorte (5 locuteurs contre 40 dans l'exemple précédent), du choix de la population (deux populations différentes) ou du critère de sélection (mesure de similarité/hasard) ne peut être envisagée. De surcroît, ces facteurs sont probablement très dépendants de la population de locuteurs clients ainsi que de l'application visée.

Une des justifications, apportée par [Higgins et al., 1991], sur le choix d'une cohorte proche des locuteurs clients est de rendre le système de VAL plus robuste face à des imposteurs présentant des voix similaires à celles des locuteurs clients. Néanmoins, l'effet inverse n'est pas controlé par cette approche, rendant le système très vulnérable à des voix d'imposteurs très différentes de celles des locuteurs clients – une femme usurpant l'identité d'un homme par exemple. Dans un tel cas de figure, les vraisemblances $L_{\mathcal{X}}(y_N)$ et $L_{\overline{\mathcal{X}}}(y_N)$ sont très faibles et présentent des valeurs similaires. Le rapport des vraisemblances conduit à des valeurs avoisinant l'unité, assimilées par le système – selon la valeur du seuil de décision – à une acceptation.

[Reynolds, 1995] apporte une première solution en proposant une cohorte dans laquelle sont réunis des locuteurs proches et des locuteurs éloignés du locuteur client (illustration

[1] La sélection est réalisée à l'aide d'une distance entre modèles.

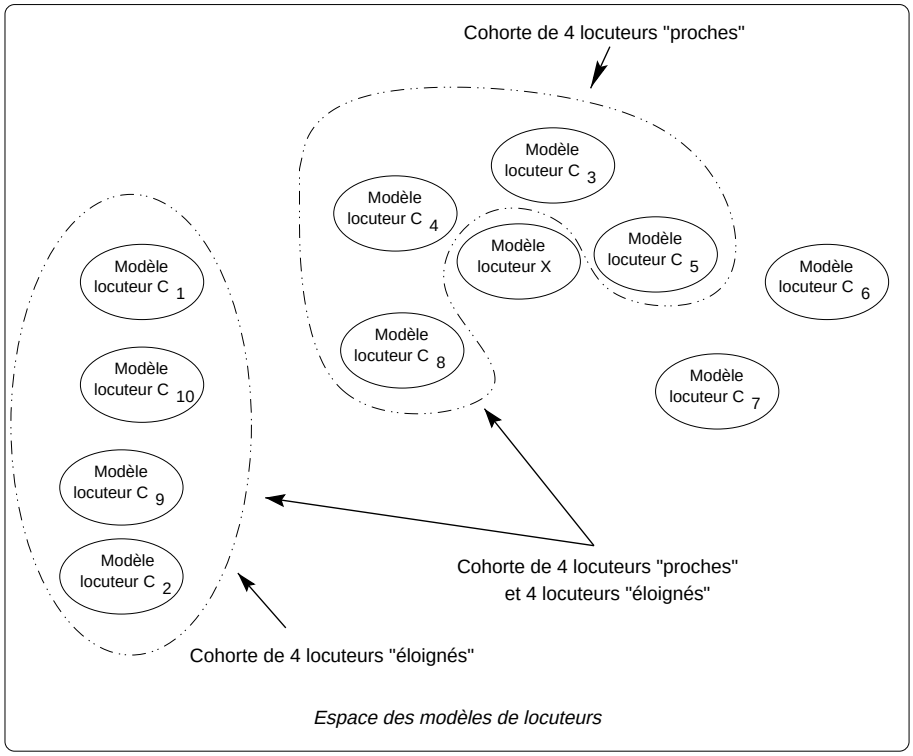

Figure 9.3: Cohortes de locuteurs. Illustration de la construction de plusieurs cohortes pour le locuteur X par distance entre modèles de locuteurs : cohorte de locuteurs proches, cohorte de locuteurs éloignés et cohorte de locuteurs proches et éloignés.

figure 9.3). Des expériences, dans ce contexte, montrent que la cohabitation de ces deux sous-ensembles de locuteurs améliore les performances, comparée à une cohorte de locuteurs proches.

[Liu et al., 1996] suggèrent l'utilisation d'imposteurs pour construire le modèle $\overline{\mathcal{X}}$. Sur le principe de la cohorte, un ensemble d'imposteurs est sélectionné dans la population des locuteurs clients[2] lors de l'apprentissage. Cette cohorte d'imposteurs reste dépendante du locuteur.
Des expériences comparatives entre les approches de [Liu et al., 1996] et [Rosenberg et al., 1996] montrent un avantage, en termes de performances, pour la cohorte d'imposteurs.

Une autre alternative, proposée par [Ariyaeeinia et al., 1997], consiste à sélectionner la cohorte de locuteurs lors de la phase de reconnaissance. En d'autres termes, la cohorte de locuteurs est formée en fonction du signal de test. Sous cette condition, le modèle $\overline{\mathcal{X}}$ est implicitement issu d'une cohorte de locuteurs proches du locuteur X dans le cas d'un test client (le signal de test appartenant au locuteur X) ou d'une cohorte d'imposteurs dans le cas d'un test imposteur (le signal de test appartenant à un autre locuteur que X). Comparée à l'utilisation d'une cohorte classique [Rosenberg et al., 1992], cette approche apporte un gain pertinent en performances. L'inconvénient majeur de cette technique est

[2]Pour un locuteur issu d'une population de N individus, les $N-1$ autres locuteurs peuvent être considérés comme des imposteurs potentiels.

que le modèle $\overline{\mathcal{X}}$ ne peut être pré-calculé avant la phase de reconnaissance.

2.1.2 Cohorte "virtuelle"

Dans [Isobe et al., 1999], les cohortes de locuteurs sont jugées peu satisfaisantes pour fournir une bonne approximation du modèle de non-locuteur. En effet, la sélection des locuteurs de la cohorte par simple distance entre modèles est remise en cause. Les auteurs considèrent que cette sélection n'est pas assez fine pour conduire à une bonne estimation du modèle de non-locuteur.
Dans cette optique, les auteurs proposent de construire une cohorte "virtuelle" en sélectionnant, parmi les modèles de locuteurs potentiels, certaines composantes jugées proches des composantes du modèle client. Dans le cas de modèles GMM par exemple, la sélection est réalisée, non plus au niveau du modèle global mais au niveau des distributions. La cohorte virtuelle se compose au final d'un ensemble de distributions, chacune représentant des caractéristiques acoustiques proches de celles du locuteur client.

2.1.3 Modèle générique

Les travaux de [Carey et al., 1992] introduisent la notion de modèle du monde[3] pour l'approximation du modèle de non-locuteur. Ce modèle du monde a pour objectif de représenter une population générique de locuteurs.
Le principal avantage de cette approche est de considérer un modèle $\overline{\mathcal{X}}$ indépendant des locuteurs clients. Contrairement à l'utilisation d'une cohorte, aucune procédure de sélection n'est nécessaire durant la phase d'apprentissage ou de test.

Il est néanmoins important de noter que la pertinence du modèle du monde dépend fortement de la composition de la population générique. Cette dernière doit contenir un nombre suffisant de locuteurs pour couvrir un espace acoustique le plus large possible sans pour autant sur-représenter l'espace acoustique d'un client particulier.

En pratique, la population générique est choisie indépendamment de la population des locuteurs clients (et imposteurs) [Reynolds, 1997].

2.2 Probabilité *a posteriori*

Outre le rapport de vraisemblances évoqué précédemment, [Matsui et al., 1994a] proposent une seconde méthode de normalisation de l'espace des mesures de similarité. Cette méthode repose sur la définition en VAL de la probabilité *a posteriori* du locuteur X, ayant observé un échantillon de parole y_N. Cette probabilité est définie par :

$$p(X|y_N) = \frac{p(y_N|X).P(X)}{\sum_i p(y_N|X_i).P(X_i)} \tag{9.10}$$

où X_i est un locuteur quelconque, $p(y_N|X)$ est la probabilité que le signal de parole y_N soit émis par le locuteur X et $P(X_i)$ est la probabilité *a priori* d'un locuteur X_i, considérée equi-probable pour tous les locuteurs.

[3] Plusieurs dénominations sont données dans la littérature pour ce modèle : modèle générique, modèle du monde ou encore modèle universel (Universal Background Model : UMB).

En omettant les probabilités *a priori* $P(X_i)$ et $P(X)$ supposées constantes, l'équation 9.10 peut s'écrire plus simplement :

$$p(X|y_N) = \frac{p(y_N|X)}{\sum_i p(y_N|X_i)} \tag{9.11}$$

Pour un échantillon de test donné, une approximation du terme $\sum_i p(y_N|X_i)$ est obtenue, dans [Matsui et al., 1994a], par la somme des n plus grandes valeurs de vraisemblances, $\sum_{i=1}^{n} L_{\mathcal{X}_i}(y_N)$, calculées sur l'ensemble de la population des locuteurs clients. La vraisemblance du locuteur X est systématiquement incluse dans cette somme.

L'inconvénient majeur de cette approche est qu'elle requiert, pour un signal de parole y_N, d'estimer la vraisemblance $L_{\mathcal{X}_i}(y_N)$ pour chaque locuteur i de la population. Pour une large population, cette procédure peut être très coûteuse en temps de calcul.

L'approche par "probabilité *a posteriori*" diffère de l'approche par "rapport de vraisemblances" par le simple fait que le locuteur X intervient lors de la normalisation ($\sum_i p(y_N|X_i)$). Des expériences comparatives montrent néanmoins que l'inclusion ou l'exclusion du locuteur X n'a pas d'influence sur la qualité de la normalisation dès lors que l'ensemble des locuteurs impliqués dans cette normalisation est suffisamment large [Matsui et al., 1994a].

3 Espace des seuils

La dernière grande classe de normalisations s'attache à minimiser la variabilité intra-locuteur des vraisemblances. Les fondements de cette normalisation sont dérivés des travaux de [Li et al., 1988] sur l'observation des distributions des vraisemblances issues de tests clients (vraisemblances intra-locuteurs) et imposteurs (vraisemblances inter-locuteurs). Comme indiqué au chapitre précédent (section 4), [Li et al., 1988] constatent de très grandes variances au sein des distributions des vraisemblances intra- et inter-locuteurs rendant impossible l'utilisation d'un seuil global fixé *a priori*.

3.1 Normalisation des vraisemblances imposteurs

La variance des distributions des vraisemblances inter-locuteurs (imposteurs) étant sensiblement plus importante, la solution proposée par [Li et al., 1988] est de normaliser cette distribution afin de "contrôler" plus facilement la variance globale (client et imposteur) au sein des distributions des vraisemblances.

Cette normalisation consiste à rendre la distribution des vraisemblances issues de tests imposteurs centrée et réduite (distribution gaussienne de moyenne 0 et de variance 1). Elle s'applique sur chaque vraisemblance, produite par le système de VAL, de la façon suivante :

$$L_{\mathcal{X}}(y_N)_{norm} = \frac{L_{\mathcal{X}}(y_N) - \mu}{\sigma} \tag{9.12}$$

où μ et σ – représentant respectivement une moyenne et une variance – sont les paramètres de la normalisation.

Il est à noter qu'une approche similaire fut appliquée par [Furui, 1981] pour la recherche de seuils dépendants du locuteur.

3.1.1 Znorm

La technique Znorm estime les paramètres μ et σ en étudiant la réponse de chaque modèle client face à des signaux de parole appartenant à un autre individu (tests d'imposture) [Gravier et al., 1998], [Gravier et al., 2000]. Une représentation de cette technique est donnée sur la figure 9.4.

Soit un client X (et son modèle $\mathcal{X}$) et une population d'imposteurs[4], chaque imposteur ayant produit un signal de parole, représenté par y_N^i.
À partir des vraisemblances $L_{\mathcal{X}}(y_N^i)$ calculées pour chaque signal de parole y_N^i, une distribution de vraisemblances peut être estimée pour le locuteur X. Cette distribution dépendante du locuteur X est caractérisée par les paramètres μ_X et σ_X. Ces paramètres sont utilisés pour normaliser les vraisemblances issues du modèle client $\mathcal{X}$, lors des tests de vérification, suivant l'équation 9.13 :

$$L_{\mathcal{X}}(y_N)_{Znorm} = \frac{L_{\mathcal{X}}(y_N) - \mu_{\mathcal{X}}}{\sigma_{\mathcal{X}}} \tag{9.13}$$

3.1.2 Tnorm

Une autre façon d'estimer les paramètres μ et σ est d'étudier la réponse de modèles imposteurs face au signal de test [Auckenthaler et al., 2000].
Cette technique, appelée Tnorm, conduit à des paramètres μ et σ dépendants du signal de test et non plus du modèle de locuteur, comme le montre la figure 9.5.

Un des avantages de cette approche est d'assurer un niveau de variation équivalent entre :

1. le signal de parole y_N et les signaux de parole impliqués dans la construction des modèles d'imposteurs, intervenant dans l'estimation des paramètres μ et σ.
2. le signal de parole et les signaux d'apprentissage utilisés pour la construction du modèle client, intervenant dans l'estimation de la vraisemblance à normaliser.

Cet aspect de l'approche Tnorm fait défaut à l'approche Znorm où les paramètres μ et σ sont estimés indépendamment du signal de parole y_N.

Soit y_N un signal de parole issu d'un test de vérification et une population d'imposteurs. Chaque imposteur est caractérisé par son modèle $\mathcal{I}_i$. Dans ces conditions, les vraisemblances $L_{\mathcal{I}_i}(y_N)$ forment une distribution de vraisemblances, de moyenne μ_{y_N} et de variance σ_{y_N}. Ces paramètres sont utilisés pour normaliser les vraisemblances relatives au signal y_N indépendamment du modèle de locuteur concerné par le test. L'équation 9.12 devient :

$$L_{\mathcal{X}}(y_N)_{Tnorm} = \frac{L_{\mathcal{X}}(y_N) - \mu_{y_N}}{\sigma_{y_N}} \tag{9.14}$$

[4]Le locuteur X n'appartient pas à cette population. Le genre des imposteurs de cette population est fonction du genre du client X.

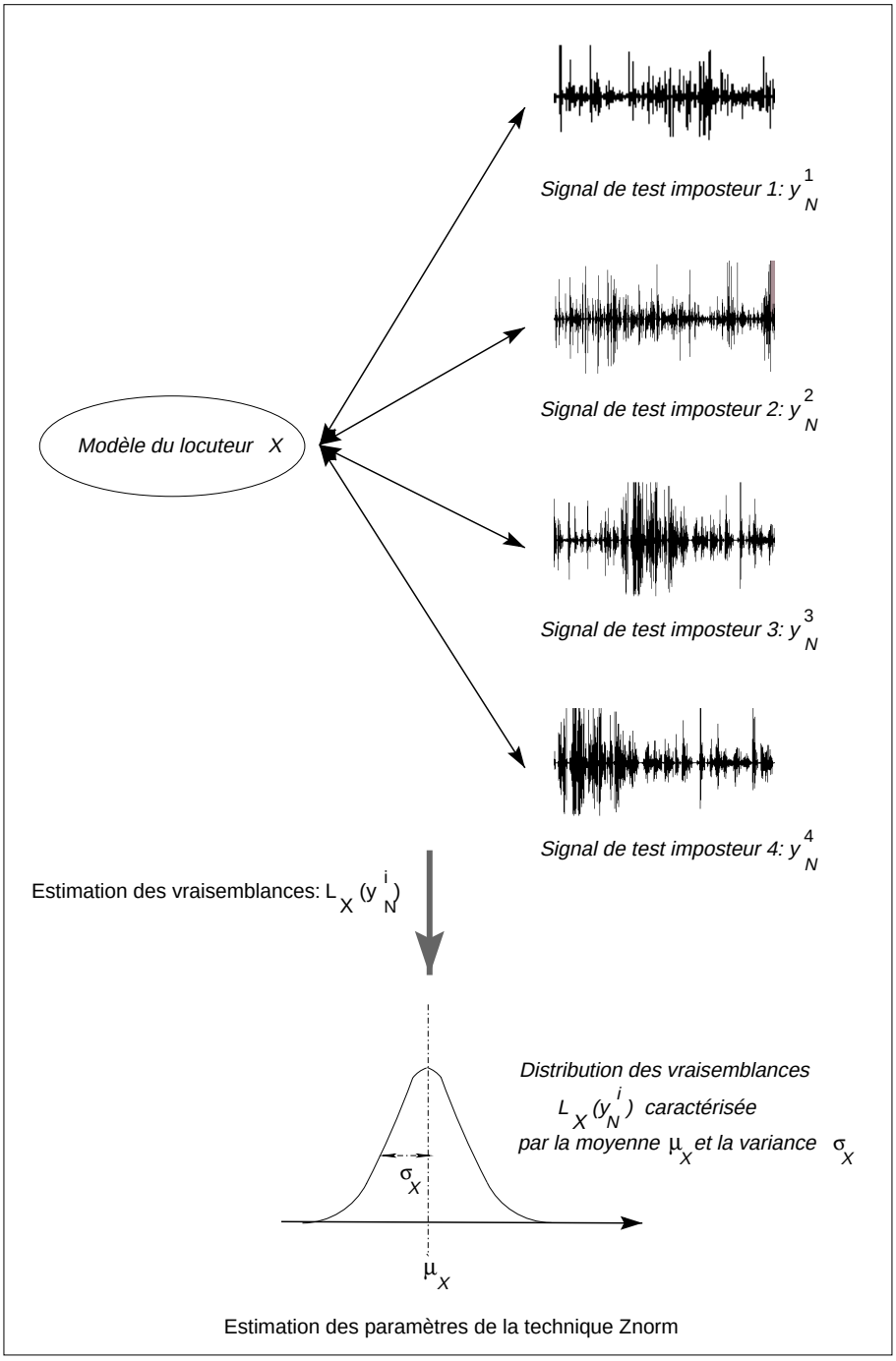

Figure 9.4: Technique Znorm. Représentation schématique de l'estimation des paramètres de la technique de normalisation : Znorm.

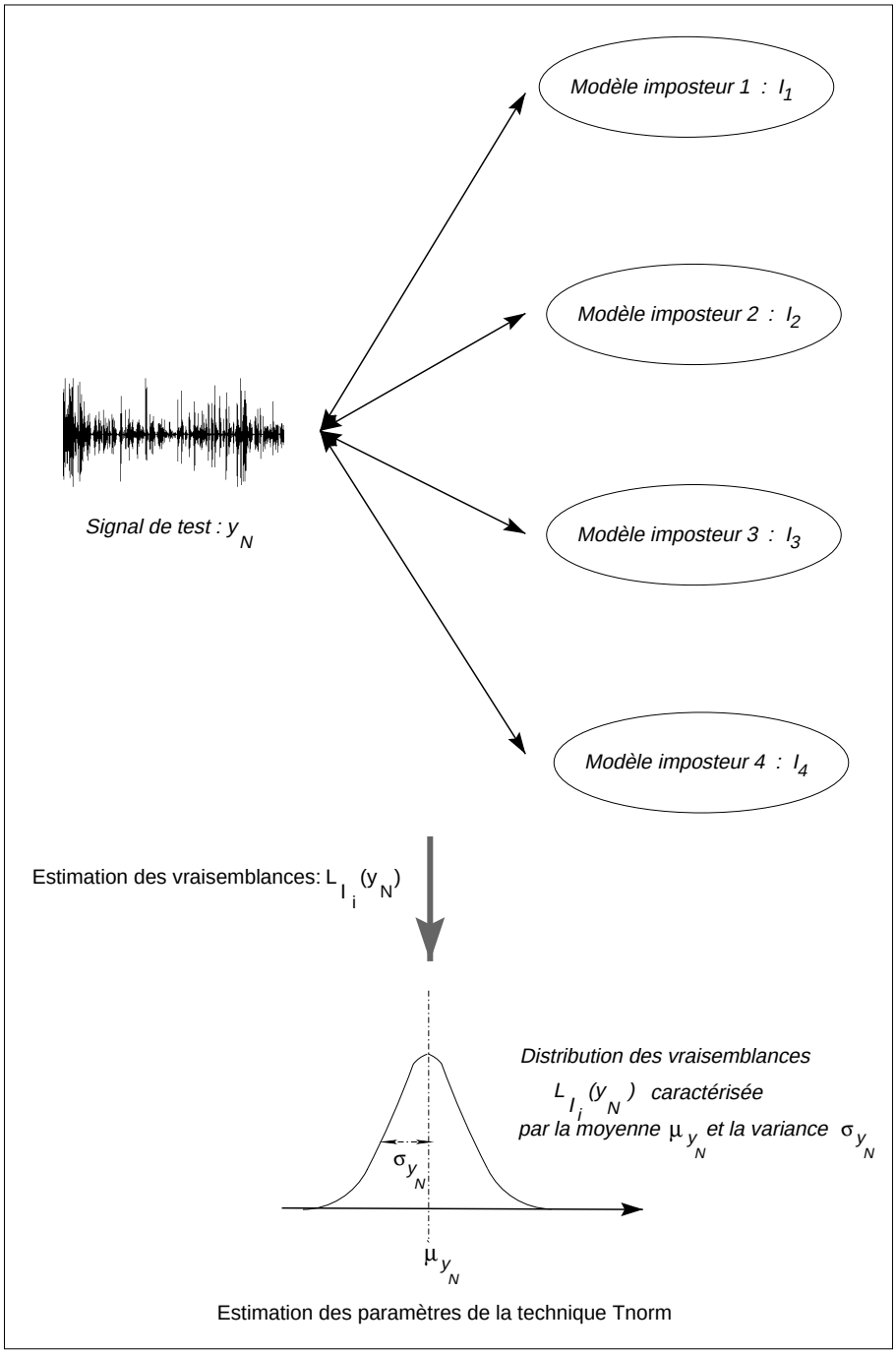

Figure 9.5: Technique Tnorm. Représentation schématique de l'estimation des paramètres de la technique de normalisation : Tnorm.

3.2 Remarque sur la normalisation des vraisemblances clients

Dans les techniques que nous venons de citer, la normalisation a pour objectif de rendre la distribution des vraisemblances imposteurs centrée et réduite. D'un point de vue théorique, cette approche est tout à fait applicable à la distribution des vraisemblances clients. D'un point de vue pratique, l'insuffisance des données des clients rend très souvent impossible l'estimation des paramètres μ et σ. En effet, cette estimation demande à être réalisée sur un ensemble de données séparé des données d'apprentissage et de test.

D'autre part, l'essence même de cette normalisation est de réduire la variabilité des vraisemblances intra-locuteur. Il est, par conséquent, nécessaire que les données utilisées à cette fin reflètent une forte variabilité de la voix du client et ce, dans différentes conditions d'utilisation du système. Or, nous avons déjà souligné le fait qu'il est difficile d'imposer à des clients d'une application de multiples sessions d'enrôlements. Par ailleurs, si nous disposions de telles données pour chaque client, les modèles de locuteurs seraient mieux appris et une telle normalisation s'avérerait alors superflue.

3.3 Ajustement des rapports de vraisemblances

Dans [Bimbot et al., 1997], les auteurs proposent un ajustement, dépendant du locuteur, du seuil de décision global fixé *a priori*. Cet ajustement s'appuie sur des distributions de vraisemblances estimées à partir du modèle de locuteur et d'un modèle du monde pour des signaux de parole appartenant au locuteur et à des imposteurs.

Les techniques précédentes s'attachent à réduire la variabilité des vraisemblances (due au locuteur) pour permettre l'utilisation d'un seuil global. Ici, l'approche est inversée. Le seuil de décision est ajusté, locuteur par locuteur, pour intégrer cette variabilité.

4 Connaissances et normalisation

L'intégration de connaissances durant le processus de normalisation permet d'améliorer l'efficacité de ce dernier. Ces connaissances peuvent concerner le genre du locuteur ou le type de combiné téléphonique utilisé. Cette proposition est motivée par [Reynolds, 1996] qui montre que de graves dégradations de performances sont observées sur les systèmes de RAL en présence d'une variation de combinés[5] téléphoniques entre les signaux d'apprentissage et de test.

La prise en compte de ces informations supplémentaires intervient principalement lors de la normalisation de l'espace des mesures de similarité ou de l'espace des seuils.

Par exemple, différents modèles génériques dépendants du genre du locuteur ou/et dépendants du type de combinés peuvent être utilisés pour une normalisation basée sur le rapport de vraisemblances [Heck et al., 1997], [Gravier et al., 1998], [Gravier et al., 2000]. Dans [Heck et al., 1997] par exemple, quatre populations de locuteurs sont impliquées pour l'apprentissage de modèles génériques, suivant le genre des locuteurs et le type de combiné téléphonique employé : "electret" ou "carbon". Quatre modèles de monde : $\overline{\mathcal{X}_{f-e}}$

[5] La variation est plus exactement liée au type de microphone intégré dans le combiné téléphonique.

(femme-electret), $\overline{\mathcal{X}_{h-e}}$ (homme-electret), $\overline{\mathcal{X}_{f-c}}$ (femme-carbon) et $\overline{\mathcal{X}_{h-c}}$ (homme-carbon) sont appris. Lors d'un test de vérification, le modèle générique est choisi parmi les quatre modèles possibles, suivant les caractéristiques du modèle client considéré. Si le modèle client correspond à un homme dont les données d'apprentissage ont été collectées au moyen d'un combiné téléphonique de type "carbon", le modèle $\overline{\mathcal{X}_{h-c}}$ est utilisé pour la normalisation de la vraisemblance.
Dans ce cas de figure, l'équation 9.7 devient :

$$\begin{aligned} \frac{L_{\mathcal{X}}(y_N)}{L_{\overline{\mathcal{X}_{h-c}}}(y_N)} &\geq \Theta \quad \Rightarrow \textit{Acceptation} \\ \frac{L_{\mathcal{X}}(y_N)}{L_{\overline{\mathcal{X}_{h-c}}}(y_N)} &< \Theta \quad \Rightarrow \textit{Rejet} \end{aligned} \tag{9.15}$$

De même, une variante de Znorm, appelée Hnorm, permet la prise en compte d'informations relatives aux types de combinés téléphoniques [Reynolds, 1996], [Reynolds et al., 2000]. Dans ce cas, la réponse des modèles clients face à des tests d'imposture est étudiée suivant différents types de combinés. Plusieurs paires de paramètres μ et σ – selon le type de combiné – sont ainsi estimées.
Par exemple, [Reynolds, 1996] suggère d'utiliser deux populations d'imposteurs selon le type de combiné considéré : electret ou carbon. Pour un modèle client $\mathcal{X}$, deux paires de paramètres sont estimées ($\mu_{\mathcal{X}_e}$, $\sigma_{\mathcal{X}_e}$) pour le type electret et ($\mu_{\mathcal{X}_c}$, $\sigma_{\mathcal{X}_c}$) pour le type carbon. Lors d'un test de vérification, si le signal de parole y_N est collecté au moyen d'un combiné téléphonique de type electret, la paire ($\mu_{\mathcal{X}_e}$, $\sigma_{\mathcal{X}_e}$) est utilisée pour la normalisation de la vraisemblance. L'équation 9.13 s'écrit alors :

$$L_{\mathcal{X}}(y_N)_{norm} = \frac{L_{\mathcal{X}}(y_N) - \mu_{\mathcal{X}_e}}{\sigma_{\mathcal{X}_e}} \tag{9.16}$$

5 Discussion

Cet état de l'art recense un très grand nombre de techniques de normalisation pour la VAL. Ces techniques, qui ont pour rôle de réduire la variabilité des vraisemblances, interviennent à différents niveaux du processus de VAL : lors de la paramétrisation (normalisation de l'espace des paramètres acoustiques), lors de l'estimation des vraisemblances (normalisation de l'espace des mesures de similarité) et lors de la décision (normalisation de l'espace des seuils).

De nombreux travaux expérimentaux font état de l'utilisation conjointe de ces différentes approches de normalisation. Dans [Heck et al., 1997], le retrait de la soustraction cepstrale (normalisation de l'espace des paramètres acoustiques) est associé à la normalisation par rapport de vraisemblances (normalisation de l'espace des mesures de similarité) s'appuyant sur des modèles du monde dépendants du type de combiné téléphonique. Dans [Reynolds, 1997], le retrait de la soustraction cepstrale ainsi que l'application d'un filtre RASTA sont associés à l'utilisation conjointe du rapport de vraisemblances (par modèle du monde) et de la technique Hnorm. D'après les résultats obtenus, il semblerait que l'utilisation en cascade de plusieurs approches de normalisation améliore les systèmes de VAL ; la variabilité des vraisemblances est minimisée à différents niveaux, permettant l'emploi d'un seuil de décision global fixé *a priori*.

Néanmoins, nous pouvons reprocher à la majorité de ces techniques d'imposer un seuil de décision dont la valeur est difficile à interpréter. En effet, si le seuil de décision est fixé *a priori* pour répondre aux contraintes d'une application, sa valeur est choisie dans l'espace des vraisemblances normalisées produites par le système de VAL. Une vraisemblance, définie dans le domaine logarithmique, peut varier de $-$ l'infini à $+$ l'infini. Quelle interprétation peut-on donner à un seuil de décision fixé à $0,5$ pour une application et à 2 pour une autre ?

Chapitre 10

Normalisation World+MAP

Afin d'apporter une solution à ce problème d'interprétation du seuil de décision, nous proposons, dans ce chapitre, une approche de normalisation originale, appelée World+MAP. Cette normalisation concilie les avantages d'un modèle du monde et de l'approche Bayesienne afin de "projeter" les vraisemblances dans un espace probabiliste.

---

1 Introduction

Les techniques de normalisation actuelles ont pour objectif principal de réduire la variabilité des vraisemblances produites par un système de VAL et impliquées dans le processus de décision. Cette réduction de la variabilité permet l'utilisation d'un seuil de décision unique (indépendant du locuteur), choisi *a priori* pour satisfaire les contraintes d'une application donnée. Ces contraintes peuvent être, par exemple, un taux de fausse acceptation ($p(FA)$) très faible pour sécuriser l'accès à des données sensibles par téléphone.

Les systèmes de VAL, s'appuyant sur de telles techniques de normalisation, montrent une amélioration notable des performances. Dans ce contexte, le seuil de décision est choisi dans le nouvel espace des vraisemblances normalisées comme point de fonctionnement du système. En considérant le domaine logarithmique, l'espace des vraisemblances normalisées s'étend de − l'infini à + l'infini. Selon les contraintes de l'application considérée, le seuil de décision peut, par conséquent, varier également dans ce même intervalle. Dans ces conditions, il est hasardeux de donner une signification à un seuil de décision sorti de son contexte applicatif.

Pourtant, attribuer une signification à un seuil de décision rendrait plus facile l'intégration des systèmes de VAL dans des contextes applicatifs différents.

Dans cette voie, nous proposons une approche originale de normalisation, appelée World+MAP. Cette approche suit les mêmes objectifs de minimisation de la variabilité des vraisemblances que les techniques "état de l'art" présentées dans le chapitre précédent. De surcroît, elle a pour rôle de donner une signification aux vraisemblances normalisées et au seuil de décision sous-jacent. De par les principes qu'elle met en œuvre, cette nouvelle approche de normalisation peut être rattachée soit aux techniques de normalisation de l'espace des mesures de similarité soit aux techniques de normalisation de l'espace des seuils.

2 Aspects théoriques de la normalisation World+MAP

La normalisation World+MAP repose sur deux concepts théoriques majeurs :

- la normalisation des vraisemblances par application du rapport de vraisemblances et utilisation d'un modèle du monde (World) ;
- l'estimation de probabilités *a posteriori* (MAP).

L'utilisation du rapport de vraisemblances a pour objectif de réduire la variabilité des vraisemblances, due notamment aux variations de matériel et des conditions d'utilisation du système entre les tests. Le deuxième concept choisi a pour objectif de projeter les vraisemblances dans un espace probabiliste. Cette projection permet de manipuler non plus des vraisemblances dans un espace $[0, +\inf]$ mais des probabilités lors du processus de décision. Dans ce contexte, la plage de variation des valeurs du seuil de décision se restreint à l'intervalle $[0, 1]$. Le seuil de décision prend toute sa signification dans le domaine probabiliste et devient facilement interprétable.

Formalisme de l'approche World+MAP

Soit la vraisemblance $L_{\mathcal{X}}(y_N)$ pour que le signal de parole, représenté par y_N, soit produit par le modèle client $\mathcal{X}$ appartenant au locuteur X.
Soit un modèle de non-locuteur, dont une approximation est donnée, dans ce chapitre, par un modèle du monde noté $\mathcal{W}$ et la vraisemblance $L_{\mathcal{W}}(y_N)$ pour que le signal y_N soit produit par le modèle $\mathcal{W}$.

La première phase de normalisation (World) de l'approche World+MAP – l'application du rapport de vraisemblances – se décrit sous la forme :

$$LR_{\mathcal{X}}(y_N) = \frac{L_{\mathcal{X}}(y_N)}{L_{\mathcal{W}}(y_N)} \tag{10.1}$$

La deuxième phase de la normalisation (MAP) s'appuie sur la théorie Bayesienne. Elle consiste à remplacer la vraisemblance normalisée (ou rapport de vraisemblances) $LR_{\mathcal{X}}(y_N)$ par la probabilité *a posteriori* pour que le locuteur du modèle $\mathcal{X}$ ait prononcé le signal y_N connaissant la vraisemblance $LR_{\mathcal{X}}(y_N)$, en d'autres termes, la probabilité d'être en présence d'un accès client sachant la valeur de la vraisemblance $LR_{\mathcal{X}}(y_N)$[1]. Cette probabilité est notée $p(X = Y|LR_{\mathcal{X}})$ où Y est le locuteur ayant prononcé le signal de parole y_N. Par application de la règle de bayes, elle s'exprime sous la forme :

$$p(X = Y|LR_{\mathcal{X}}) = \frac{p(LR_{\mathcal{X}}|X = Y).P(X = Y)}{p(LR_{\mathcal{X}}|X = Y).P(X = Y) + p(LR_{\mathcal{X}}|X \neq Y).P(X \neq Y)} \tag{10.2}$$

où $p(LR_{\mathcal{X}}|X = Y)$ (resp. $p(LR_{\mathcal{X}}|X \neq Y)$) est la probabilité de la vraisemblance $LR_{\mathcal{X}}$ sachant que l'accès est client (resp. accès imposteur) et $P(X = Y)$ (resp. $P(X \neq Y)$) est la probabilité *a priori* d'un accès client (resp. accès imposteur).

La règle de décision, définie au chapitre 8, prend ici la forme suivante :

$$\begin{aligned} p(X = Y|LR_{\mathcal{X}}) &\geq \Theta \quad \Rightarrow acceptation \\ p(X = Y|LR_{\mathcal{X}}) &< \Theta \quad \Rightarrow rejet \end{aligned} \tag{10.3}$$

2.1 Autres avantages de l'approche World+MAP

Outre l'interprétation probabiliste donnée au seuil de décision, l'approche World+MAP peut être très avantageuse dans le contexte d'un système multi-reconnaisseur (architecture multi-reconnaisseur). Dans une telle architecture, chaque reconnaisseur fournit un score (vraisemblances par exemple) en vue de la décision. Ces scores sont combinés, lors d'une étape de fusion des reconnaisseurs, afin de produire un score final au processus de décision. En appliquant l'approche World+MAP, ces scores se résument à des probabilités. Manipuler des probabilités à la place de vraisemblances doit faciliter grandement l'étape de fusion.

De plus, une des motivations de l'approche World+MAP est d'intégrer implicitement la qualité intrinsèque de chaque reconnaisseur dans les probabilités produites. Dans ces conditions, des opérateurs simples, telles que la moyenne arithmétique ou géométrique, peuvent alors être utilisés pour la fusion [Fredouille et al., 1998].

[1] Dans la suite de cette thèse, la vraisemblance normalisée $LR_{\mathcal{X}}(y_N)$ sera notée plus simplement $LR_{\mathcal{X}}$.

3 Mise en œuvre de l'approche World+MAP

Au regard des sections précédentes, l'approche World+MAP repose sur trois grandes composantes :

- le **modèle du monde** $\mathcal{W}$ est impliqué dans la première phase de normalisation (World) pour le calcul des rapports de vraisemblances.
- les **probabilités** $p(LR_{\mathcal{X}}|X = Y)$ **et** $p(LR_{\mathcal{X}}|X \neq Y)$ interviennent dans l'équation 10.2 pour la deuxième phase de normalisation (MAP). Elles sont estimées à partir de distributions de probabilités notées $\mathcal{F}_{X=Y}$ et $\mathcal{F}_{X\neq Y}$ respectivement.
- **les probabilités** ***a priori*** $P(X = Y)$ **et** $P(X \neq Y)$ sont constantes et fixées pour une application donnée. Elles offrent l'avantage d'intégrer systématiquement dans le calcul des probabilités *a posteriori* les conditions d'utilisation du système, concernant la prévision du nombre d'accès clients et imposteurs.

Ces trois composantes interviennent directement dans la mise en œuvre de l'approche World+MAP. Un jeu de données est également nécessaire impliquant des accès clients et imposteurs. Cette mise en œuvre consiste en une phase d'apprentissage d'une fonction de normalisation, utilisée par la suite pour normaliser les rapports de vraisemblances lors des tests de vérification. Comme indiqué sur la figure 10.1, différentes étapes sont nécessaires pour l'apprentissage de la fonction de normalisation :

1. des tests de vérification sont conduits sur le jeu de données résultant en l'estimation d'une vraisemblance pour chaque accès client ou imposteur.
2. ces vraisemblances sont normalisées par application du rapport de vraisemblances et l'utilisation du modèle du monde W (première phase de normalisation : World).
3. les distributions de probabilités $\mathcal{F}_{X=Y}$ et $\mathcal{F}_{X\neq Y}$ sont calculées respectivement à partir des rapports de vraisemblances (vraisemblances normalisées) clients et imposteurs.
4. suivant l'équation 10.2, la fonction de normalisation, appelée F_{WMap}, est apprise à partir des distributions $\mathcal{F}_{X=Y}$ et $\mathcal{F}_{X\neq Y}$ et des probabilités *a priori* $P(X = Y)$ et $P(X \neq Y)$ (deuxième phase de normalisation : MAP). Cette fonction de normalisation permet d'assigner une probabilité *a posteriori* à un rapport de vraisemblances donné.

Lors du processus de vérification, les étapes 1 et 2 sont reproduites (étapes 1' et 2' sur la figure 10.1). La fonction de normalisation F_{WMap} est ensuite appliquée sur chaque rapport de vraisemblances afin de produire des probabilités pour la phase de décision (étape 5 sur la figure 10.1).

3.1 Distributions $\mathcal{F}_{X=Y}$ et $\mathcal{F}_{X\neq Y}$ et modèle du monde

L'efficacité de la deuxième phase de normalisation (MAP) repose en grande partie sur l'estimation des distributions $\mathcal{F}_{X=Y}$ et $\mathcal{F}_{X\neq Y}$. Cette estimation est facilitée et rendue plus fiable par l'intervention du modèle du monde lors de la normalisation par rapport de vraisemblances (World). En effet, deux cas de figure peuvent se présenter :

- en présence d'une grande disparité entre les signaux d'apprentissage et de test du jeu de données, la normalisation par rapport de vraisemblances permet de réduire la variabilité des vraisemblances et de garantir des variances relativement faibles pour les distributions $\mathcal{F}_{X=Y}$ et $\mathcal{F}_{X\neq Y}$.

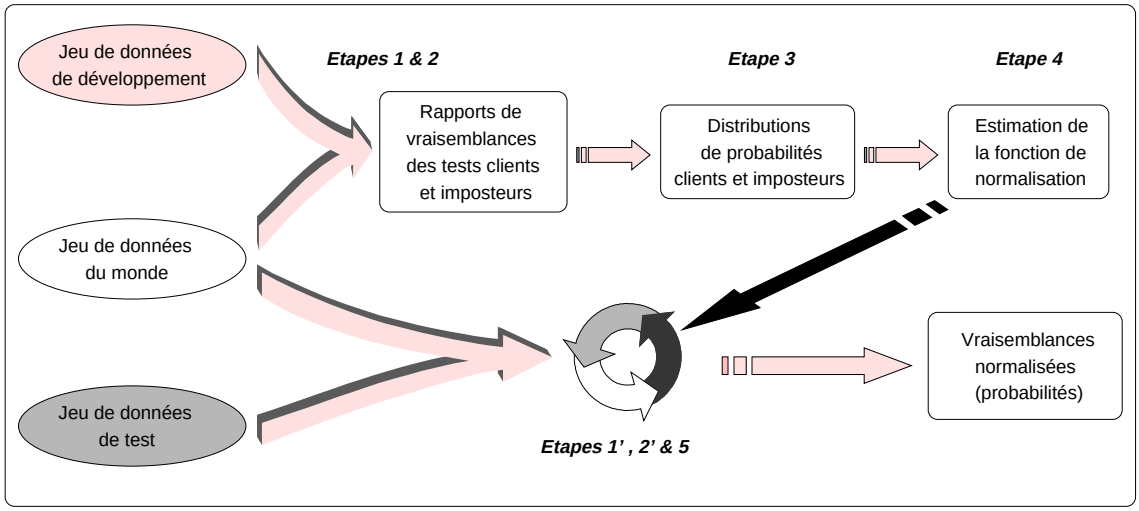

Figure 10.1: Mise en œuvre de l'approche World+MAP. Description des différentes étapes de l'approche World+MAP : apprentissage de la fonction de normalisation (étapes 1, 2, 3 et 4) et de son application (étapes 1', 2' et 5).

- en situation où la fonction de normalisation F_{WMap} doit être indépendante des locuteurs clients[2], le jeu de données nécessaire à l'apprentissage de F_{WMap} doit provenir d'une population différente des locuteurs clients. Dans ce cas de figure, l'application seule de la phase de normalisation MAP nécessiterait une grande quantité de signaux d'apprentissage afin d'assurer une bonne représentativité des locuteurs (clients et imposteurs) ainsi que des conditions d'utilisation du système au sein des distributions de probabilités $\mathcal{F}_{X=Y}$ et $\mathcal{F}_{X\neq Y}$. L'intervention du modèle du monde (dans le calcul des rapports de vraisemblances) vient pallier cette difficulté. En effet, la normalisation des vraisemblances permet de réduire considérablement la quantité de signaux d'apprentissage requise.

3.2 Approche bloc-segmentale

L'estimation des distributions de probabilités $\mathcal{F}_{X=Y}$ et $\mathcal{F}_{X\neq Y}$ est confrontée à un autre problème. La longueur des signaux de test, en nombre de trames, peut varier d'un test à l'autre. Or, il est bien connu que la variance des rapports de vraisemblances, calculée sur l'ensemble des trames d'un signal, est très dépendante de ce facteur. Une simple normalisation des valeurs des rapports de vraisemblances par le nombre de trames du test demeure souvent insuffisante. La conséquence directe est un problème d'homogénéité des valeurs de rapports de vraisemblances qui peut être préjudiciable pour l'estimation des distributions $\mathcal{F}_{X=Y}$ et $\mathcal{F}_{X\neq Y}$. Un moyen de contourner cette problématique est de considérer le calcul du rapport de vraisemblances sur une unité segmentale plus petite que l'ensemble des trames d'un signal et de taille constante.

Considérant la trame comme une unité segmentale trop réduite, nous préférons utiliser des blocs de trames d'une longueur suffisante pour conduire à des rapports de

[2]Ce cas peut se produire par exemple si très peu de signaux de parole sont disponibles par client. Il est par ailleurs imposé lors des campagnes d'évaluation NIST.

vraisemblances pertinents pour l'estimation des distributions. Ce procédé est appelé approche "bloc-segmentale" et est illustré en figure 10.2.

Considérons un signal de parole représenté par la séquence de vecteurs $\{y_t\}_{1\leq t\leq N}$. Soit B_i un des blocs résultant de l'approche bloc-segmentale. B_i est composé de T trames de signal de parole $\{y_t\}_{i\times T+1\leq t\leq (i+1)\times T}$. L'application du rapport de vraisemblances sur le bloc B_i se résume alors par la formulation suivante :

$$LR_{\mathcal{X}}(B_i) = f(LR_{\mathcal{X}}(y_{i\times T+1}), \ldots, LR_{\mathcal{X}}(y_{(i+1)\times T})) \tag{10.4}$$

où $LR_{\mathcal{X}}(\{y_t\}_{i\times T+1\leq t\leq (i+1)\times T})$ représente les rapports de vraisemblances sur chacune des trames composant le bloc et f est une fonction de fusion de ces rapports.

Dans cette thèse, la fonction de fusion est une simple moyenne géométrique (ou arithmétique dans le domaine logarithmique) :

$$f(LR_{\mathcal{X}}(y_{i\times T+1}), \ldots, LR_{\mathcal{X}}(y_{(i+1)\times T}) = \left(\prod_{j=1}^{T} LR_{\mathcal{X}}(y_{i\times T+j})\right)^{\frac{1}{T}} \tag{10.5}$$

D'autres fonctions f peuvent être envisagées dans ce cadre. En particulier, l'approche bloc-segmentale se prête parfaitement à des fonctions d'élagage de type "N-meilleurs" par exemple [Kittler et al., 1997], [Besacier, 1998] afin d'éliminer les trames les moins pertinentes.

Le calcul des rapports de vraisemblances au niveau d'un bloc ne change rien au processus décrit précédemment pour la mise en œuvre de l'approche World+MAP. À présent, la fonction de normalisation F_{WMap} est simplement apprise et appliquée bloc à bloc.

4 Un exemple d'application de World+MAP

Les différentes étapes de mise en œuvre de l'approche World+MAP couplée à l'approche bloc-segmentale sont à présent décrites dans un contexte expérimental afin d'étudier le comportement de la fonction F_{WMap}.

Des tests de vérification sont réalisés sur un ensemble de données de développement. Les vraisemblances sont estimées sur des blocs d'une longueur de 0,3 seconde soit 30 trames de signal de parole. Le choix de cette taille de blocs est arbitraire. Il est supposé qu'une séquence de 30 trames de signal de parole véhicule suffisamment d'informations pertinentes pour une estimation correcte des vraisemblances.

4.1 Jeu de données

L'ensemble de données est issu de la base de données Switchboard [Campbell et al., 1998]. Il comprend des signaux de parole enregistrés lors de conversations téléphoniques provenant de deux populations différentes.

Population pour l'estimation des modèles du monde

La première population est nécessaire pour l'apprentissage des modèles du monde, dépendants du genre et du type de combiné téléphonique. Cette population comprend 100

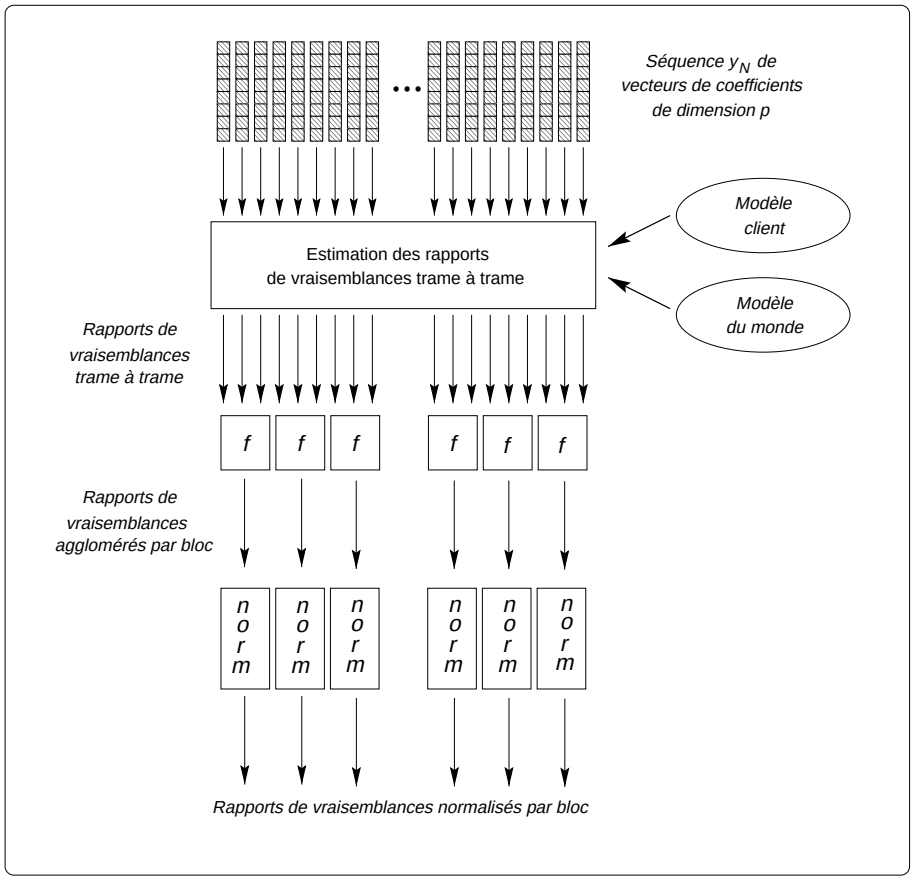

Figure 10.2: Approche bloc-segmentale. Illustration schématique des différentes étapes de l'approche bloc-segmentale.

hommes et 100 femmes. Des signaux de parole, d'une longueur de 30 secondes, ont été enregistrés pour chaque locuteur. Deux types de combinés téléphoniques sont utilisés dans la base de données Switchboard : electret ou carbon.

Population de développement

La seconde population est utilisée pour l'apprentissage de la fonction de normalisation F_{WMap}. Elle est constituée de 50 hommes et 50 femmes.
Deux sous-ensembles de signaux de parole sont construits pour tous les locuteurs. Le premier sous-ensemble est constitué de signaux de parole, enregistrés lors de deux sessions d'enrôlement, d'une durée totale de l'ordre de deux minutes. Il est destiné à l'apprentissage des modèles clients. Le deuxième sous-ensemble comprend des signaux de parole multi-sessions d'environ 30 secondes dédiés aux tests de vérification.

Cet ensemble de données, appelé ensemble de développement et noté *Dev*, est finalement constitué de 1190 tests clients (572 tests femmes et 618 tests hommes) et 8992 tests imposteurs (4506 tests femmes et 4486 tests hommes)[3].

[3]Cet ensemble de données ne comprend pas de tests inter-genres.

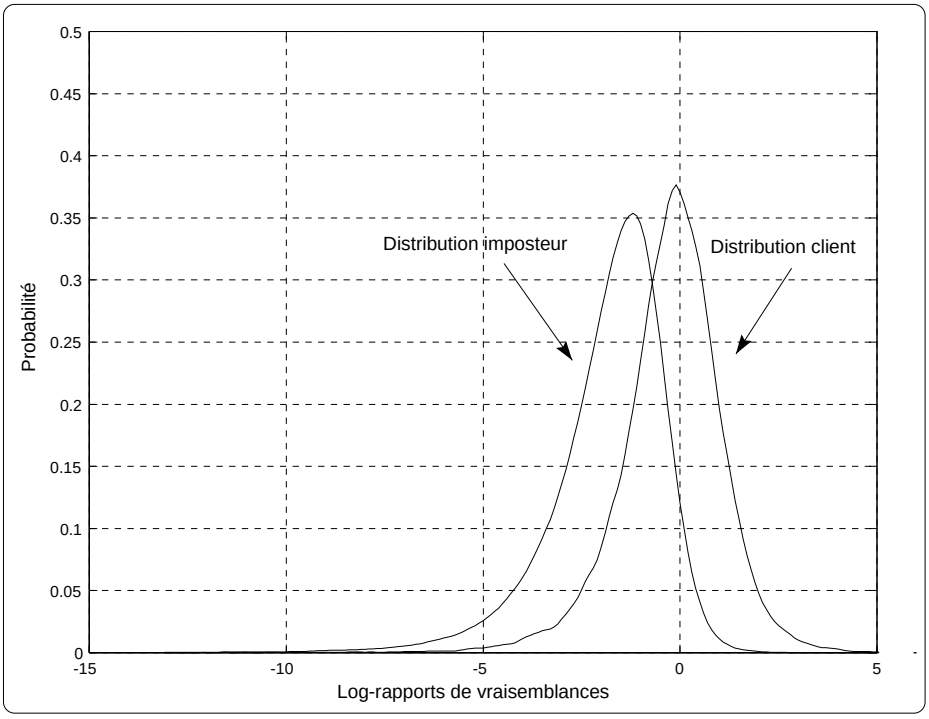

Figure 10.3: Dev : Distributions sans normalisation. Distributions des log-rapports de vraisemblances clients et imposteurs estimées sur l'ensemble Dev (femmes) (environ 17000 blocs clients et 135000 blocs imposteurs).

Les populations de développement et du monde sont entièrement séparées.

Les modèles clients ainsi que les modèles du monde reposent sur des mélanges de gaussiennes (GMM) à 16 composantes. Chaque composante est caractérisée par un vecteur moyen et une matrice de covariance pleine. L'estimation de ces modèles est réalisée au moyen de l'algorithme EM couplé à l'approche EMV.

4.2 Distributions des rapports de vraisemblances

Des tests de vérification sont menés sur l'ensemble de développement *Dev*, mettant en jeu les tests clients et imposteurs. Lors de chaque test, la vraisemblance du signal de test pour le modèle client considéré est estimée. Parallèlement, une estimation similaire est réalisée entre le signal de test et le modèle du monde correspondant.

La figure 10.3 illustre les distributions des log- rapports de vraisemblances issus des tests clients et imposteurs. Ces distributions concernent uniquement la population des femmes du jeu de données *Dev*.

Ces deux distributions présentent des moyennes très proches l'une de l'autre et des variances très importantes. Cette configuration conduit à un fort recouvrement entre les deux distributions, pénalisant les performances du système de VAL.

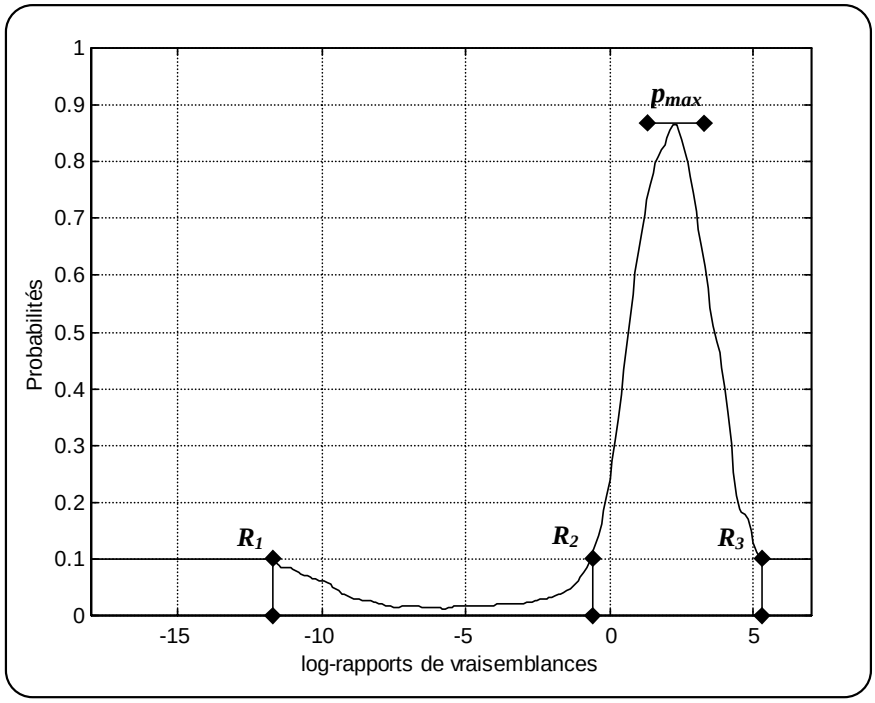

Figure 10.4: Dev : Fonction de normalisation F_{WMap}. Fonction de normalisation estimée à partir des distributions des log-rapports de vraisemblances clients et imposteurs calculés sur Dev (femmes) et des probabilités a priori $P(X = Y) = 0.1$ et $P(X \neq Y) = 0.9$ (environ 17000 blocs clients et 135000 blocs imposteurs).

4.3 Estimation de la fonction de normalisation F_{WMap}

Comme décrit en section 2, la normalisation World+MAP repose sur les distributions des rapports de vraisemblances $\mathcal{F}_{X=Y}$ et $\mathcal{F}_{X\neq Y}$ et sur les probabilités *a priori* $P(X = Y)$ et $P(X \neq Y)$. Ces probabilités *a priori* sont fixées en fonction des conditions de test attendues. Dans ce contexte expérimental, les probabilités $P(X = Y) = 0.1$ et $P(X \neq Y) = 0.9$ sont choisies en adéquation avec la composition du corpus de test.

La figure 10.4 présente la fonction de normalisation estimée à partir des distributions des log-rapports de vraisemblances clients et imposteurs obtenues sur la population féminine de l'ensemble de développement *Dev* (figure 10.3) et des probabilités *a priori*. Cette fonction de normalisation est obtenue par application de l'équation 10.2, qui permet d'attribuer à un rapport de vraisemblances donné une probabilité d'appartenance à un test client.

Trois parties, délimitées par les valeurs (en abscisse) de log-rapports de vraisemblances R_1, R_2 et R_3 sur la figure 10.4, caractérisent la fonction de normalisation.

Intervalle $[R_2; R_3]$

Des probabilités supérieures à la probabilité *a priori* $P(X = Y)$ sont assignées aux log-rapports de vraisemblances compris dans l'intervalle $[R_2; R_3]$. Cette première partie désigne les accès clients.

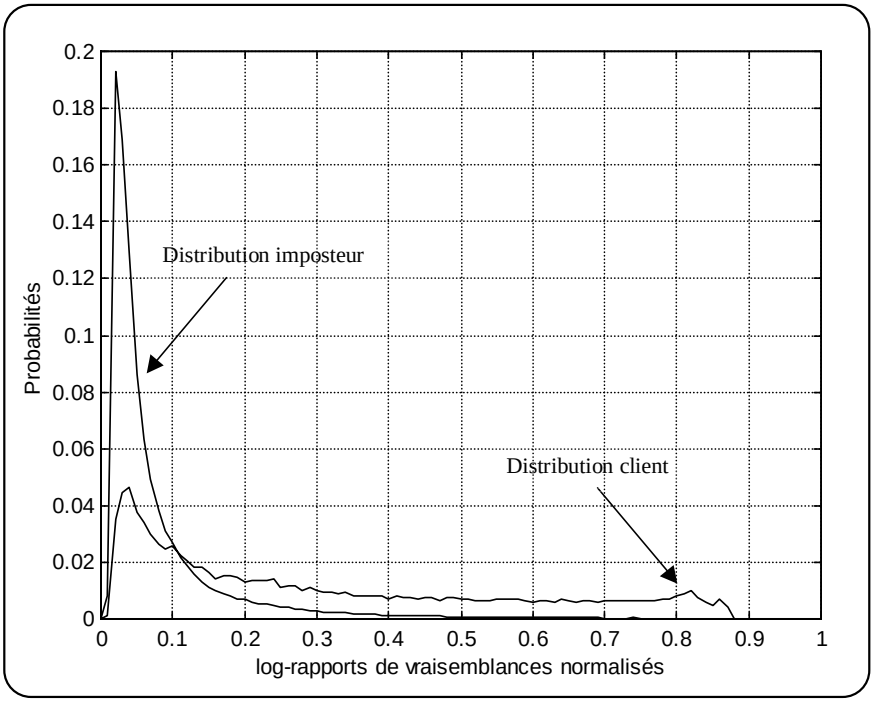

Figure 10.5: Dev : distributions après normalisation. Distributions des log-rapports de vraisemblances normalisés (probabilités) clients et imposteurs calculées sur Dev (femmes) (environ 17000 blocs clients et 135000 blocs imposteurs).

Intervalle $[R_1; R_2[$

Des probabilités inférieures à la probabilité *a priori* $P(X = Y)$ sont assignées aux log-rapports de vraisemblances compris dans l'intervalle $[R_1; R_2]$. Cette deuxième partie désigne les accès imposteurs.

Intervalles $]\inf; R_1[$ **et** $]R_3; \inf[$

La probabilité *a priori* $P(X = Y)$ est attribuée aux log-rapports de vraisemblances situés dans les intervalles $]\inf; R_1[$ et $]R_3; \inf[$. Il s'agit de valeurs jugées aberrantes sur le plan statistique (valeurs très peu ou pas représentées).

4.4 Distributions des log-rapports de vraisemblances normalisés

La figure 10.5 montre les distributions des log-rapports de vraisemblances, issues des mesures clients et imposteurs du jeu de données *Dev* (femmes), après application de la fonction de normalisation estimée précédemment (figure 10.4).

Plusieurs remarques émergent de l'observation de ces distributions :

- Les scores normalisés provenant des tests imposteurs sont en majeure partie concentrés dans l'intervalle $[0; P(X = Y)]$ en adéquation avec la fonction de normalisation.
- La distribution des log-rapports de vraisemblances normalisés (probabilités) issus des tests clients peut être assimilée à la combinaison de deux distributions.

La première distribution concerne les probabilités comprises dans l'intervalle $[P(X = Y); p_{max}]$. Ces probabilités *a posteriori* sont effectivement attribuées à des tests clients avec un facteur de confiance directement lié à leur valeur. Une probabilité proche de p_{max} se voit accordé un facteur de confiance très élevé. À l'opposé, le facteur de confiance associé à une probabilité proche de $P(X = Y)$ est faible.
La deuxième distribution représente des probabilités *a posteriori* variant dans l'intervalle $[0; P(X = Y)]$. Selon la fonction de normalisation, ces probabilités font référence à des tests imposteurs. Cette erreur d'appréciation de la fonction de normalisation – assignation de log-rapports de vraisemblances issus de tests clients à des probabilités caractérisant des tests imposteurs – peut s'expliquer par le fait que l'information spécifique du locuteur n'est pas également répartie dans le domaine temporel [Besacier et al., 2000a]. Certains blocs, peu ou pas informatifs, sont par conséquent assignés à de faibles probabilités, inférieures à la probabilité *a priori* $P(X = Y)$.

Les distributions des log-rapports de vraisemblances normalisés, i.e. les distributions des probabilités, montrent, comme la fonction de normalisation, une manière simple de discriminer les tests clients des tests imposteurs.

4.5 Mesure de la qualité de la fonction F_{WMap}

Le rôle de la fonction de normalisation F_{WMap} est d'assigner à un rapport de vraisemblances une probabilité *a posteriori* pour que ce rapport soit issu d'un test client. Attribuer une probabilité de 0,2 à un rapport de vraisemblances signifie que, sur 10 tests de vérification, ce rapport est issu 2 fois d'un test client et 8 fois d'un test imposteur.

Un moyen de mesurer la précision de la fonction F_{WMap} est de vérifier si la signification d'une probabilité se révèle exacte sur le jeu de données sur lequel la fonction est appliquée. Pour exemple, sur le jeu de données *Dev*, il suffit de vérifier que pour toute probabilité (rapport de vraisemblances normalisé) de valeur 0,2 fournie par F_{WMap}, 20% de ces valeurs sont effectivement attribuées à des tests clients et 80% à des tests d'imposteurs.

Dans cette optique, une probabilité $p_{Dev}(p_{F_{WMap}}|X = Y)$ est calculée sur le jeu de données *Dev* et comparée à la probabilité $p_{F_{WMap}}$ fournie par la fonction de normalisation F_{WMap}. Cette probabilité $p_{Dev}(p_{F_{WMap}}|X = Y)$ est calculée en comptant le nombre d'attributions, sur le jeu de développement *Dev*, de la probabilité $p_{F_{WMap}}$ à un test client – $C_{client}(p_{F_{WMap}})$ – ou à un test imposteur – $C_{imp}(p_{F_{WMap}})$. Cette probabilité se définit par l'expression :

$$p_{Dev}(p_{F_{WMap}}|X = Y) = \frac{C_{client}(p_{F_{WMap}})}{C_{client}(p_{F_{WMap}}) + C_{imp}(p_{F_{WMap}})} \tag{10.6}$$

La figure 10.6 fournit le résultat de l'analyse du jeu de données *Dev*. Les probabilités $p_{Dev}(p_{F_{WMap}}|X = Y)$ (axe des ordonnées) sont données ici en fonction des probabilités $p_{F_{WMap}}$ (axe des abscisses). La droite $x = y$ représente la précision optimale que peut atteindre la fonction de normalisation F_{WMap}.

La proximité des points autour de la droite $x = y$ démontre une précision très satisfaisante de la fonction de normalisation F_{WMap}. Notamment, on peut remarquer une progression des probabilités $p_{Dev}(p_{F_{WMap}}|X = Y)$ ostensiblement linéaire. La perte de précision, bien que peu sensible, peut être raisonnablement imputée à l'estimation de la

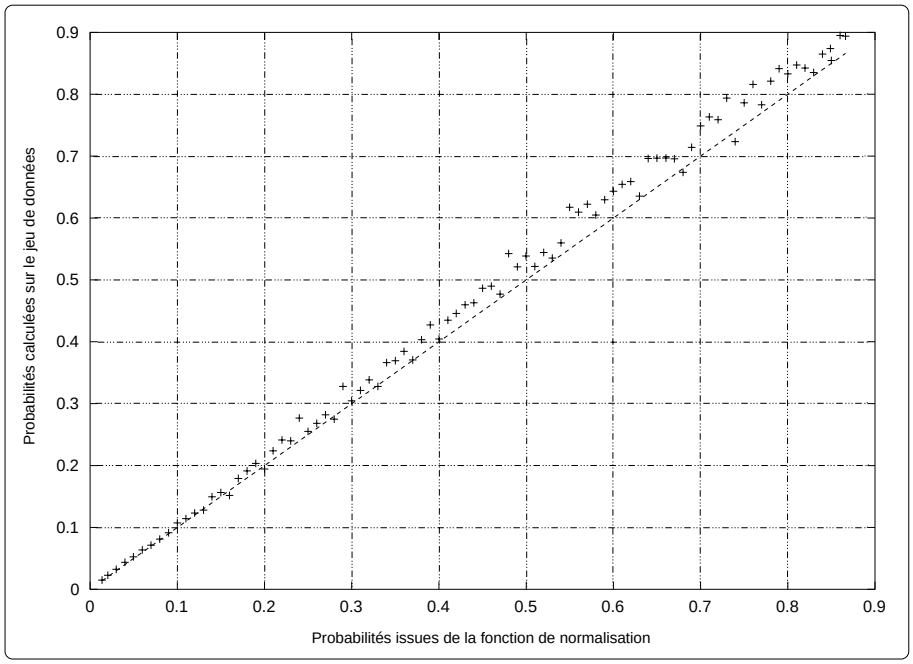

Figure 10.6: Dev : précision de la fonction de normalisation F_{WMap}. Comparaison des probabilités fournies par la fonction de normalisation F_{WMap} estimée sur Dev (femmes) et des probabilités réelles calculées sur Dev (femmes) (17000 blocs clients et 135000 blocs imposteurs).

fonction F_{WMap} à partir des distributions de probabilités $\mathcal{F}_{X=Y}$ et $\mathcal{F}_{X\neq Y}$ (qui n'est pas triviale).

Néanmoins, ce haut niveau de précision est ici prévisible par le fait que la fonction de normalisation est apprise et appliquée sur le même jeu de données (Dev). En effet, dans ce cas de figure, la fonction de normalisation peut être considérée comme dépendante des locuteurs et facilite la normalisation des rapports de vraisemblances.

4.6 Évaluation de la fonction F_{WMap}

Les sections précédentes montrent que la fonction de normalisation F_{WMap} est efficace pour discriminer les tests clients et imposteurs. La comparaison des probabilités issues de cette fonction avec les probabilités réelles calculées sur le jeu de données Dev souligne le haut niveau de précision de cette fonction.

Néanmoins, il est important de noter que la fonction de normalisation est apprise et testée sur le même ensemble de données (Dev). Il serait à présent intéressant d'observer le comportement de la fonction de normalisation F_{WMap} sur un jeu de données séparé, i.e. composé de locuteurs différents de ceux du jeu de données Dev.

4.6.1 Jeu de données d'évaluation : Eva

Le jeu de données d'évaluation Eva est d'une composition similaire à celle du jeu de données Dev. Seule la population de locuteurs est différente.

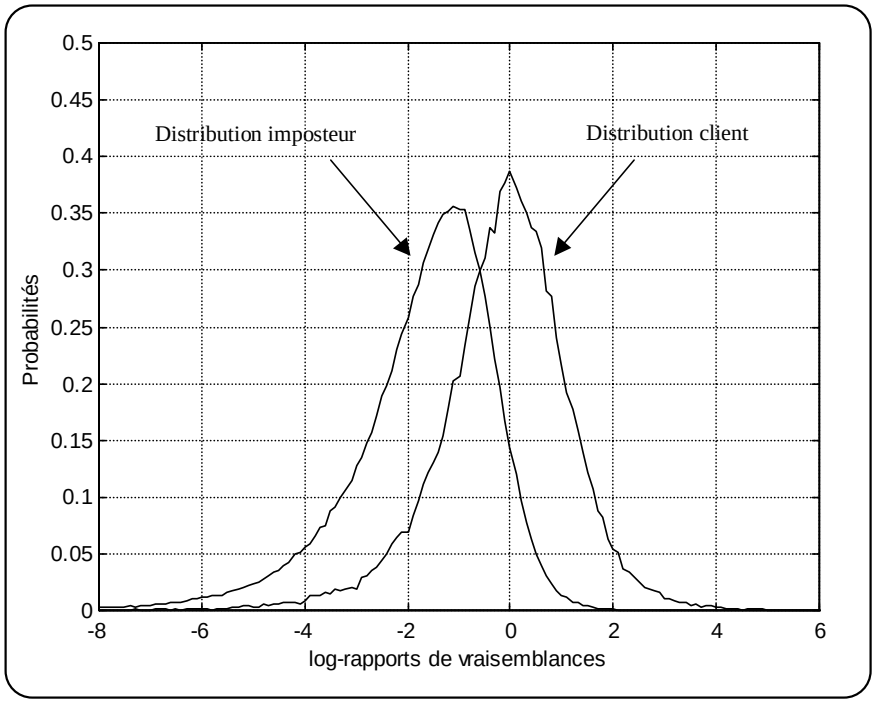

Figure 10.7: Eva : distributions sans normalisation. Distributions des log-rapports de vraisemblances clients et imposteurs calculées sur Eva (femmes) (environ 12000 blocs clients et 135000 blocs imposteurs).

Elle se compose de 100 locuteurs, également répartis en genre. Les signaux d'apprentissage, enregistrés lors de deux sessions d'enrôlement, sont d'une durée de 2 minutes et les signaux de test, multi-sessions, d'une durée de 30 secondes. Ce jeu de données d'évaluation comprend 948 tests clients (403 femmes et 545 hommes) et 8965 tests imposteurs (4475 femmes et 4490 hommes)[4].

NB : il est à noter que les modèles du monde (dépendants du genre et du type de combiné téléphonique) sont identiques à ceux utilisés précédemment.

4.6.2 Application de F_{WMap} sur Eva

La fonction de normalisation F_{WMap}, estimée sur le jeu de données Dev (figure 10.4), est à présent appliquée pour normaliser les log-rapports de vraisemblances produits lors des tests clients et imposteurs du jeu de données d'évaluation Eva. Les distributions de ces log-rapports de vraisemblances avant application de la fonction F_{WMap} sont fournies en figure 10.7 (population féminine du jeu de données Eva).

En outre, la figure 10.8 reporte les distributions des log-rapports de vraisemblances normalisés, i.e. les distributions des probabilités après application de la fonction F_{WMap}.

Nous pouvons constater que ces distributions présentent des caractéristiques comparables à celles observées sur les distributions issues du jeu de données de développement Dev (figure 10.5). L'application de la fonction de normalisation F_{WMap} sur un jeu de

[4]Cet ensemble de données ne comprend pas de tests inter-genres.

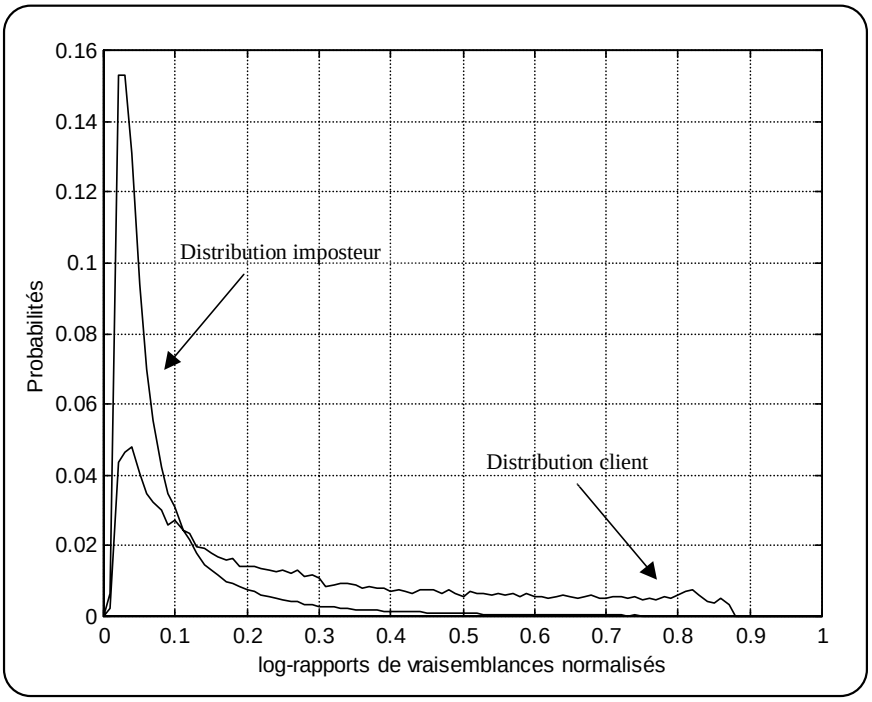

Figure 10.8: Eva : distributions après normalisation. Distributions des log-rapports de vraisemblances normalisés (probabilités) clients et imposteurs calculées sur Eva (femmes) après application de la fonction de normalisation F_{WMap} estimée sur Dev (femmes) (environ 12000 blocs clients et 135000 blocs imposteurs).

données différent de celui utilisé pour son apprentissage ne semble pas perturber le comportement de la normalisation.

4.6.3 Précision de la fonction F_{WMap} avec des populations différentes

La précision de la fonction F_{WMap} est à présent mesurée sur le jeu de données d'évaluation *Eva*. Il s'agit, ici, de calculer les probabilités $p_{Eva}(p_{F_{WMap}}|X = Y)$ pour chaque probabilité $p_{F_{WMap}}$ assignée par la fonction de normalisation à un rapport de vraisemblances calculés sur le jeu de données *Eva*.

Le graphe, présenté en figure 10.9, reporte les probabilités $p_{Eva}(p_{F_{WMap}}|X = Y)$ (axe des ordonnées) en fonction des probabilités $p_{F_{WMap}}$ (axe des abscisses).

La précision de la fonction F_{WMap} est dégradée par rapport à celle obtenue sur le jeu de données *Dev* (figure 10.6). Une différence de 5 à 10 % est observée entre les probabilités fournies par la fonction F_{WMap} et celles relatives au jeu de données *Eva*. Néanmoins, la progression des probabilités $p_{Eva}(p_{F_{WMap}}|X = Y)$ conserve une linéarité manifeste.

Cette dégradation est justifiée par le fait que la fonction F_{WMap} est estimée et appliquée sur deux jeux de données différents : *Dev* et *Eva* i.e. sur deux populations de locuteurs totalement séparées. La différence de 5 à 10 % peut être considérée comme une perte

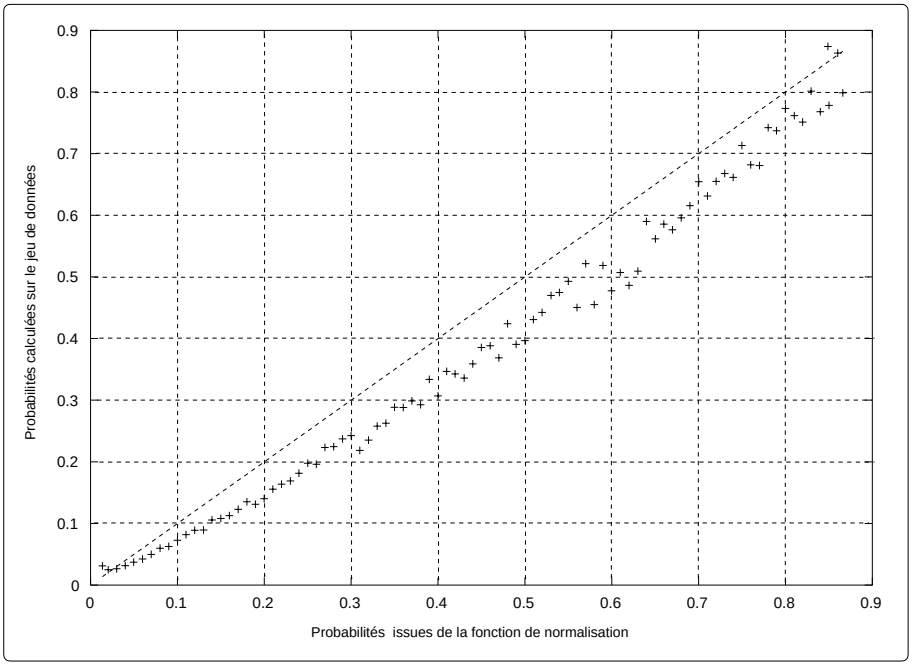

Figure 10.9: Eva : Précision de la fonction de normalisation F_{WMap}. Comparaison des probabilités fournies par la fonction de normalisation F_{WMap} estimée sur Dev (femmes) et des probabilités réelles calculées sur Eva (femmes) (12000 blocs clients et 135000 blocs imposteurs).

raisonnable dans ces conditions. Par ailleurs, une perte de précision peut de nouveau incomber au processus d'apprentissage de la fonction F_{WMap} à partir des distributions $\mathcal{F}_{X=Y}$ et $\mathcal{F}_{X\neq Y}$.

5 Premières conclusions

Au cours de ce chapitre, nous avons montré de quelle manière, grâce à l'approche World+MAP, les vraisemblances entre signaux de test et modèles clients peuvent être projetées dans un espace probabiliste, conférant une interprétation directe au seuil de décision.
En outre, l'étude de la fonction de normalisation F_{WMap} et son application sur les jeux de données *Dev* et *Eva* ont permis de mettre en évidence la capacité de cette fonction à discriminer les accès clients des accès imposteurs.

Néanmoins, il reste à présent à démontrer que l'approche World+MAP permet d'assurer de bonnes performances à un système de VAL ou, du moins, des performances comparables aux techniques de normalisation issues de l'état de l'art.

Chapitre 11

Evaluation de l'approche World+MAP

Le comportement de l'approche World+MAP est à présent évalué au sein du système de VAL. Il s'agit de montrer, au travers de résultats d'expériences, que la projection des vraisemblances dans un espace probabiliste est conciliable avec le maintien des performances du système.

Les expériences présentées dans ce chapitre ont pour but d'évaluer l'effet de l'approche World+MAP au sein d'un système de VAL. En première expérience, nous étudions l'impact d'une fonction de normalisation apprise et appliquée sur deux jeux de données différents. L'approche World+MAP est ensuite comparée à différentes techniques de normalisation. Finalement, l'intérêt de l'approche World+MAP dans une architecture multi-reconnaisseur est exploré.

La majorité des résultats présentés dans ce chapitre sont donnés sous forme de courbes DET (voir le chapitre 1 pour plus de détails sur ce type de courbes).

1 Contexte expérimental

1.1 Bases de données

Les diverses expériences présentées ici sont conduites sur les jeux de données introduits au chapitre précédent : une population de locuteurs pour la construction des modèles du monde dépendants du genre et du type de combiné téléphonique ainsi que les jeux de données de développement (*Dev*) et d'évaluation (*Eva*).

Pour certaines expériences mettant en jeu notamment les techniques classiques de normalisation de type znorm ou hnorm, un jeu de données supplémentaire, entièrement disjoint des précédents, est nécessaire. Ce jeu de données, noté *Imp*, se compose de signaux de parole issus d'une population de 100 imposteurs (50 femmes et 50 hommes). Ces signaux, d'une durée moyenne de 30 secondes, se répartissent équitablement en deux sous-ensembles suivant le type de combiné téléphonique, electret ou carbon.

1.2 Système de VAL

Le système de VAL, utilisé dans ces expériences, repose sur une paramétrisation du signal de parole en vecteurs de 16 coefficients cepstraux, issus d'une analyse en banc de filtres. Ces vecteurs cepstraux sont normalisés par application du retrait de la moyenne cepstrale (CMS).

Les modèles clients et du monde sont des mixtures de gaussiennes à 16 composantes. Chaque composante est caractérisée par un vecteur moyen et une matrice de covariance pleine.

Basé sur l'approche bloc-segmentale (voir détails en section 3 du chapitre précédent), le système de VAL fournit un score (probabilité dans le cas de l'approche World+MAP, log-rapport de vraisemblances dans le cas d'une normalisation par le modèle du monde uniquement) pour chaque bloc – composé de 30 trames (0,3 seconde) – issu du signal de test. Ces scores sont ensuite fusionnés, par simple moyenne arithmétique, pour donner un score unique de décision.

2 Intérêt d'une normalisation des vraisemblances

Cette première expérience est destinée à mettre en évidence l'intérêt, en termes de performances, d'une normalisation des vraisemblances. Dans cette optique, le système de

VAL est testé avec et sans normalisation des vraisemblances[1]. Dans ce contexte, nous avons choisi d'utiliser une technique de normalisation "état de l'art" : normalisation par rapport de vraisemblances (modèles du monde dépendants du genre et du type de combiné téléphonique).

La figure 11.1 reporte les courbes DET obtenues sans (No Norm.) et avec normalisation (Rap. Vrais.) des vraisemblances, sur les jeux de données *Dev* (figure du haut) et *Eva* (figure du bas). Une très nette amélioration des performances est observée avec la normalisation des vraisemblances et ce pour les deux jeux de données. Un gain absolu d'environ 10 % de taux d'égale erreur est constaté dans les deux cas.

Ces résultats démontrent sans équivoque l'intérêt d'une normalisation des vraisemblances afin d'améliorer les performances des systèmes de VAL.

3 Approche World+MAP et modèle du monde

3.1 Approche MAP vs. World+MAP

Les expériences proposées ici ont pour but de mesurer l'apport du rapport de vraisemblances (World) pour la deuxième phase de normalisation (MAP) dans le cadre de l'approche World+MAP.

Des expériences comparatives sont conduites, opposant l'approche MAP (appliquée directement sur les log-vraisemblances) à l'approche World+MAP. Ces expériences sont réalisées sur les jeux de données *Dev* et *Eva*. Concernant l'approche World+MAP, la fonction de normalisation F_{WMap}, présentée au chapitre précédent, est employée lors de la normalisation des log-rapports de vraisemblances. Nous rappelons que cette fonction de normalisation est apprise sur le jeu de données *Dev*. Pour l'approche MAP, des distributions clients et imposteurs de log-vraisemblances sont estimées sur le jeu de données *Dev* et utilisées pour l'apprentissage d'une nouvelle fonction de normalisation.

La figure 11.2 présente les courbes DET obtenues, sur les jeux de données *Dev* et *Eva*, lors de l'utilisation de l'approche MAP ou World+MAP.

L'approche World+MAP obtient de bien meilleurs résultats que l'approche MAP utilisée seule, quel que soit le jeu de données considéré. Cette constatation souligne l'intérêt d'une réduction préalable de la variabilité des vraisemblances clients et imposteurs, apportée par le modèle du monde, pour l'apprentissage et l'utilisation de la fonction de normalisation F_{WMap}.

3.2 Approche World vs. World+MAP

Comme le souligne le chapitre précédent, le principal avantage de l'approche World+MAP est de "projeter" les vraisemblances, produites par le système de VAL, ainsi que le seuil de décision, dans un espace probabiliste. Il est cependant nécessaire de montrer que cet avantage de l'approche World+MAP est compatible avec le maintien des

[1] En vérité, une normalisation est déjà appliquée au travers de la CMS. Néanmoins, nous estimons que ses effets sur le signal ne peuvent pas compromettre les résultats des expériences présentées dans ce chapitre.

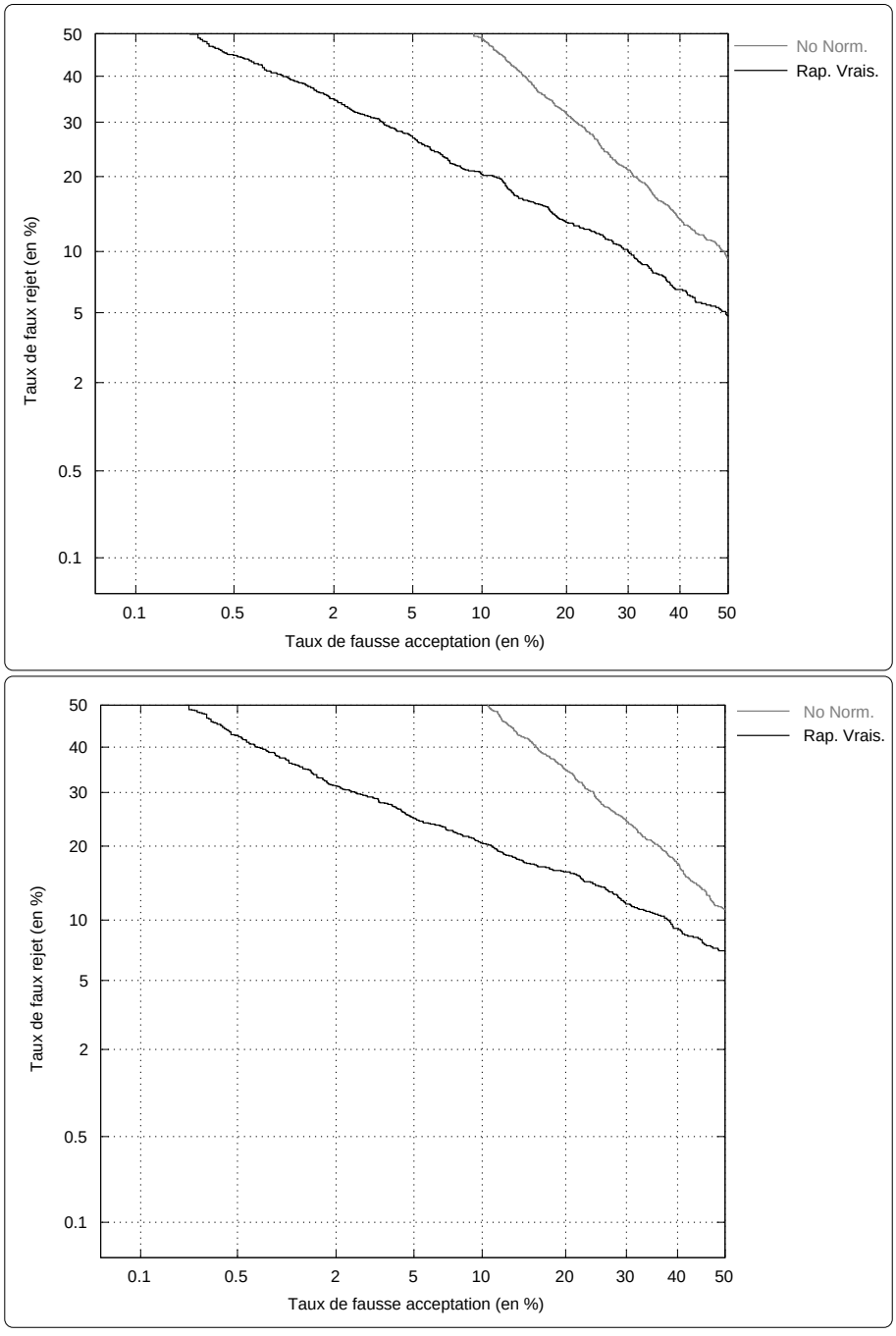

Figure 11.1: Intérêt d'une normalisation des vraisemblances. Comparaison des courbes DET obtenues, sur Dev (figure du haut) et Eva (figure du bas) avec et sans normalisation des vraisemblances produites par le système de VAL (Tâche de vérification sur Switchboard – Dev – 1190 tests clients et 8992 tests imposteurs – Eva – 948 tests clients et 8965 tests imposteurs).

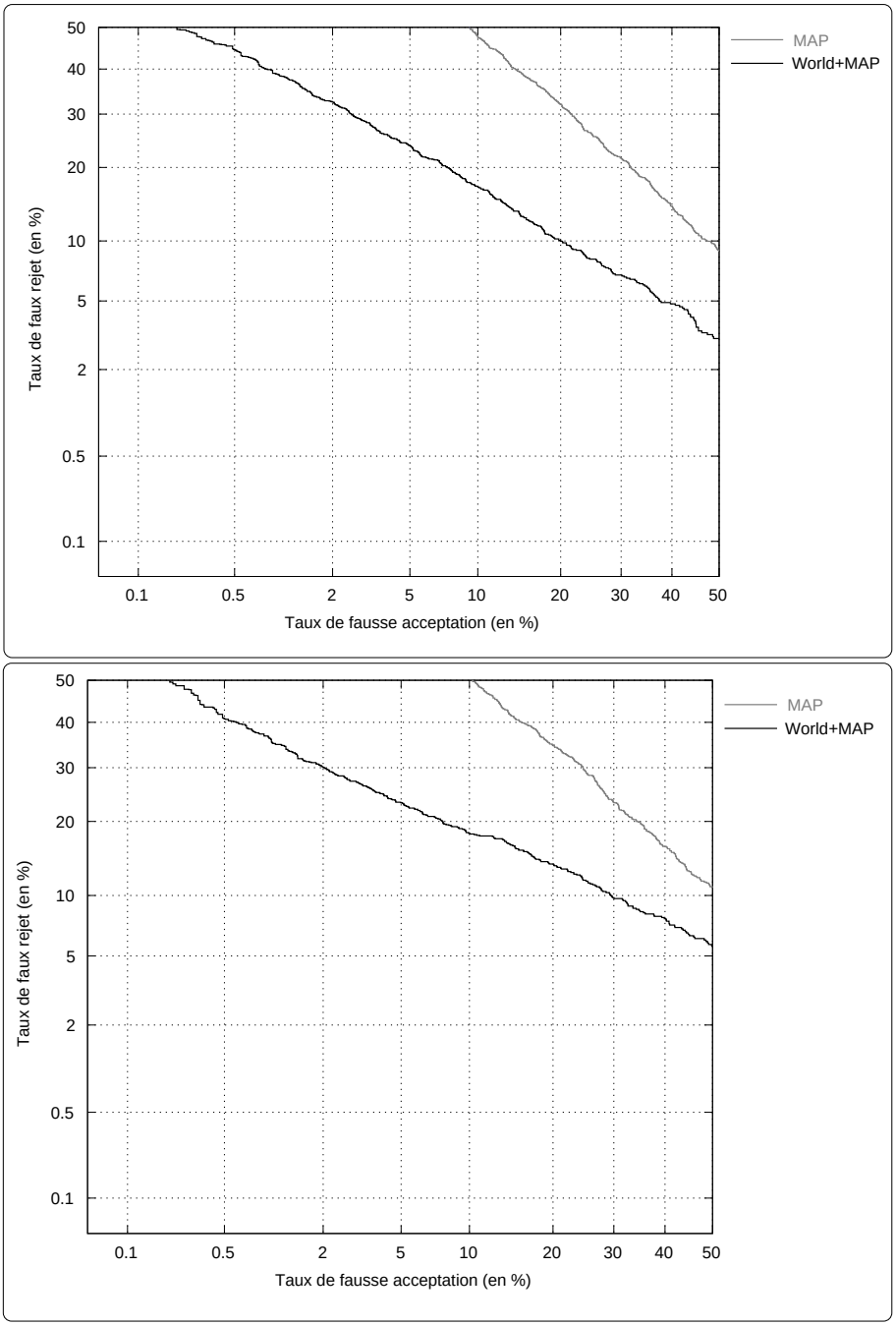

Figure 11.2: Apport du modèle du monde dans l'approche World+MAP. Comparaison des courbes DET obtenues, sur Dev (figure du haut) et Eva (figure du bas), en utilisant soit l'approche MAP, soit l'approche World+MAP (Tâche de vérification sur Switchboard – Dev – 1190 tests clients et 8992 tests imposteurs – Eva – 948 tests clients et 8965 tests imposteurs).

performances d'un système de VAL. En d'autres termes, il faut vérifier que l'approche World+MAP ne dégrade pas les performances du système.

Dans cette voie, nous avons comparé les performances de la technique "état de l'art" du rapport de vraisemblances (basé sur les modèles du monde dépendants du genre et du type de combiné téléphonique) à l'approche World+MAP. Les courbes DET issues de cette comparaison sont fournies en figure 11.3 pour les deux jeux de données *Dev* et *Eva*.

Au regard de ces courbes, nous pouvons juger d'un gain significatif, en termes de performances, de l'approche World+MAP sur la normalisation par rapport de vraisemblances (Rap. Vrais.). Le deuxième point intéressant relève des résultats obtenus sur le jeu d'évaluation *Eva* (figure du bas). Nous pouvons constater que l'utilisation de la fonction de normalisation F_{WMap}, apprise sur *Dev* et appliquée sur *Eva*, ne dégrade pas les performances du système et conduit, au contraire, à une légère amélioration des performances par comparaison avec la technique du rapport de vraisemblances.

Ces résultats montrent que le processus de "projection" des vraisemblances dans un espace probabiliste s'intègre pleinement au sein d'un système de VAL, améliorant ses performances par rapport à l'utilisation seul d'un modèle du monde. La généralisation de la fonction de normalisation, apprise et appliquée sur des jeux de données séparés, est également vérifiée par les résultats obtenus sur *Eva*.

4 Comparaison de différentes normalisations

L'approche World+MAP est à présent comparée à différentes techniques de normalisation issues de la littérature :

- Normalisation par rapport de vraisemblances (expérience présentée dans la section précédente) ;
- Normalisation par rapport de vraisemblances couplée à la normalisation Znorm (détaillée dans le chapitre 9 section 3.1.1) ;
- Normalisation par rapport de vraisemblances couplée à la normalisation Hnorm (détaillée dans le chapitre 9 section 4).

L'estimation des paramètres μ et σ, nécessaire à l'application des normalisations Znorm et Hnorm, est réalisée sur le jeu de données *Imp*. Ce jeu de données, décrit en section 1.1, est entièrement séparé des jeux de données *Dev* et *Eva*.

Les expériences, mettant en jeu les différentes techniques de normalisation, sont conduites sur les jeux de données *Dev* et *Eva*. La figure 11.4 regroupe les différentes courbes DET, obtenues pour chacune d'elles.

Les courbes DET provenant du jeu de données *Dev* (figure du haut) montrent que l'approche World+MAP parvient à des résultats significativement meilleurs que la normalisation par rapport de vraisemblances seule ou couplée à la technique Znorm. De plus, l'approche World+MAP obtient des résultats comparables à ceux du couple rapport de vraisemblances – technique Hnorm.

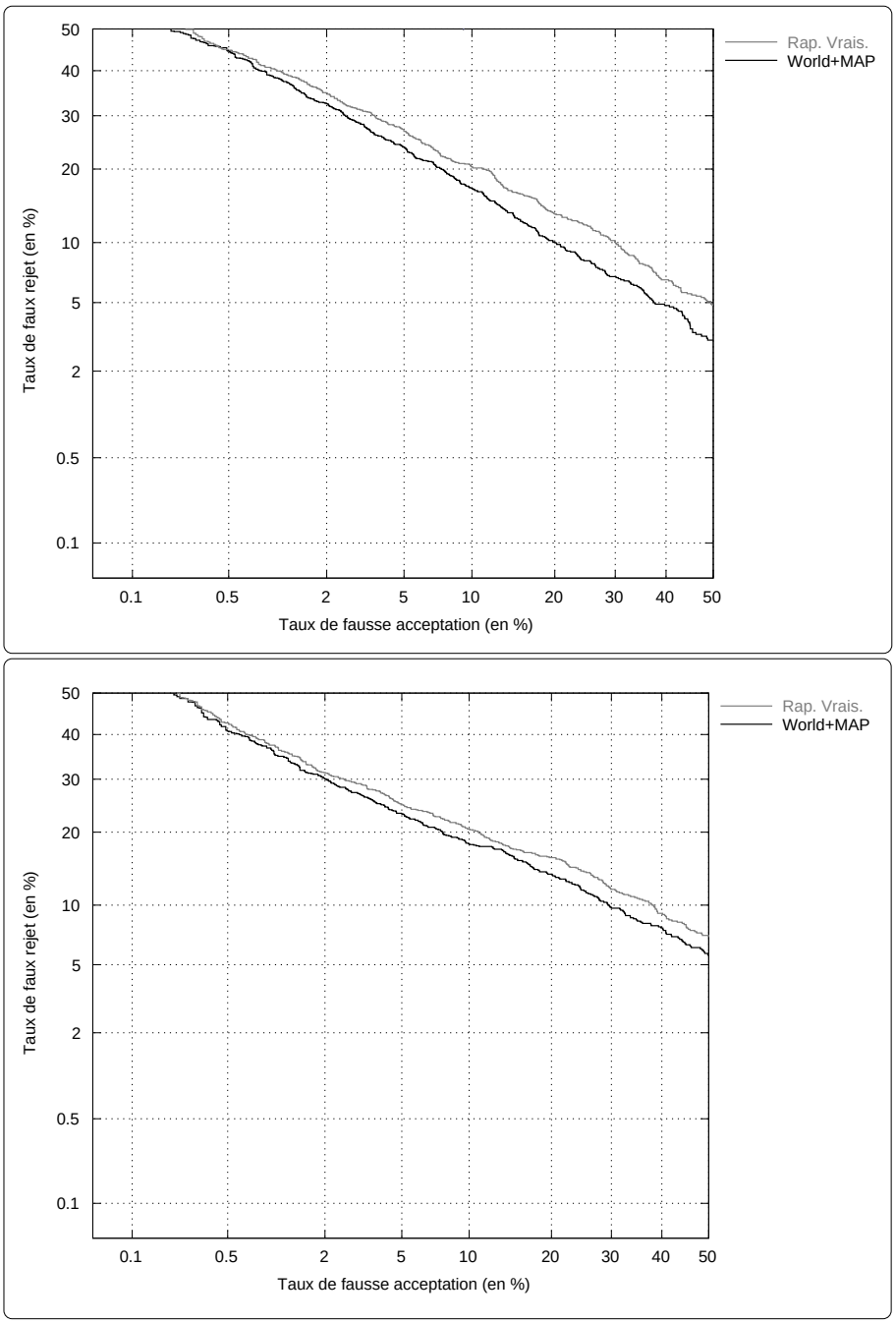

Figure 11.3: Rapport de vraisemblances vs. World+MAP. Comparaison des courbes DET obtenues, sur Dev (figure du haut) et Eva (figure du bas) en utilisant soit la technique du rapport de vraisemblances (basé sur les modèles du monde dépendants du genre et du type de combiné téléphonique) soit l'approche World+MAP (Tâche de vérification sur Switchboard – Dev – 1190 tests clients et 8992 tests imposteurs – Eva – 948 tests clients et 8965 tests imposteurs).

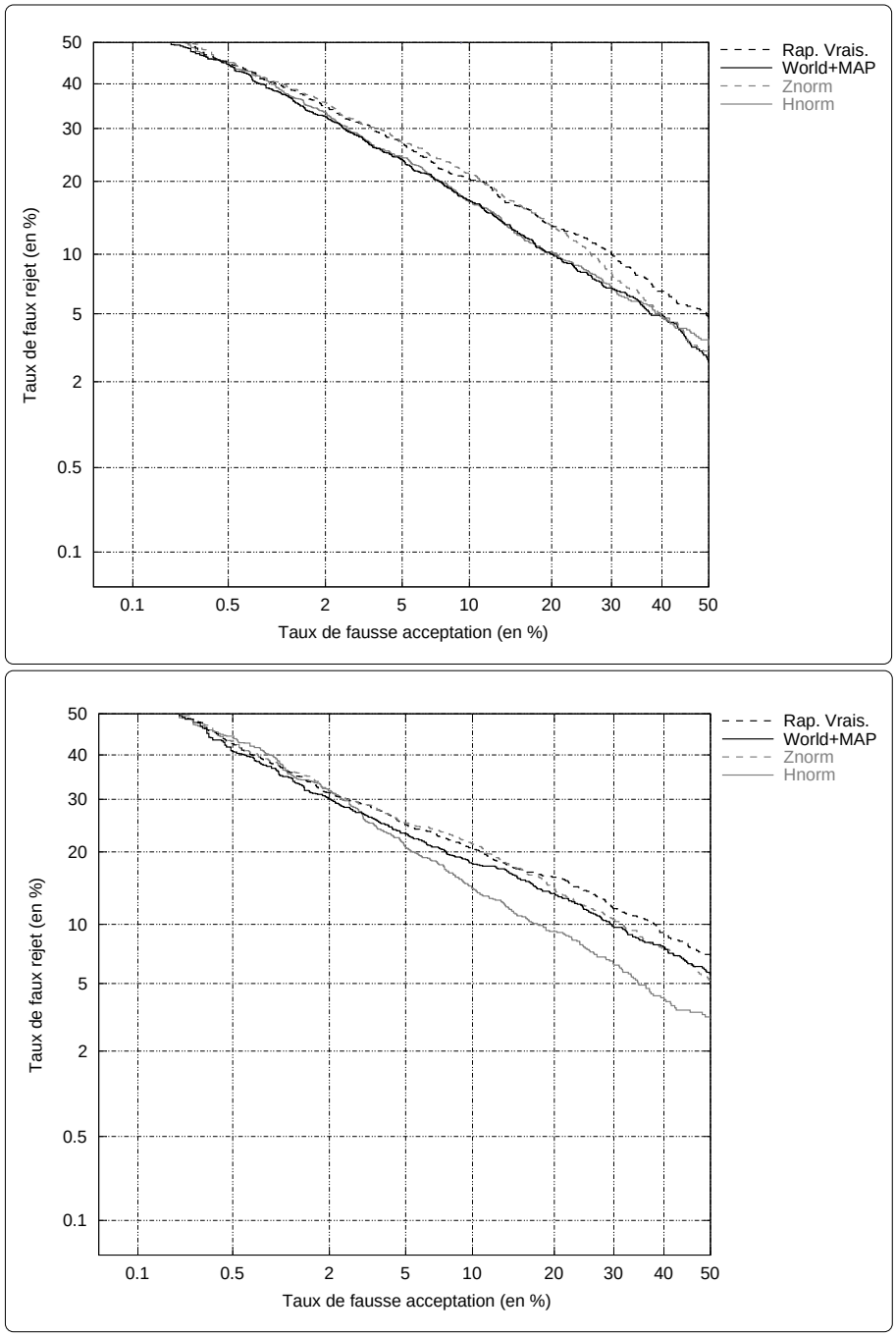

Figure 11.4: Comparaison de différentes techniques de normalisation. Courbes DET obtenues, sur Dev (figure du haut) et Eva (figure du bas) suivant différentes techniques de normalisation : normalisation par rapport de vraisemblances, rapport de vraisemblances couplé à Znorm, couplé à Hnorm et finalement l'approche World+MAP (Tâche de vérification sur Switchboard – Dev – 1190 tests clients et 8992 tests imposteurs – Eva – 948 tests clients et 8965 tests imposteurs).

	Taux d'égale erreur : EER (en %)	
Normalisation	Jeu de données *Dev*	Jeu de données *Eva*
Rap. Vrais.	15,8	16,7
World+MAP	13,7	15,6
Znorm	15,9	16,3
Hnorm	13,7	12,8

Tableau 11.1: Taux d'EER de différentes techniques de normalisation. Comparaison, en termes de taux d'égale erreur obtenus sur Dev et Eva, suivant différentes techniques de normalisation : normalisation par rapport de vraisemblances, rapport de vraisemblances couplé à Znorm, couplé à Hnorm et finalement l'approche World+MAP (Tâche de vérification sur Switchboard – Dev – 1190 tests clients et 8992 tests imposteurs – Eva – 948 tests clients et 8965 tests imposteurs).

En revanche, les résultats relatifs au jeu de données *Eva* soulignent une efficacité moindre de la part de l'approche World+MAP. En effet, l'approche World+MAP obtient de meilleures performances que la normalisation par rapport de vraisemblances seule ou couplée à la technique Znorm. Néanmoins, la différence est moins marquée que celle observée sur le jeu de données *Dev*. D'autre part, le couple – normalisation par rapport de vraisemblances et technique Hnorm – présente un net avantage (sur certaines portions des courbes DET), notamment en termes de taux d'égale erreur (reportés dans le tableau 11.1), par rapport à l'approche World+MAP.

La comparaison des résultats obtenus sur les jeux de données *Dev* et *Eva* indique une différence de comportement de l'approche World+MAP. Il semble raisonnable d'attribuer cette différence à la fonction de normalisation F_{WMap}. La fonction de normalisation est apprise sur le jeu de données *Dev*. Elle intègre, par conséquent, intrinsèquement les caractéristiques de ce jeu de données (caractéristiques des locuteurs, des signaux de tests, des conditions d'utilisation du système). Dans ces conditions, l'application de la fonction de normalisation conduit à de très bonnes performances sur le jeu de données *Dev*, comparables à celles de la normalisation par rapport de vraisemblances couplée à la technique Hnorm. Pour les deux techniques – World+MAP (sur *Dev*) et Hnorm – l'obtention de bonnes performances est attribuée à la dépendance des méthodes aux locuteurs clients (et à leurs modèles).

À l'opposé, lors de l'application de la fonction F_{WMap} sur le jeu *Eva*, l'approche World+MAP devient entièrement indépendante des locuteurs clients. Cette propriété de l'approche World+MAP se traduit par un gain de performances moindre sur *Eva*.

Ce dernier point met en avant la principale limite de l'approche World+MAP dès lors que deux jeux de données sont imposés pour l'apprentissage et l'application de la fonction de normalisation. Dans ce cas de figure, ces deux jeux de données doivent présenter des caractéristiques très similaires pour assurer l'efficacité de l'approche World+MAP.

Les résultats obtenus sur *Eva* sont toutefois très convenables. La généralisation de la fonction de normalisation F_{WMap} d'une population de locuteurs (*Dev*) à l'autre (*Eva*) montre des résultats tout à fait satisfaisants.

5 Architecture multi-reconnaisseur

Dans le chapitre précédent, nous soulignions le fait que l'utilisation de l'approche World+MAP peut être très intéressante dans le cadre d'une architecture multi-reconnaisseur pour deux raisons majeures :

1. des probabilités seraient manipulées lors de l'étape de fusion des scores produits par chacun des reconnaisseurs de l'architecture,
2. une fonction de normalisation dépendante du reconnaisseur permettrait d'intégrer implicitement la qualité intrinsèque de chaque reconnaisseur au niveau des probabilités émises.

Pour vérifier ces deux points, nous avons mené une expérience mettant en jeu une architecture multi-bande (voir chapitre 5 section 6 ou la référence [Besacier, 1998] pour plus de détails sur les architectures multi-bandes). L'architecture multi-bande utilisée ici repose sur :

- une bande totale (BT) composée de 16 coefficients cepstraux (issus d'une analyse en banc de filtres) représentant la bande de fréquence 300-3400 Hz ;
- trois sous-bandes (SB1, SB2, SB3), composée chacune de 8 coefficients cepstraux (issus d'une analyse en banc de filtres) représentant respectivement des bandes de fréquences de 300-1600Hz, 1100-3100Hz et 2500-4000Hz. La construction des sous-bandes SB1, SB2 et SB3 est explicitée par la figure 11.5.

Cette architecture multi-bande est couplée, ici, à l'approche bloc-segmentale. Dans ce contexte, un reconnaisseur fournit des scores[2] pour chaque bloc de 30 trames de signal. Comme illustrée sur la figure 11.6, cette configuration particulière requiert deux étapes de fusion :

1. *Fusion des reconnaisseurs :* pour un bloc donné, les scores produits par chaque reconnaisseur sont fusionnés. Cette première étape de fusion est répétée pour tous les blocs.
2. *Fusion des blocs :* les scores associés aux blocs sont finalement fusionnés pour fournir un score global au processus de décision. Dans cette expérience, cette étape de fusion des blocs repose sur une simple moyenne arithmétique.

Les expériences présentées par la suite ont été conduites sur les jeux de données *Dev* et *Eva*. Pour chaque expérience, l'approche World+MAP est comparée à la normalisation par rapport de vraisemblances. Ces deux techniques de normalisation emploient les modèles du monde dépendants du genre et du type de combiné téléphonique. Concernant l'approche World+MAP, une fonction de normalisation spécifique est apprise pour chaque reconnaisseur sur le jeu de données *Dev*.

5.1 Performances individuelles des reconnaisseurs

Des tests de vérification ont été menés sur chacun des reconnaisseurs présentés précédemment afin de connaître leurs performances individuelles. Les résultats de ces tests de vérification sont donnés, en termes de taux d'égale erreur (EER), par le tableau 11.2 pour chacune des techniques de normalisation : normalisation par rapport de vraisemblances (Rap. Vrais.) et approche World+MAP.

[2]Par exemple, des log-rapports de vraisemblances normalisés (probabilités) si l'approche World+MAP est utilisée.

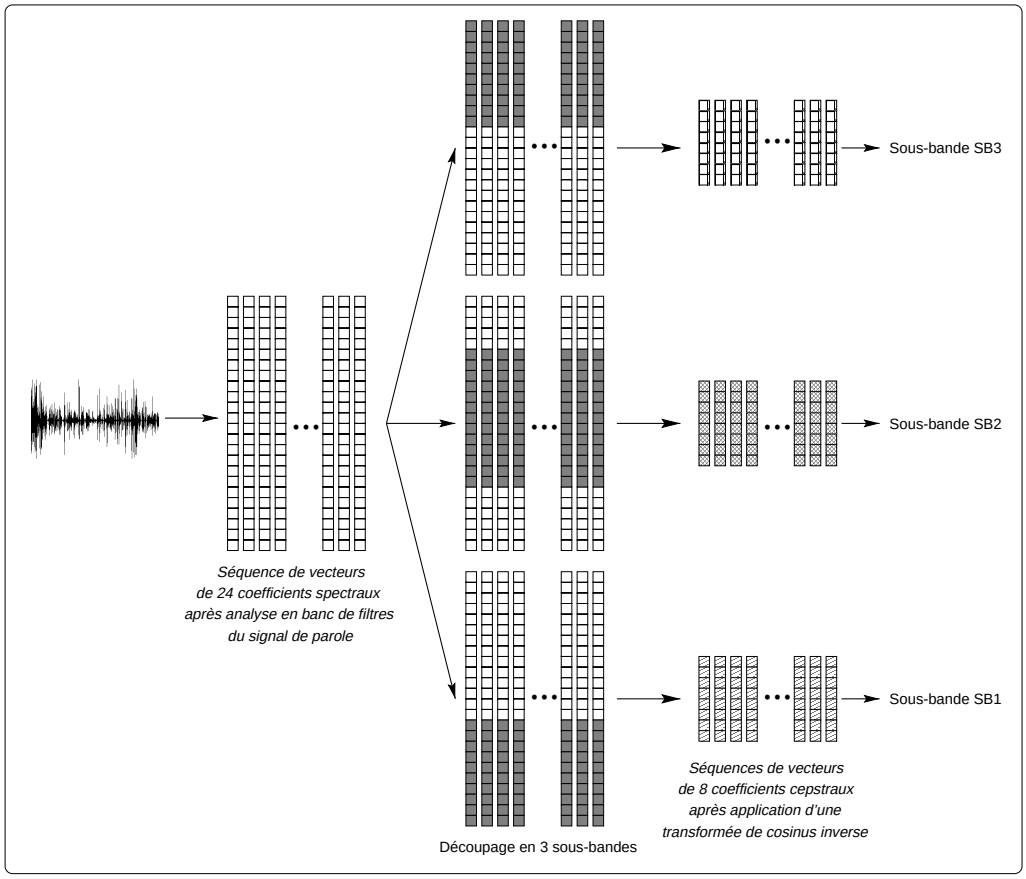

Figure 11.5: Construction des sous-bandes cepstrales. Détail des différentes étapes de construction des sous-bandes cepstrales.

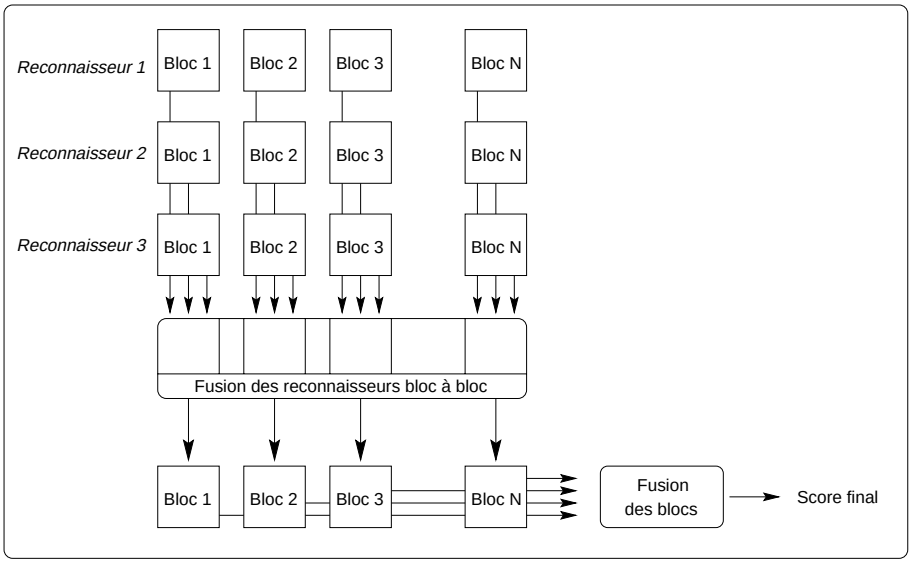

Figure 11.6: Etapes de fusion des reconnaisseurs et des blocs. Schéma des deux étapes de fusion impliquées lors de l'association d'une architecture multi-reconnaisseur à l'approche bloc-segmentale.

	Taux d'égale erreur : EER (en %)			
	Jeu de données *Dev*		Jeu de données *Eva*	
Reconnaisseurs	Rap. Vrais.	World+MAP	Rap. Vrais.	World+MAP
BT	16	15	17	16
SB1	28	27	27	27
SB2	20	20	22	23
SB3	30	29	30	29

Tableau 11.2: Performances individuelles des reconnaisseurs. Comparaison des performances, en termes de taux d'égale erreur, obtenues par chacun des reconnaisseurs selon l'utilisation de la normalisation par rapport de vraisemblances ou de l'approche World+MAP (Tâche de vérification sur Switchboard – Dev – 1190 tests clients et 8992 tests imposteurs – Eva – 948 tests clients et 8965 tests imposteurs).

D'après ce premier tableau de résultats, nous pouvons constater que les reconnaisseurs parviennent, pour chacun des jeux de données, à des performances comparables pour les deux techniques de normalisation considérées.

Ces taux d'égale erreur soulignent, en outre, des niveaux de qualité très disparates entre reconnaisseurs (les taux d'égale erreur peuvent varier de 0,15 à 0,30 suivant les reconnaisseurs). Ce dernier point est très important pour la phase de fusion des reconnaisseurs.

5.2 Performances du système multi-reconnaisseur

Le système multi-reconnaisseur est à présent testé sur le jeu de données *Eva*. Ces tests de vérification ont pour objectif d'évaluer l'intérêt de l'approche World+MAP dans l'étape de fusion des reconnaisseurs. Dans ce cadre, la normalisation par rapport de vraisemblances est proposée comme référence.

Deux schémas sont étudiés pour la combinaison des scores[3] issus de chaque reconnaisseur, pour un bloc donné :

1. combinaison des scores par simple moyenne arithmétique,
2. combinaison des scores par une moyenne arithmétique pondérée suivant la qualité intrinsèque de chaque reconnaisseur. Cette qualité se mesure ici en fonction du taux d'égale erreur (EER) obtenu par le reconnaisseur. Pour un reconnaisseur i, le poids associé est calculé suivant l'équation : $\frac{1-EER_i}{\sum_{j=1}^{N} 1-EER_j}$ avec N le nombre de reconnaisseurs.

Le tableau 11.3 reporte les taux d'égale erreur obtenus par chaque technique de normalisation suivant les deux schémas de fusion des scores issus de chaque reconnaisseur : moyenne arithmétique simple et moyenne arithmétique pondérée par les performances intrinsèques de chaque reconnaisseur.

Quel que soit le schéma de fusion considéré, nous pouvons constater que l'approche World+MAP conduit à un gain relatif moyen de 10% par rapport à la normalisation par

[3]probabilités pour l'approche World+MAP, log-rapports de vraisemblances pour la normalisation par rapport de vraisemblances.

	Schémas de fusion des scores	
Normalisation	Moyenne arithmétique	Moyenne pondérée (EER)
Rap. Vrais.	0,18	0,17
World+MAP	0,158	0,158

Tableau 11.3: Performances du système multi-reconnaisseur. Comparaison des performances, en termes de taux d'égale erreur, obtenues par le système multi-reconnaisseur en utilisant soit une normalisation par rapport de vraisemblances soit l'approche World+MAP, suivant le schéma de combinaison des scores employé (Tâche de vérification sur Switchboard – Eva – 948 tests clients et 8965 tests imposteurs).

rapport de vraisemblances.

Une analyse comparative des résultats entre les deux schémas de fusion, pour les deux approches considérées, montre que :

- une légère amélioration des taux d'égale erreur est observée pour l'approche par rapport de vraisemblances dès lors que la moyenne arithmétique pondérée est appliquée.
- les deux schémas de fusion conduisent à des performances identiques pour l'approche World+MAP.
- l'approche World+MAP parvient à des performances comparables à celles obtenues individuellement par le meilleur reconnaisseur (BT) et ce malgré la qualité relativement faible des autres reconnaisseurs (SB1, SB2 et SB3). Dans ce contexte, la normalisation par rapport de vraisemblances affiche une perte de performances.

Ces différents points démontrent que l'approche World+MAP est capable de prendre en compte la qualité intrinsèque de chaque reconnaisseur. Cette signification supplémentaire attribuée aux scores devenus probabilités permet l'application d'opérateurs de fusion simples tels que la moyenne arithmétique lors de l'étape de fusion des reconnaisseurs.

Chapitre 12

Conclusion sur World+MAP

Dans cette deuxième partie, notre attention s'est portée sur la problématique liée au processus de décision de la tâche de VAL. Rendre un système de VAL opérationnel dans le domaine applicatif implique le choix d'un seuil de décision global (indépendant des clients de l'application) fixé *a priori* pour répondre aux contraintes de l'application visée (point de fonctionnement). Or, la variabilité des vraisemblances - mesures de similarité entre un signal de test et un modèle client - fréquemment observée en présence de variations entre signaux d'apprentissage et de test par exemple, va à l'encontre de ce paradigme.

En réponse à cette problématique, la littérature propose trois grandes classes de normalisations ayant pour but de réduire la variabilité des vraisemblances. Appliquées à des niveaux différents au sein du système de VAL, ces normalisations s'attachent à réduire la variabilité due aux canaux de transmission (normalisation de l'espace des paramètres acoustiques), à des changements de matériel ou des conditions d'enregistrement (normalisation de l'espace des mesures de similarité) et au locuteur (normalisation de l'espace des seuils). Seules ou combinées, ces techniques de normalisation parviennent à de nettes améliorations de performances des systèmes de VAL.

Malgré la réduction notable de la variabilité des vraisemblances, ces techniques ne nous semblent pas satisfaisantes dans le sens où elles ne permettent pas de donner une réelle signification au seuil de décision. Dans ces conditions, le seuil de décision doit être choisi dans l'espace des vraisemblances normalisées et peut varier de 0 à $+\inf$. Il est, par conséquent, difficile de donner une interprétation à une valeur de seuil donnée.

L'approche World+MAP, que nous proposons dans cette deuxième partie, apporte une solution à ce problème. Elle répond à deux objectifs précis. En tant que technique de normalisation, elle permet de réduire la variabilité des vraisemblances, produites par le système de VAL et de doter les vraisemblances normalisées et, par association, le seuil de décision d'une signification. En effet, en combinant deux approches différentes : normalisation par rapport de vraisemblances et théorie bayésienne, l'approche World+MAP permet de transposer les vraisemblances dans un espace probabiliste, conférant au seuil une réelle signification dans le domaine des probabilités.

D'une manière plus pragmatique, nous avons montré, dans cette deuxième partie, que l'approche World+MAP s'appuie principalement sur l'estimation d'une fonction de normalisation (F_{WMap}). Cette estimation nécessite un jeu de données d'apprentissage sur lequel sont estimées des distributions de probabilités de rapports de vraisemblances clients et imposteurs ainsi que des probabilités *a priori* caractéristiques des conditions d'utilisation du système de VAL.

Les expériences relatives à l'application de cette fonction de normalisation lors de tests de vérification ont permis de démontrer le potentiel de l'approche World+MAP. Cette approche s'est révélée particulièrement efficace dès lors que la fonction de normalisation est apprise et appliquée sur le même jeu de données (*Dev*). Dans ce cas de figure, les performances de l'approche World+MAP sont comparables à plusieurs techniques de normalisation issues de l'état de l'art. En revanche, les résultats obtenus lors de l'apprentissage et l'application de la fonction de normalisation sur deux jeux de données séparés (*Dev* et *Eva*) indiquent une perte d'efficacité de l'approche World+MAP bien que cette dernière se maintienne à un niveau de performances meilleur que certaines techniques classiques de normalisation.

Cette perte raisonnable d'efficacité, due aux variations non prises en compte entre jeux de données, souligne la limite majeure de l'approche World+MAP. Dans le cas d'une application où la fonction de normalisation est apprise sur un jeu de données différent des conditions d'exploitation, l'efficacité de l'approche World+MAP est dépendante du degré de similitude des deux jeux de données considérés.

Finalement, des expériences menées dans le cadre d'un système multi-reconnaisseur ont révélé un avantage supplémentaire à l'utilisation de l'approche World+MAP. En effet, l'estimation de fonctions de normalisation dépendantes des reconnaisseurs permet aux probabilités émises lors de la normalisation "d'hériter" de la qualité intrinsèque de chaque reconnaisseur. Cet héritage permet l'utilisation d'opérateurs de fusion simples tels que la moyenne arithmétique et/ou géométrique lors de l'étape de fusion des scores (probabilités) produits par chacun des reconnaisseurs.

Chapitre 13

Campagnes d'évaluation NIST : Validation de l'approche "dynamique" et de la normalisation World+MAP

Ce chapitre présente des résultats complémentaires issus des campagnes d'évaluation NIST 1999 et 2000 concernant les approches "dynamique" et World+MAP. La tâche visée est ici la Vérification Automatique du Locuteur en milieu conversationnel et téléphonique (Switchboard).

---

1 Introduction

Les deux approches proposées dans ce travail de thèse ont été testées lors des dernières campagnes d'évaluation NIST. L'intérêt de plusieurs architectures multi-reconnaisseurs, impliquant l'approche "dynamique" ainsi que l'approche World+MAP, a été mesuré au cours de la campagne d'évaluation 1999 [Besacier et al., 2000a], [Fredouille et al., 2000a]. L'année suivante (campagne 2000), l'approche World+MAP seule a été testée.

Ce chapitre concerne exclusivement la tâche de Vérification Automatique du Locuteur (VAL) dans un contexte conversationnel et téléphonique. L'objectif principal est de montrer le comportement de l'approche "dynamique" dans le cadre de la VAL[1]. Par ailleurs, il s'agit de valider la capacité de généralisation de l'approche World+MAP.

2 Campagnes d'évaluation NIST

Depuis 1996, l'institut américain NIST (National Institute of Standards and Technologies) organise, tous les ans, des campagnes d'évaluation des systèmes de RAL indépendants du texte [Przybocki et al., 1998], [Przybocki et al., 1999], [Martin et al., 2000][2]. Ces évaluations sont réalisées dans un contexte conversationnel en milieu téléphonique (base de données Switchboard). Ces campagnes ont pour objectif d'être simples de mise en œuvre, de cibler des conditions d'exploitation des systèmes (jugées problématiques) et d'être accessibles à tout laboratoire de recherche public ou privé. À la suite de ces campagnes, un workshop réunissant tous les participants des évaluations permet de discuter des difficultés de cette dernière, des techniques mises en œuvre et des nouvelles alternatives à prévoir pour les campagnes suivantes.

Les objectifs techniques

La tâche principale des campagnes d'évaluation est, depuis 1996, la détection de locuteurs. Cette tâche de détection considère soit un signal de parole mono-locuteur (One speaker) soit un signal de parole bi-locuteur (Two speakers). Récemment, les campagnes ont intégré de nouvelles tâches comme le suivi de locuteurs (Speaker tracking) en 1999, ou l'Indexation par Locuteur de flux audio (Indexing) en 2000.

Les motivations technologiques de ces campagnes sont les suivantes :

- explorer de nouveaux concepts pour la RAL ;
- développer de nouvelles méthodes sur la base de ces concepts ;
- mesurer les performances de ces nouvelles méthodes.

Outre le contexte conversationnel en milieu téléphonique, ces campagnes se sont intéressées à des problématiques spécifiques, à l'origine des dégradations de performances des systèmes de VAL, telles que :

- la durée des signaux d'apprentissage et de test ;

[1] Les résultats relatifs à l'approche "dynamique", fournis dans la première partie de cette thèse, reposent uniquement sur la tâche d'Identification Automatique du Locuteur.

[2] Des informations concernant ces campagnes d'évaluation sont également consultables sur le site web : http ://www.nist.gov/speech/spkrinfo.htm.

- les variations de sessions d'enregistrement, de lignes téléphoniques ou de combinés téléphoniques entre les signaux d'apprentissage et de test ;
- l'impact de certaines caractéristiques des locuteurs (âge, genre, fréquence fondamentale) ;
- la langue.

Mise en place des évaluations

Une campagne d'évaluation se tient en deux temps. Dans un premier temps, les laboratoires participants reçoivent, pour chaque tâche visée, un jeu de données de développement (signaux d'apprentissage et de test) pour le réglage des systèmes. Dans un deuxième temps, un jeu de données d'évaluation, dépendant de la tâche, est distribué aux participants. Ce jeu comprend des signaux d'apprentissage ainsi qu'une large série de tests à réaliser en un temps donné. Une comparaison des systèmes présentés et des techniques mises en œuvre est finalement proposée durant le workshop de clôture.

3 Campagne NIST 1999

3.1 Jeux de données

La campagne d'évaluation 1999 repose sur deux jeux de données différents (un jeu de développement pour la mise au point des systèmes de VAL et un jeu d'évaluation pour les tests), dérivés du corpus Switchboard II.

3.1.1 Jeu de développement

Le jeu de développement utilisé ici est identique à celui décrit dans le chapitre 10 section 4.1. Il est extrait du jeu de données de la campagne d'évaluation 1998. Nous rappelons ici ses principales caractéristiques.

Ce jeu comprend deux populations distinctes de locuteurs :

- La première population est destinée à l'estimation des modèles du monde dépendants du genre et du type de combiné téléphonique. Elle se compose de signaux de parole de 30 secondes, produits par 100 femmes et 100 hommes. Ces signaux de parole sont également répartis suivant le type de combinés utilisés durant les sessions d'enregistrement : "electret" ou "carbon".
- La seconde population comprend 100 locuteurs également répartis en genre (50 hommes et 50 femmes). Un premier ensemble de signaux est disponible pour la phase d'apprentissage des modèles locuteurs. Il s'agit de signaux d'environ deux minutes de parole enregistrés lors de deux sessions. Le deuxième ensemble de signaux multi-sessions, d'une durée moyenne de 30 secondes, est quant à lui destiné aux tests de vérification. En vue de la campagne 1999, il est utilisé pour le développement et le réglage du système de VAL.

3.1.2 Jeu d'évaluation

Les tests d'évaluation de la campagne 1999 sont réalisés sur une population formée de 230 locuteurs masculins et de 309 locuteurs féminins. Pour chacun d'eux, des signaux de parole, d'une longueur moyenne de deux minutes et issus de deux sessions, sont fournis

pour la phase d'apprentissage des modèles ainsi que des signaux[3] de longueurs variables - de 2 secondes à 1 minute - pour la phase de test.

L'évaluation des systèmes de VAL repose sur un total de 3420 tests clients et 34200 tests imposteurs.

3.2 Système de VAL

Le système de base présenté à la campagne 1999 est le système AMIRAL développé au Laboratoire Informatique d'Avignon (LIA). Ce système, illustré par la figure 13.1, intègre différents modules dont certains sont développés au sein du consortium ELISA [Gravier et al., 1999], [ELISA, 2000].

Paramétrisation acoustique

Les différentes architectures multi-reconnaisseurs, testées dans les sections suivantes, sont définies au niveau du module de paramétrisation, développé au sein du consortium ELISA.
Chaque reconnaisseur représente une bande fréquentielle particulière (sous-bande), représentée par un nombre variable de coefficients cepstraux. Ces coefficients cepstraux sont tous dérivés d'une analyse en banc de filtres, normalisés par retrait de la moyenne cepstrale. Ces reconnaisseurs sont les suivants :

- Bande Totale Statique (BTS) : reconnaisseur statique représentant la bande fréquentielle totale (300-3400Hz) et composé de 16 coefficients cepstraux.
- Bande Totale Dynamique (BTD) : reconnaisseur dynamique représentant la bande fréquentielle totale (300-3400Hz) et composé de 16 coefficients cepstraux pris sur 9 trames successives de signal, soit 144 coefficients suivant l'approche "dynamique".
- 3 Sous Bandes Dynamiques (SBD) : reconnaisseurs dynamiques représentant les bandes fréquentielles 300-1600Hz, 1100-3100Hz, 2500-4000Hz. Les coefficients cepstraux qui les composent sont le résultat de la procédure de sélection (critère CritEmer) de l'approche "dynamique" appliquée initialement sur un ensemble de 8 coefficients cepstraux pris sur 9 trames de signal. Ces sous-bandes dynamiques sont identiques à celles présentées dans le chapitre 6 section 3.1.2. Les sous-ensembles de coefficients sélectionnés sont donnés en annexe B.

Modélisation

Les modèles clients et du monde reposent sur des mélanges de Gaussiennes (GMM) estimées par l'algorithme EM et l'approche EMV (cf. chapitre d'Introduction pour des informations relatives aux GMM).

Deux configurations sont utilisées pour les modèles : des GMM à 16 composantes, caractérisées par des matrices de covariance pleines (reconnaisseurs statiques) et des GMM à 128 composantes associées à des matrices de covariance diagonales (reconnaisseurs dynamiques).

[3] Ces signaux sont enregistrés au cours de plusieurs sessions.

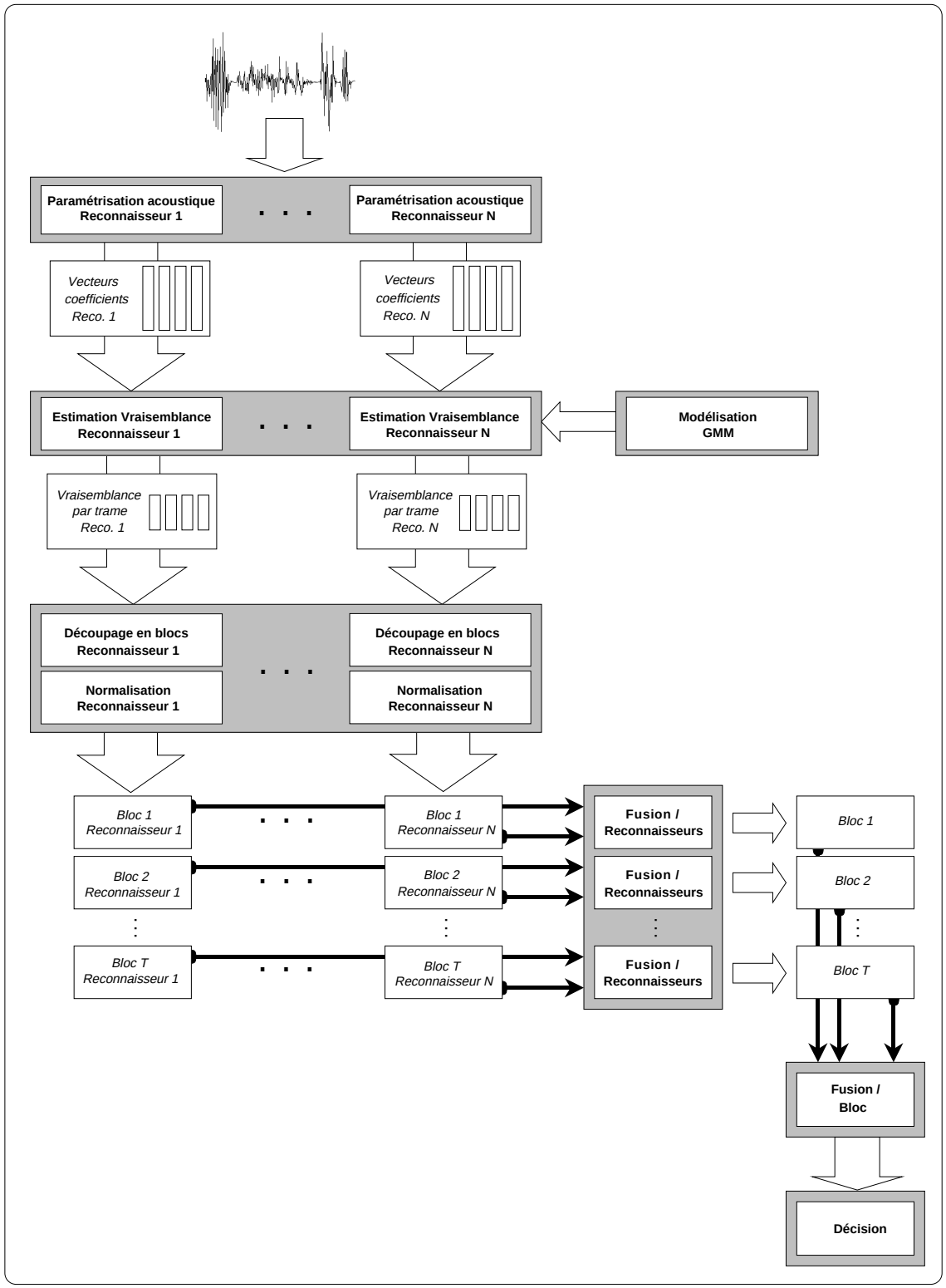

Figure 13.1: Système AMIRAL. Illustration des différents modules du système AMIRAL.

Segmentation/Normalisation

Le système AMIRAL repose sur les approches bloc-segmentale et World+MAP. Dans ce sens, l'approche World+MAP est appliquée, bloc à bloc[4], pour la normalisation des

[4]Les blocs demeurent d'une longueur de 30 trames de signal de parole.

vraisemblances entre un signal de test et le modèle du locuteur considéré. La fonction de normalisation F_{WMap} est apprise sur le jeu de développement. Elle reste, par conséquent, identique à celle présentée dans le deuxième volet de cette thèse.

Étapes de fusion

Comme présenté au chapitre 11 section 5, l'association d'une architecture multi-reconnaisseur et de l'approche bloc-segmentale nécessite deux étapes de fusion - fusion des reconnaisseurs et fusion des blocs. Ces deux étapes reposent ici sur une simple moyenne arithmétique.

Décision

Le seuil de décision, indépendant du locuteur, est fixé *a priori* sur le jeu de développement.

3.3 Architectures multi-reconnaisseurs

Trois architectures multi-reconnaisseurs sont évaluées au cours de la campagne 1999 :

- Architecture statique : BTS. Considérée comme l'architecture de référence, elle se résume au reconnaisseur BTS.
- Architecture hybride : BTS+SBD. Elle est composée du reconnaisseur statique BTS couplé aux trois reconnaisseurs dynamiques SBD [5].
- Architecture dynamique : BTD+SBD. Elle inclut les quatre reconnaisseurs dynamiques : BTD et SBD.

L'objectif de ces trois architectures est d'évaluer d'une part la corrélation entre informations statiques et dynamiques (architecture hybride) et d'autre part d'évaluer les performances de reconnaisseurs dynamiques (architecture dynamique) par comparaison avec une architecture statique (architecture statique).

La figure 13.2 reporte les courbes DET obtenues par chacune des architectures. L'architecture dynamique – BTD+SBD – fournit les meilleures performances. Les architectures hybride - BTS+SBD - et statique - BTS - parviennent à des résultats très similaires avec un léger avantage pour l'architecture hybride sur certaines portions de la courbe DET.

Par ailleurs, la figure 13.3 compare la courbe DET obtenue par l'architecture dynamique à celle obtenue par la bande totale dynamique, testée seule. En termes de taux d'égale erreur, les résultats sont très proches. Néanmoins, l'architecture dynamique montre une légère amélioration des résultats sur certaines portions de la courbe. En particulier, nous pouvons noter une amélioration pour des taux faibles de fausses acceptations qui correspondent au point de fonctionnement choisi pour les évaluations NIST en vue de comparer les systèmes.

Ces observations généralisent les résultats obtenus en IAL et confirment l'intérêt des informations dynamiques pour la RAL.

D'autre part, ces résultats semblent de nouveau souligner le manque d'efficacité de la procédure de sélection dans un tel contexte (conversationnel, multi-session et

[5] Comme indiqué à la section 3.2, la procédure de sélection de l'information dynamique utile est appliquée uniquement sur ces trois reconnaisseurs dynamiques.

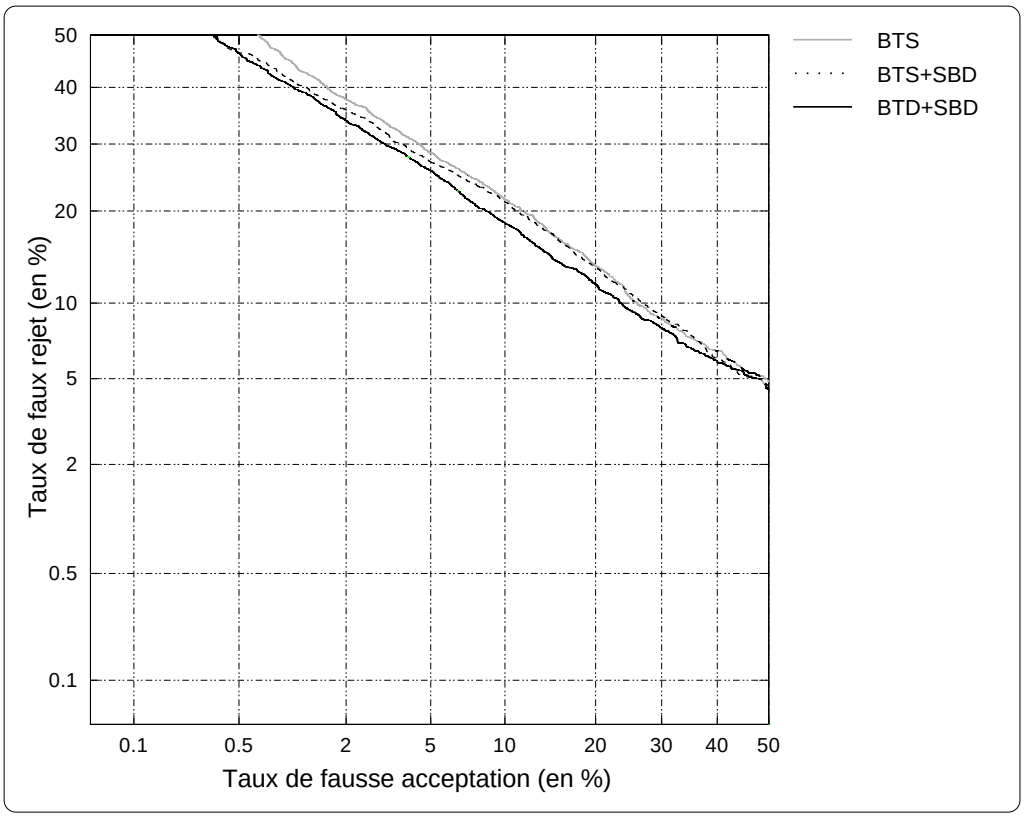

Figure 13.2: Architectures multi-reconnaisseurs. Comparaison des courbes DET obtenues par différentes architectures multi-reconnaisseurs : architecture statique (BTS), architecture hybride (BTS+SBD) et architecture dynamique (BTD+SBD). (Tâche de Vérification – Campagne d'évaluation 1999 – 3420 tests clients et 34200 tests imposteurs.)

téléphonique). La sélection sur les différentes sous-bandes semble n'amener que peu de gain en performances.

4 Campagne NIST 2000

4.1 Jeux de données

Comme précédemment, deux jeux de données sont fournis pour la campagne 2000, extraits du corpus Switchboard II.

Jeu de développement

Ce jeu de données est d'une structure similaire à celle du jeu de la campagne 1999. Les deux populations présentent des caractéristiques identiques. Seuls les locuteurs de chacune des populations sont différents. Ils proviennent du jeu de données de la campagne 1999.

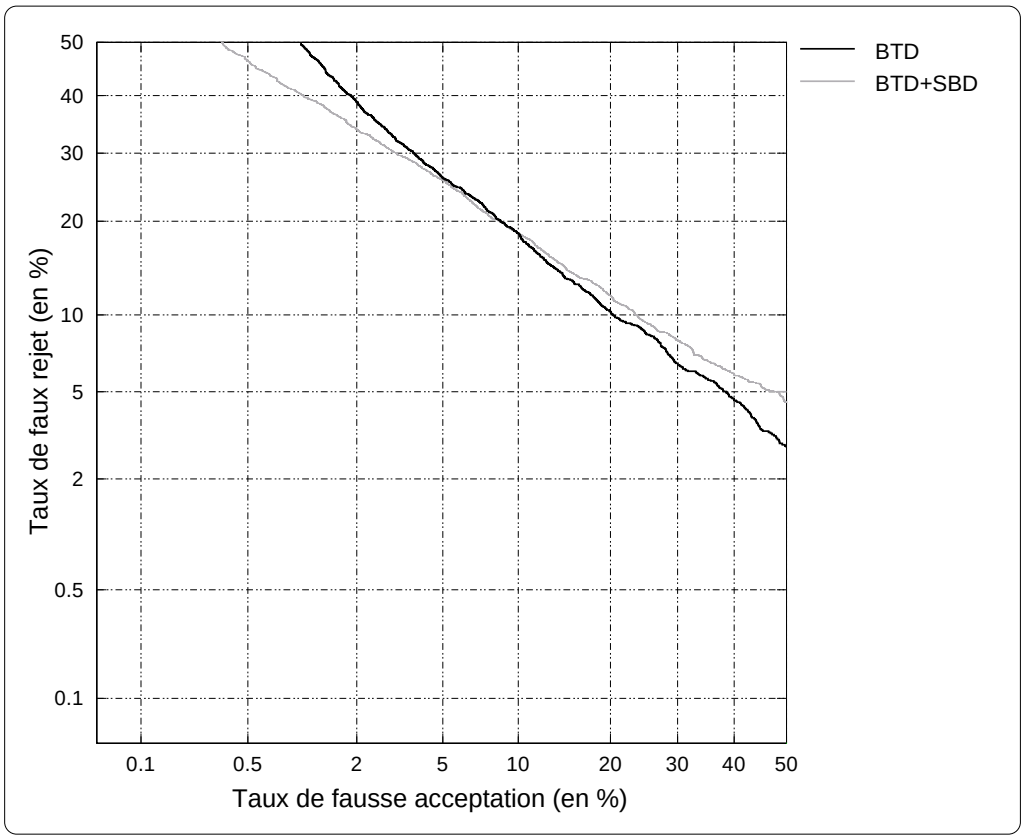

Figure 13.3: Architectures dynamiques. Comparaison des courbes DET obtenues par l'architecture dynamique (BTD+SBD) et la bande totale dynamique seule (BTD). (Tâche de Vérification – Campagne d'évaluation 1999 – 3420 tests clients et 34200 tests imposteurs.)

Jeu d'évaluation

Le jeu de données de la campagne 2000[6] repose sur une population de 500 locuteurs également répartis en genre. Des signaux de parole d'une durée d'environ deux minutes, mono-sessions[7], sont disponibles pour l'apprentissage des modèles clients ainsi que des signaux de test, multi-sessions, d'une durée moyenne de 30 secondes. L'évaluation des systèmes de VAL s'effectue sur 6052 tests clients et 60520 tests imposteurs.

4.2 Système de VAL

Le système de VAL présenté lors de la campagne 2000 reste, en grande partie, identique au système AMIRAL, détaillé en section 3.2. Seul le module de modélisation a subi des améliorations notoires. En effet, l'approche EMV, utilisée pour l'estimation des paramètres[8] des modèles clients, est remplacée par un processus d'adaptation des paramètres d'un modèle du monde à partir des signaux d'apprentissage des clients. Dans

[6]La constitution de la base de données Switchboard étant achevée, la campagne 2000 n'a pu bénéficier, cette année, de nouveaux locuteurs. Elle reprend, en fait, les locuteurs utilisés durant la campagne 1998.

[7]Il est à noter que les conditions de test sont ici plus difficiles par comparaison à la campagne NIST 1999 qui faisait état de signaux d'apprentissage enregistrés sur deux sessions.

[8]Ces paramètres sont le vecteur moyen, la matrice de covariance et le poids de chaque gaussienne du modèle.

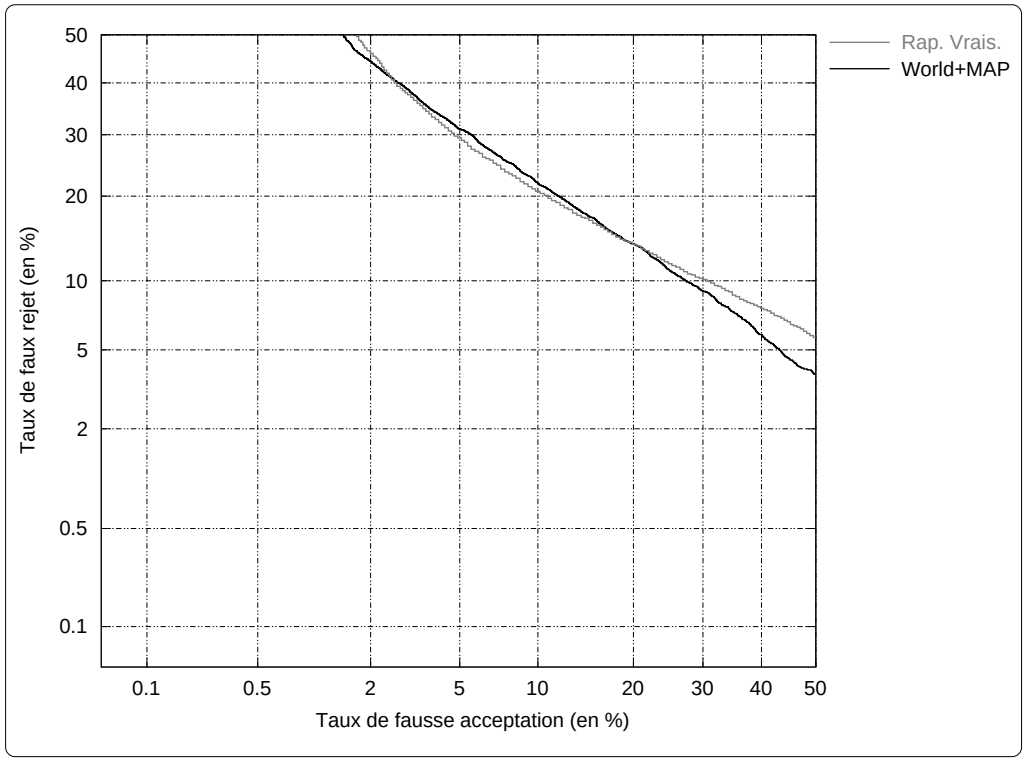

Figure 13.4: Apport de l'approche World+MAP. Comparaison des courbes DET obtenues par normalisation à l'aide du rapport de vraisemblances seul et de l'approche World+MAP. (Tâche de Vérification – Campagne d'évaluation 2000 – 6052 tests clients et 60520 tests imposteurs.)

ce sens, les paramètres des modèles clients sont dérivés des paramètres du modèle du monde, estimés sur une plus grande quantité de données. La technique d'adaptation est basée sur l'approche MAP [Gauvain et al., 1994].

Le système de VAL, dont les résultats sont présentés par la suite, se restreint à un seul reconnaisseur, représenté par 16 coefficients cepstraux et 16 coefficients Delta.

Concernant l'approche World+MAP, la fonction de normalisation est apprise sur le jeu de développement (données de la campagne 1999).

4.3 Apport de l'approche World+MAP

Afin de mesurer l'apport de l'approche World+MAP, les résultats obtenus lors de la campagne 2000 sont comparés à des résultats provenant d'expériences supplémentaires. Ces expériences, menées dans des conditions identiques, concernent l'utilisation du rapport de vraisemblances (modèles du monde dépendants du genre et du type de combinés téléphoniques) comme technique de normalisation.

La figure 13.4 reporte les courbes DET obtenues avec le rapport de vraisemblances seul (Rap. Vrais.) et l'approche World+MAP.

La comparaison de ces deux courbes montre des performances similaires pour les deux

approches.

Cette constatation confirme la bonne généralisation de l'approche World+MAP dès lors que la fonction de normalisation est apprise sur un ensemble de données caractérisant le jeu d'exploitation. Dans ce cas de figure, les deux jeux de données présentent des variations plus marquées que lors des expériences du chapitre 11.

Cette différence entre jeux de données se manifeste naturellement sur les fonctions de précision, opposant les probabilités réelles calculées sur les données de tests et les probabilités fournies par les fonctions de normalisation dépendantes du genre (figure 13.5).

Les fonctions de normalisation accusent un comportement relativement satisfaisant dans l'intervalle $[0; 0.65]$ (similaire au comportement de la fonction de normalisation, relevé dans le chapitre 10, pour le jeu de données *Eva*). À l'opposé, nous pouvons remarquer que ces fonctions sur-estiment très largement les probabilités réelles dans l'intervalle ($[0.65; 1]$). Néanmoins, cet intervalle concerne uniquement 0.1% des scores produits sur le jeu de tests et n'a, par conséquent, qu'une très légère incidence sur les performances du système.

5 Conclusions

La participation aux campagnes d'évaluation NIST 1999 et 2000 nous a permis de valider le comportement des approches proposées dans ce travail de thèse – approche "dynamique" et approche World+MAP.

Les résultats obtenus lors de ces campagnes confirment amplement les conclusions tirées des travaux présentés tout au long de ce document. Ces résultats ont, une nouvelle fois, démontré :

- la pertinence des informations de nature dynamique dans le cadre d'un système de RAL ;
- le maintien des performances d'un système de VAL dans le cadre de l'utilisation de l'approche World+MAP malgré la projection des vraisemblances dans un espace probabiliste ;
- la bonne généralisation de l'approche World+MAP malgré l'estimation de la fonction de normalisation sur un jeu de données différent du jeu d'exploitation.

Par ailleurs, le manque d'efficacité de la procédure de sélection, en présence de signaux multi-sessions et multi-canaux (base de données Switchboard), s'est de nouveau manifesté.

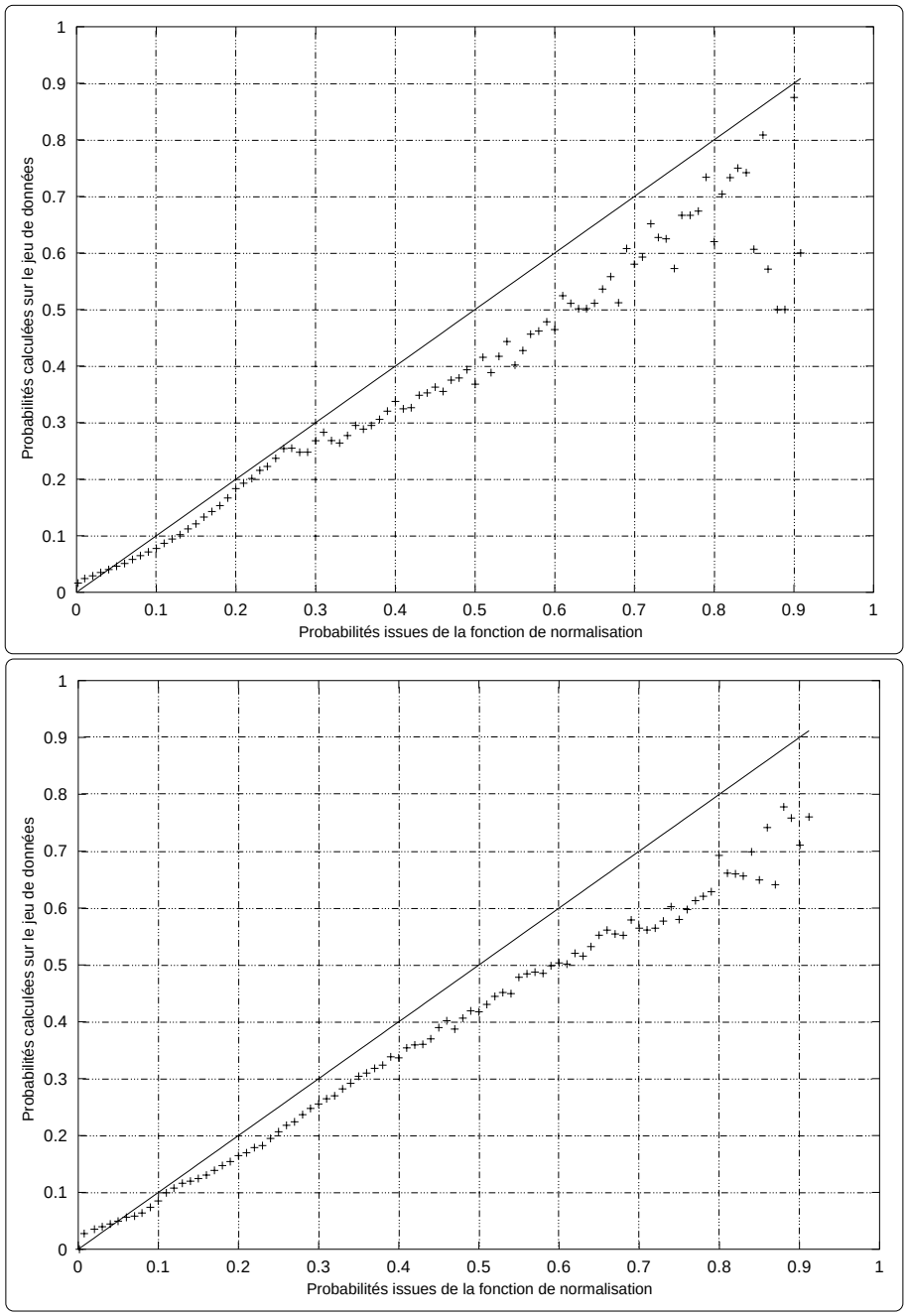

Figure 13.5: Précision de la fonction de normalisation – NIST 2000. Comparaison des probabilités fournies par les fonctions de normalisation hommes (figure du haut) et femmes (figure du bas) apprises sur les données NIST 1999 et des probabilités réelles calculées sur les données NIST 2000. (Tâche de Vérification – Campagne d'évaluation 2000 – 6052 tests clients et 60520 tests imposteurs.)

Chapitre 14

Conclusions – Perspectives

Ce travail de thèse s'inscrit dans le domaine de la Reconnaissance Automatique du Locuteur (RAL). La RAL consiste à reconnaître l'identité d'un individu à l'aide de sa voix. Essentiellement axée sur les problèmes d'authentification et de confidentialité, la RAL est très souvent sollicitée dans un contexte téléphonique, pour lequel la voix demeure le seul élément disponible pour reconnaître une personne. Pourtant, l'environnement téléphonique est à l'origine de graves perturbations au sein des systèmes de RAL (bande téléphonique, variation de combinés entre signaux d'apprentissage et de test, etc.. Ce point est d'ailleurs une des principales orientations des campagnes NIST, qui sont destinées, depuis 1996, à évaluer la robustesse des systèmes de RAL dans un tel contexte téléphonique (base de données Switchboard). L'ensemble des travaux présentés dans ce document ont été réalisés et validés dans cette direction.

Deux thèmes sont abordés dans cette thèse : l'utilisation des caractéristiques dynamiques du locuteur véhiculées par le signal de parole ainsi que la normalisation des vraisemblances en vue d'améliorer le processus de décision des systèmes de VAL. Bien qu'ils se rapportent à des processus différents au sein d'un système de RAL – paramétrisation et modélisation pour les informations dynamiques et processus de décision pour la normalisation – ces thèmes présentent deux points communs :

- le cadre général des travaux proposés dans l'un ou l'autre thème repose sur une approche statistique ;
- les deux thèmes apportent leur contribution respective au sein d'une architecture multi-reconnaisseur.

Les information dynamiques

Les différents travaux consacrés aux informations dynamiques (et cités dans le chapitre 4) ont montré l'intérêt de cette source d'informations dans le cadre de la RAL. Dans cette voie, nous avons proposé une approche originale pour leur traitement. Cette approche repose sur la concaténation de trames successives issues d'une fenêtre temporelle glissant le long du signal de parole. Une procédure de sélection permet, dans un deuxième temps, de déterminer la part d'information spécifique du locuteur.

Les travaux présentés dans ce document ont confirmé la nécessité d'une large fenêtre temporelle, de l'ordre de 100 milli-secondes, pour une exploitation correcte des informations dynamiques véhiculées par le signal de parole. Nous avons montré que, malgré sa mise en œuvre simple et peu coûteuse en temps de calcul, l'approche dynamique proposée est tout à fait adaptée à une telle taille de fenêtre au contraire d'une grande part des méthodes usuelles comme les techniques prédictives (modèles ARV).

En outre, l'utilisation d'un protocole particulier, basé sur le mélange aléatoire des trames du signal dans le plan temporel, a démontré la nature manifestement dynamique des informations manipulées par notre approche.

La procédure de sélection de l'information utile a permis une amélioration significative des performances sur la base de données TIMIT (gain relatif de 6 à 20% des taux d'identification pour le critère CritId et de 8 à 31% pour le critère CritConf par comparaison avec l'utilisation d'informations statiques).

Par ailleurs, des expériences relatives à l'utilisation de l'approche dynamique sans procédure de sélection sur la base de données Switchboard ont également conduit à une amélioration des performances (gain relatif de 10 à 20% des taux d'identification comparés à l'utilisation d'informations statiques). Néanmoins, aucune amélioration n'a pu être apportée par la procédure de sélection (une dégradation des résultats a même été observée dans certains cas). Ce comportement de la sélection, contraire aux résultats obtenus sur la base TIMIT, provient majoritairement du caractère multi-session et multi-canal de la base Switchboard. Dans ce contexte, l'utilisation d'un jeu de données de sélection différent du jeu d'exploitation rend apparemment la sélection de l'information utile inopérante.

Finalement, il est intéressant de souligner que l'approche dynamique dans le contexte NIST/Switchboard est équivalente, en termes de performances, aux techniques dynamiques "état de l'art" telles que l'utilisation des coefficients dérivés Delta et Delta-Delta.

L'approche World+MAP

Dans la deuxième partie de cette thèse, nous nous sommes intéressés au processus de décision pour la tâche de Vérification du Locuteur. Outre la nécessité d'une normalisation des vraisemblances en vue de la décision, nous avons souligné une lacune des techniques de normalisation proposées dans la littérature : ces techniques ne permettent pas de doter le seuil de décision d'une interprétation directe, pourtant utile d'un point de vue applicatif. Afin de pallier cette déficience, nous avons proposé une technique de normalisation originale nommée World+MAP. Cette approche est fondée essentiellement sur une technique de normalisation classique et sur l'utilisation de probabilités *a posteriori*. Elle présente comme premier avantage de projeter les vraisemblances dans un espace de probabilités, conférant au seuil de décision une signification directement interprétable. D'autre part, l'approche World+MAP permet, dans le cadre d'une architecture multi-reconnaisseur, d'intégrer de manière implicite la qualité intrinsèque de chaque reconnaisseur dans les scores de reconnaissance, devenus des probabilités. Cette deuxième propriété de l'approche World+MAP facilite amplement l'étape de fusion des scores des reconnaisseurs au sein de l'architecture.

Nous avons, dans un premier temps, démontré que la projection à l'aide de l'approche World+MAP des vraisemblances dans un espace de probabilités (intervalle $[0, 1]$) n'entraîne aucune perte de performances lors d'expériences en vérification du locuteur sur la base de données Switchboard. De plus, ces expériences ont souligné la généralisation tout à fait satisfaisante de la fonction de normalisation dès lors qu'elle est apprise sur un ensemble de données présentant des propriétés similaires aux conditions d'exploitation.

D'autre part, la signification des probabilités fournies par la fonction de normalisation a été validée, sur les données d'apprentissage de la fonction comme sur les données d'exploitation (ainsi que durant les évaluations NIST2000) par l'estimation d'une mesure de précision. Parallèlement, nous avons démontré, au cours d'expériences conduites sur une architecture multi-reconnaisseur, la capacité de l'approche World+MAP à assimiler la qualité intrinsèque de chaque reconnaisseur dans la valeur des probabilités.

Finalement, il est intéressant de noter que l'approche World+MAP, bien qu'indépendante du locuteur (dès lors que la fonction de normalisation est apprise

sur un jeu de données séparé), obtient des performances de même niveau que des techniques de normalisation classiques dépendantes du locuteur (Znorm, Hnorm).

Perspectives

Bien que la procédure de sélection se soit avérée très pertinente sur la base de données TIMIT, la base Switchboard a montré les limites de l'algorithme de sélection.

Une solution potentielle consisterait à travailler directement au niveau des matrices de covariance. En effet, les matrices sont de forme connue (matrices bloc-Toeplitz) et acceptent certainement un algorithme formel permettant de mettre en évidence les partitions optimales.

Par ailleurs, une étude approfondie des ensembles de coefficients sélectionnés à partir des signaux de TIMIT et Switchboard, sur un plan phonétique, permettrait de mieux cibler la part utile d'information dynamique présente dans tout signal de parole. Cette analyse phonétique pourrait conduire à l'élaboration de procédures de sélection mieux adaptées à ce contexte. D'autre part, cette étude pourrait être généralisée dans le cadre de la caractérisation du locuteur pour lequel l'analyse phonétique pourrait apporter des connaissances supplémentaires pour renseigner sur l'état émotionnel ou pathologique d'un individu par exemple.

Les développements futurs des travaux réalisés sur l'approche World+MAP peuvent porter sur différents points. Ils concernent tout d'abord la mise en œuvre de l'approche World+MAP au niveau ingénierie (estimation de la fonction de normalisation). Ensuite, un approfondissement de l'approche World+MAP dans le contexte bloc-segmental et d'une architecture multi-reconnaisseur peut être envisagé. Finalement, l'utilisation de l'approche World+MAP pour des tâches autres que la VAL est à considérer.

- L'approche World+MAP s'est révélée très pertinente dans le cadre d'une architecture multi-bande. L'utilisation de probabilités tenant compte de la qualité intrinsèque de chaque sous-bande fréquentielle a simplifié l'étape de fusion des reconnaisseurs. Une extension de ce travail à une architecture mettant en jeu des reconnaisseurs différents (architectures multi-classificateurs et multi-modalités) est en préparation.

- Dans ce document, l'approche World+MAP est couplée à l'approche bloc-segmentale. Dans cette configuration, le signal de parole est décomposé en une séquence de blocs, chacun de ces blocs étant associé à une probabilité. Une étape de fusion permet de combiner les probabilités pour fournir un score (probabilité) final au processus de décision. Cette étape repose ici sur une simple moyenne arithmétique. Pourtant, il serait intéressant d'étendre ce travail par la recherche d'opérateurs de fusion plus intelligents s'appuyant davantage sur les propriétés probabilistes des scores attribués à chaque bloc. Cette voie de recherche serait particulièrement fructueuse pour une application de l'approche World+MAP dans le cadre de la détection de locuteurs dans une conversation (multi-locuteur) ou dans le cadre du suivi de locuteurs.

- La littérature montre fréquemment que l'association de plusieurs techniques de normalisation, vouées à la réduction de la variabilité des vraisemblances, permet

d'améliorer les performances des systèmes de VAL. Par exemple, lors de la campagne d'évaluation NIST 2000, l'application successive du rapport de vraisemblances, de la technique Hnorm et de la technique Tnorm s'est révélée la plus pertinente. L'efficacité de cette association est due au fait que les techniques de normalisation agissent à des niveaux différents et/ou intègrent des connaissances hétérogènes pour réduire la variabilité des vraisemblances. Une étude approfondie de ces techniques de normalisation pourrait permettre de mieux appréhender les phénomènes liés à la réduction de la variabilité des vraisemblances. Celle-ci aurait pour principal objectif de proposer un schéma d'unification des différentes techniques de normalisation, élaboré dans le cadre de l'approche World+MAP.

- L'application de l'approche World+MAP peut être étendue à d'autres domaines que la RAL. En effet, toute tâche autorisant la manipulation de vraisemblances intra- et inter-classes peut s'appuyer sur l'approche World+MAP. Nous pensons notamment à la Reconnaissance Automatique de la Parole (décodage acoustico-phonétique par exemple). Dans ce cadre, notre approche offre deux avantages majeurs. De par son principe, elle permet de manipuler des probabilités au sein des systèmes. D'autre part, l'interprétation probabiliste proposée par notre approche peut être vue comme une mesure de confiance, appréciable lors d'un processus de décision.

Annexe A : Algorithmes de sélection

Présentation des algorithmes

La littérature fait état de trois grandes classes d'algorithmes de sélection. Nous présentons ici chacune de ces classes en décrivant leur principe de base ainsi que les techniques afférentes. L'utilisation de ces techniques dans le cadre qui nous intéresse – la sélection du meilleur sous-ensemble de coefficients – est finalement discutée.

Algorithmes exponentiels

Cette catégorie fait référence aux algorithmes dont la complexité est exponentielle (de type $O(2^n)$ avec n le nombre initial de coefficients). Elle comprend notamment les techniques dites "exhaustives" qui consistent à évaluer tous les sous-ensembles d'éléments possibles dès lors que le nombre de coefficients potentiels est relativement faible (technique optimale).

Algorithmes aléatoires

Deux grands types d'algorithmes sont étiquetés comme aléatoires : les algorithmes génétiques et les algorithmes de recuit simulé.

Les algorithmes génétiques, introduits par [Holland, 1975], puis repris par [Goldberg, 1989] s'inspirent de la représentation binaire de l'ADN et des mécanismes de sélection naturelle. Dans ce sens, les algorithmes génétiques de base reposent sur un codage, sous forme de chaînes binaires appelées *chromosomes*, des solutions potentielles d'un problème et sur des opérations de *reproduction* (un chromosome est sélectionné[1] et reproduit à l'identique), de *mutation* (un ou plusieurs bits d'un chromosome sont inversés) et de *croisement* (création d'un chromosome enfant à partir de deux chromosomes parents). Le processus est le suivant : à chaque *génération* un nouvel ensemble de chromosomes est généré par *reproduction* – *croisement* – *mutation* et évalué. Si le critère d'arrêt est atteint, l'algorithme génétique s'arrête et rend le meilleur chromosome produit sinon le processus est réitéré.

Dans le contexte qui nous intéresse – la recherche du meilleur sous-ensemble de coefficients dynamique – la difficulté principale réside dans la transcription du problème en un code binaire.

[1] La sélection des chromosomes est pseudo-aléatoire. En effet, les chromosomes répondant au mieux à la fonction d'évaluation auront plus de chance d'être sélectionnés que les autres.

La solution la plus utilisée, que ce soit dans le domaine du diagnostic médical [Yang et al., 1997], de la reconnaissance de textures d'images [Vafaie et al., 1993] ou de la reconnaissance du locuteur [Demirekler et al., 1999], est de considérer un coefficient potentiel comme un bit d'une chaîne. Si ce bit est à 1, le coefficient est sélectionné, s'il est à 0 il est rejeté. Dans ce contexte, un chromosome est formé de n bits, correspondant aux n coefficients potentiels. Le nombre de chromosomes varie selon les études. Dans [Demirekler et al., 1999], un chromosome unique est défini et seule l'opération de mutation est appliquée durant la procédure de sélection. Dans [Yang et al., 1997], la population de chromosomes représente tous les sous-ensembles possibles de coefficients.

Un des intérêts des algorithmes génétiques est d'éviter au maximum d'atteindre un optimum local. En effet, l'opération de *croisement*, qui permet de passer d'une population à une autre a pour effet, à partir d'un certain nombre de *générations*, d'éviter de stagner dans un espace de recherche trop restreint. De même, l'opération de *mutation* permet d'introduire de nouvelles informations dans la population existante et doit éviter ces mêmes phénomènes de stagnation. En pratique, cette dernière opération est peu utilisée et ne permet pas d'écarter entièrement le risque d'évolution vers une population trop homogène.

Les algorithmes dits évolutifs, basés sur le recuit simulé, sont considérés comme une solution à ce problème [Liu et al., 1998]. Ces derniers introduisent la notion de température – fonction décroissante du temps – qui permet d'assurer plus aisément une convergence vers un optimum global. Celle-ci est fonction du nombre de générations et agit de différentes manières. Premièrement, elle empêche les *chromosomes* trop faibles de se reproduire par l'emploi d'un critère de sélection de plus en plus restrictif. Par ailleurs, on décide, lors du croisement, de créer plusieurs enfants par couple de parents et de ne garder que le meilleur d'entre eux. Plus les frères de celui-ci sont forts, plus il pourra créer d'enfants à la *génération* suivante. Le nombre d'enfants créés diminue avec la baisse de température.

Algorithmes séquentiels.

Les algorithmes dits séquentiels se caractérisent par une complexité polynomiale (de type $O(n^2)$) comme le démontrent les approches présentées ici :

- *la sélection des N-meilleurs coefficients.* Cette approche consiste à évaluer séparément chacun des éléments selon un critère de sélection prédéfini (par exemple le F-ratio [Pruzansky et al., 1964], [Wolf, 1972]) et à sélectionner les p meilleurs coefficients au sens du critère choisi. La valeur de p étant indéterminée *a priori*, une évaluation de tous les sous-ensembles de taille p variant de 1 à n est réalisée. Le nombre d'évaluations à effectuer est par conséquent réduit à $n^2 - 1$. Cependant, ce mécanisme de sélection impose une contrainte forte pour assurer son optimalité : *le meilleur sous-ensemble de p coefficients doit contenir les p meilleurs coefficients sans tenir compte de leur éventuelle interaction* [Charlet, 1997].
- *la méthode knock-out.* Cette approche, proposée dans [Sambur, 1975], [Aha et al., 1996], commence par évaluer tous les sous-ensembles de $n-1$ coefficients. Le meilleur sous-ensemble de taille $n-1$ est ensuite sélectionné, désignant le coefficient exclu du sous-ensemble comme le coefficient le moins pertinent au sens du critère de sélection choisi. Ce coefficient est alors éliminé (knocked-out) de la procédure de sélection. Cette dernière est réitérée jusqu'à ce que les n coefficients soient éliminés. Cette approche réduit le nombre d'évaluations à $\frac{n(n+1)}{2}$ mais impose à chaque itération que le meilleur sous-ensemble de p coefficients inclut le meilleur sous-ensemble de $p-1$

coefficients pour assurer l'optimalité du sous-ensemble de coefficients sélectionnés.

- *la sélection ascendante.* Cette procédure, proposée dans [Aha et al., 1996], [Charlet, 1997], peut être considérée comme une variante de la méthode knock-out. Si cette dernière élimine au fur et à mesure des itérations les coefficients les moins pertinents, la méthode ascendante, quant à elle, rajoute les coefficients les plus significatifs. En effet, cette procédure commence par évaluer tous les sous-ensembles constitués de 1 coefficient. La sélection du meilleur sous-ensemble désigne le coefficient associé comme étant le plus significatif au sens du critère de sélection choisi. Sur la base du meilleur sous-ensemble sélectionné, cette procédure est réitérée jusqu'à ce que les n coefficients soient ajoutés. Cette approche possède les mêmes caractéristiques que la méthode knock-out tant au niveau du nombre d'évaluations que de la condition d'optimalité.

Discussions sur les techniques de sélection

La qualité d'une technique de sélection se mesure évidemment par le degré d'optimalité du sous-ensemble de coefficients sélectionnés. Cependant, la classification de [Doak, 1992] nous montre que la complexité d'une telle technique, dépendante du nombre de coefficients potentiels, est un facteur important qui doit intervenir dans la mesure de cette qualité. En effet, suivant la fonction d'évaluation de la qualité d'un sous-ensemble, la procédure de sélection peut s'avérer être très coûteuse en temps de calcul.

Nous reportons ici différents travaux comparatifs réalisés dans des domaines variés afin de tirer nos propres conclusions quant au choix de la technique de sélection adaptée à notre contexte applicatif.

Ces différentes études ne font état d'aucune comparaison avec des algorithmes exponentiels. Cette constatation s'explique naturellement par la complexité exponentielle de ce type d'algorithmes qui les rend inapplicables dès lors que le nombre de coefficients n est grand.

Les algorithmes séquentiels décrits précédemment sont testés et comparés dans [Charlet, 1997] dans le cadre de la sélection d'un jeu optimal de paramètres acoustiques pour la vérification du locuteur en mode dépendant du texte. La sélection des N-meilleurs s'avère être de loin la moins performante à cause de sa simplicité. En revanche, les trois autres techniques montrent des résultats très similaires ne permettant pas de les différencier en termes de performances.
Dans cette même étude, une variante est proposée pour les méthodes knock-out et ascendante afin d'alléger leur contrainte d'optimalité. Cette variante est basée sur une remise en cause des coefficients du meilleur sous-ensemble sélectionné à chaque itération. L'intérêt de cette approche est de remettre en question l'interaction éventuelle des coefficients à chaque itération. Cependant, en apportant une augmentation non négligeable du nombre d'évaluations à réaliser, peu de changements au sein des meilleurs sous-ensembles de coefficients sélectionnés sont observés lors d'expériences réalisées dans les mêmes conditions.
Les mêmes conclusions sont reportées dans [Aha et al., 1996] concernant la comparaison entre les méthodes knock-out et ascendante appliquées sur des données météorologiques.
Cependant, ces conclusions contredisent l'étude menée par [Doak, 1992] qui affirmait, quelques années plus tôt, que la méthode knock-out était plus performante que la sélection

ascendante, à nombre d'itérations égal, au regard des résultats expérimentaux obtenus.

Une comparaison entre un algorithme génétique et la méthode knock-out est proposée dans [Vafaie et al., 1993]. Cette comparaison révèle que les performances de l'une ou de l'autre méthode sont fortement liées au degré d'interaction entre les coefficients à sélectionner ; la méthode knock-out se montre plus performante que l'algorithme génétique dans le cas d'une faible interaction et moins adaptée dans le cas d'une forte interaction.
Cette observation est directement liée à la condition d'optimalité de l'approche knock-out qui impose que le meilleur sous-ensemble de p coefficients contienne le meilleur sous-ensemble de $p-1$ coefficients. Cette condition est entièrement respectée si les coefficients sont indépendants les uns des autres, i.e. si l'ajout (ou le retrait) d'un coefficient dans un sous-ensemble n'interagit pas sur le comportement des coefficients de ce sous-ensemble. Dans ce cas précis, le sous-ensemble de coefficients sélectionnés est considéré comme optimal et nous pouvons parler d'optimum global atteint au cours des itérations.
Á l'opposé, dans le cas d'interactions entre coefficients, la méthode knock-out peut tendre vers un optimum local, réduisant ainsi l'optimalité du meilleur sous-ensemble sélectionné. L'algorithme génétique étant délié d'une telle condition d'optimalité, une forte interaction ne perturbe pas son fonctionnement.
La méthode ascendante étant très proche de la méthode knock-out, notamment au niveau de la condition d'optimalité, nous pouvons supposer que la comparaison : sélection ascendante et algorithme génétique aboutirait aux mêmes observations.

Pour résumer, il semblerait que les techniques de knock-out et ascendante soient les plus performantes dans le cas où leur condition d'optimalité est respectée, assurant une complexité connue *a priori* quelle que soit l'application visée. Les algorithmes génétiques sont une solution de remplacement envisageable dans le cas où la condition d'optimalité de ces dernières n'est pas respectée et que les données concernées amènent à une complexité de calcul raisonnable.

Dans le cadre de nos travaux, les coefficients concernés par la sélection ne respectent pas la condition d'optimalité. En outre, leur nombre pouvant être variable et conséquent, l'utilisation d'algorithmes génétiques ne nous semble pas non plus appropriée dans ce contexte. En effet, malgré les améliorations que pourraient apporter les algorithmes basés sur le recuit simulé, aucune garantie n'est donnée concernant la convergence vers un optimum global. Aussi, nous préférons choisir les techniques knock-out/ascendante, qui, malgré leur optimalité remise en question, nous assureront une complexité de calcul quantifiable.

Annexe B : Meilleurs ensembles de coefficients dynamiques.

Les sous-ensembles de coefficients retenus par la procédure de sélection (approche "dynamique") sont ici présentés suivant la base de données concernée – TIMIT ou Switchboard – et le critère de sélection employé : CritId, CritConf ou CritEmer.

	Vecteur 1	2	3	4	5	6	7	8	9	10
SB6	1-24	2-24	3-24	4-24	5-24	6-24	7-24	8-24	9-24	10-24
	1-23	2-23	3-23	4-23	5-23	6-23	7-23	8-23	9-23	10-23
	1-22	2-22	3-22	4-22	5-22	6-22	7-22	8-22	9-22	10-22
	1-21	2-21	3-21	4-21	5-21	6-21	7-21	8-21	9-21	10-21
SB5	1-20	2-20	3-20	4-20	5-20	6-20	7-20	8-20	9-20	10-20
	1-19	2-19	3-19	4-19	5-19	6-19	7-19	8-19	9-19	10-19
	1-18	2-18	3-18	4-18	5-18	6-18	7-18	8-18	9-18	10-18
	1-17	2-17	3-17	4-17	5-17	6-17	7-17	8-17	9-17	10-17
SB4	1-16	2-16	3-16	4-16	5-16	6-16	7-16	8-16	9-16	10-16
	1-15	2-15	3-15	4-15	5-15	6-15	7-15	8-15	9-15	10-15
	1-14	2-14	3-14	4-14	5-14	6-14	7-14	8-14	9-14	10-14
	1-13	2-13	3-13	4-13	5-13	6-13	7-13	8-13	9-13	10-13
SB3	1-12	2-12	3-12	4-12	5-12	6-12	7-12	8-12	9-12	10-12
	1-11	2-11	3-11	4-11	5-11	1-11	7-11	8-11	9-11	10-11
	1-10	2-10	3-10	4-10	5-10	6-10	7-10	8-10	9-10	10-10
	1-9	2-9	3-9	4-9	5-9	1-9	7-9	8-9	9-9	10-9
SB2	1-8	2-8	3-8	4-8	5-8	6-8	7-8	8-8	9-8	10-8
	1-7	2-7	3-7	4-7	5-7	6-7	7-7	8-7	9-7	10-7
	1-6	2-6	3-6	4-6	5-6	6-6	7-6	8-6	9-6	10-6
	1-5	2-5	3-5	4-5	5-5	1-5	7-5	8-5	9-5	1-5
SB1	1-4	2-4	3-4	4-4	5-4	6-4	7-4	8-4	9-4	10-4
	1-3	2-3	3-3	4-3	5-3	6-3	7-3	8-3	9-3	10-3
	1-2	2-2	3-2	4-2	5-2	6-2	7-2	8-2	9-2	10-2
	1-1	2-1	3-1	4-1	5-1	6-1	7-1	8-1	9-1	10-1

Figure 14.1: TIMIT : CritId. Ensemble des coefficients sélectionnés par le critère CritId sur TIMIT (63 hommes).

	Vecteur 1	2	3	4	5	6	7	8	9	10
SB6	1-24	2-24	3-24	4-24	5-24	6-24	7-24	8-24	9-24	10-24
	1-23	2-23	3-23	4-23	5-23	6-23	7-23	8-23	9-23	10-23
	1-22	2-22	3-22	4-22	5-22	6-22	7-22	8-22	9-22	10-22
	1-21	2-21	3-21	4-21	5-21	6-21	7-21	8-21	9-21	10-21
SB5	1-20	2-20	3-20	4-20	5-20	6-20	7-20	8-20	9-20	10-20
	1-19	2-19	3-19	4-19	5-19	6-19	7-19	8-19	9-19	10-19
	1-18	2-18	3-18	4-18	5-18	6-18	7-18	8-18	9-18	10-18
	1-17	2-17	3-17	4-17	5-17	6-17	7-17	8-17	9-17	10-17
SB4	1-16	2-16	3-16	4-16	5-16	6-16	7-16	8-16	9-16	10-16
	1-15	2-15	3-15	4-15	5-15	6-15	7-15	8-15	9-15	10-15
	1-14	2-14	3-14	4-14	5-14	6-14	7-14	8-14	9-14	10-14
	1-13	2-13	3-13	4-13	5-13	6-13	7-13	8-13	9-13	10-13
SB3	1-12	2-12	3-12	4-12	5-12	6-12	7-12	8-12	9-12	10-12
	1-11	2-11	3-11	4-11	5-11	1-11	7-11	8-11	9-11	10-11
	1-10	2-10	3-10	4-10	5-10	6-10	7-10	8-10	9-10	10-10
	1-9	2-9	3-9	4-9	5-9	1-9	7-9	8-9	9-9	10-9
SB2	1-8	2-8	3-8	4-8	5-8	6-8	7-8	8-8	9-8	10-8
	1-7	2-7	3-7	4-7	5-7	6-7	7-7	8-7	9-7	10-7
	1-6	2-6	3-6	4-6	5-6	6-6	7-6	8-6	9-6	10-6
	1-5	2-5	3-5	4-5	5-5	1-5	7-5	8-5	9-5	1-5
SB1	1-4	2-4	3-4	4-4	5-4	6-4	7-4	8-4	9-4	10-4
	1-3	2-3	3-3	4-3	5-3	6-3	7-3	8-3	9-3	10-3
	1-2	2-2	3-2	4-2	5-2	6-2	7-2	8-2	9-2	10-2
	1-1	2-1	3-1	4-1	5-1	6-1	7-1	8-1	9-1	10-1

Figure 14.2: TIMIT : CritConf. Ensemble des coefficients sélectionnés par le critère CritConf sur TIMIT (63 hommes).

SB3	1-24	2-24	3-24	4-24	5-24	6-24	7-24	8-24	9-24
	1-23	2-23	3-23	4-23	5-23	6-23	7-23	8-23	9-23
	1-22	2-22	3-22	4-22	5-22	6-22	7-22	8-22	9-22
	1-21	2-21	3-21	4-21	5-21	6-21	7-21	8-21	9-21
	1-20	2-20	3-20	4-20	5-20	6-20	7-20	8-20	9-20
	1-19	2-19	3-19	4-19	5-19	6-19	7-19	8-19	9-19
	1-18	2-18	3-18	4-18	5-18	6-18	7-18	8-18	9-18
	1-17	2-17	3-17	4-17	5-17	6-17	7-17	8-17	9-17
SB2	1-16	2-16	3-16	4-16	5-16	6-16	7-16	8-16	9-16
	1-15	2-15	3-15	4-15	5-15	6-15	7-15	8-15	9-15
	1-14	2-14	3-14	4-14	5-14	6-14	7-14	8-14	9-14
	1-13	2-13	3-13	4-13	5-13	6-13	7-13	8-13	9-13
	1-12	2-12	3-12	4-12	5-12	6-12	7-12	8-12	9-12
	1-11	2-11	3-11	4-11	5-11	1-11	7-11	8-11	9-11
	1-10	2-10	3-10	4-10	5-10	6-10	7-10	8-10	9-10
	1-9	2-9	3-9	4-9	5-9	1-9	7-9	8-9	9-9
SB1	1-8	2-8	3-8	4-8	5-8	6-8	7-8	8-8	9-8
	1-7	2-7	3-7	4-7	5-7	6-7	7-7	8-7	9-7
	1-6	2-6	3-6	4-6	5-6	6-6	7-6	8-6	9-6
	1-5	2-5	3-5	4-5	5-5	1-5	7-5	8-5	9-5
	1-4	2-4	3-4	4-4	5-4	6-4	7-4	8-4	9-4
	1-3	2-3	3-3	4-3	5-3	6-3	7-3	8-3	9-3
	1-2	2-2	3-2	4-2	5-2	6-2	7-2	8-2	9-2
	1-1	2-1	3-1	4-1	5-1	6-1	7-1	8-1	9-1
Vecteur	1	2	3	4	5	6	7	8	9

Figure 14.3: Switchboard : CritId. Ensemble des coefficients sélectionnés par le critère CritId sur Switchboard (50 femmes).

SB3	1-24	2-24	3-24	4-24	5-24	6-24	7-24	8-24	9-24
	1-23	2-23	3-23	4-23	5-23	6-23	7-23	8-23	9-23
	1-22	2-22	3-22	4-22	5-22	6-22	7-22	8-22	9-22
	1-21	2-21	3-21	4-21	5-21	6-21	7-21	8-21	9-21
	1-20	2-20	3-20	4-20	5-20	6-20	7-20	8-20	9-20
	1-19	2-19	3-19	4-19	5-19	6-19	7-19	8-19	9-19
	1-18	2-18	3-18	4-18	5-18	6-18	7-18	8-18	9-18
	1-17	2-17	3-17	4-17	5-17	6-17	7-17	8-17	9-17
SB2	1-16	2-16	3-16	4-16	5-16	6-16	7-16	8-16	9-16
	1-15	2-15	3-15	4-15	5-15	6-15	7-15	8-15	9-15
	1-14	2-14	3-14	4-14	5-14	6-14	7-14	8-14	9-14
	1-13	2-13	3-13	4-13	5-13	6-13	7-13	8-13	9-13
	1-12	2-12	3-12	4-12	5-12	6-12	7-12	8-12	9-12
	1-11	2-11	3-11	4-11	5-11	1-11	7-11	8-11	9-11
	1-10	2-10	3-10	4-10	5-10	6-10	7-10	8-10	9-10
	1-9	2-9	3-9	4-9	5-9	1-9	7-9	8-9	9-9
SB1	1-8	2-8	3-8	4-8	5-8	6-8	7-8	8-8	9-8
	1-7	2-7	3-7	4-7	5-7	6-7	7-7	8-7	9-7
	1-6	2-6	3-6	4-6	5-6	6-6	7-6	8-6	9-6
	1-5	2-5	3-5	4-5	5-5	1-5	7-5	8-5	9-5
	1-4	2-4	3-4	4-4	5-4	6-4	7-4	8-4	9-4
	1-3	2-3	3-3	4-3	5-3	6-3	7-3	8-3	9-3
	1-2	2-2	3-2	4-2	5-2	6-2	7-2	8-2	9-2
	1-1	2-1	3-1	4-1	5-1	6-1	7-1	8-1	9-1
Vecteur	1	2	3	4	5	6	7	8	9

Figure 14.4: Switchboard : CritId. Ensemble des coefficients sélectionnés par le critère CritId sur Switchboard (50 hommes).

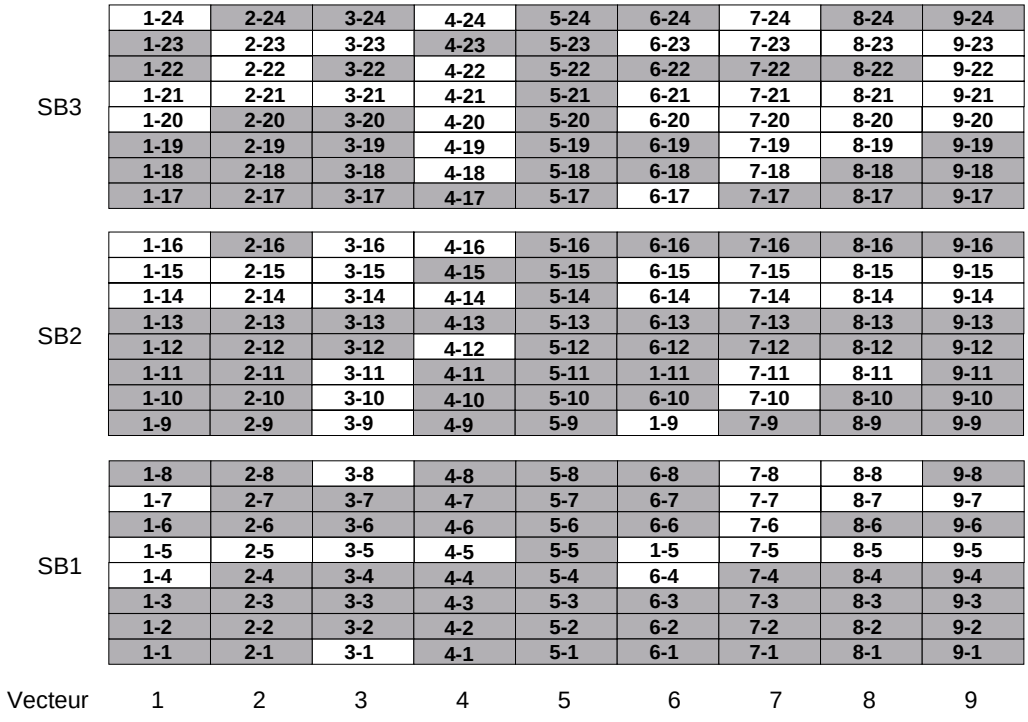

Figure 14.5: Switchboard : CritConf. Ensemble des coefficients sélectionnés par le critère CritConf sur Switchboard (50 femmes).

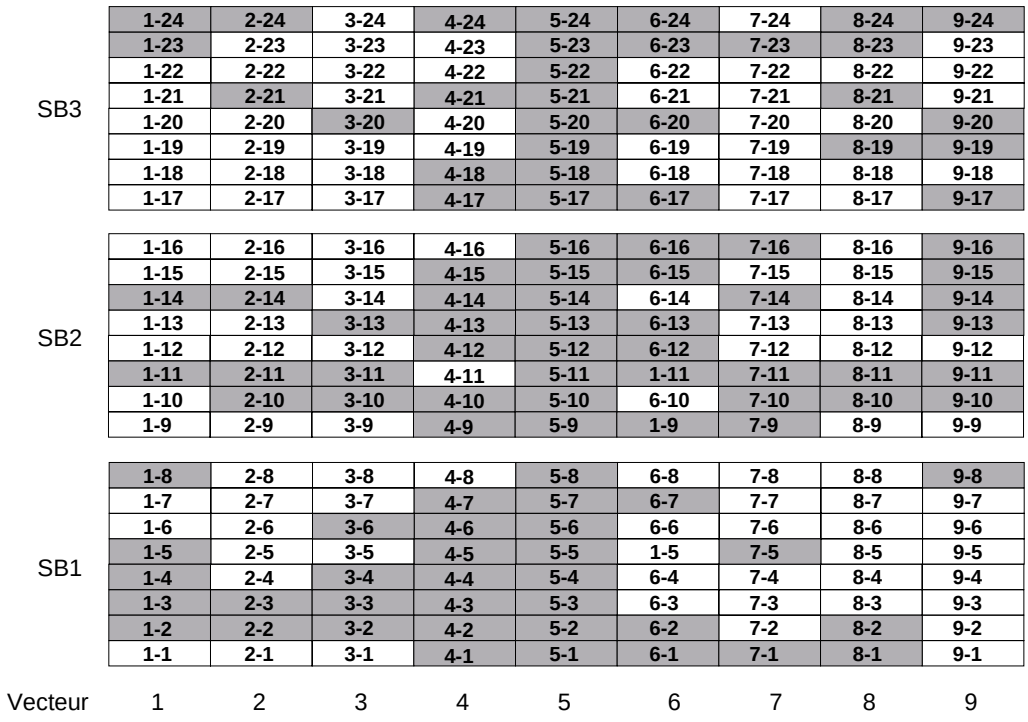

Figure 14.6: Switchboard : CritConf. Ensemble des coefficients sélectionnés par le critère CritConf sur Switchboard (50 hommes).

SB3	1-24	2-24	3-24	4-24	5-24	6-24	7-24	8-24	9-24
	1-23	2-23	3-23	4-23	5-23	6-23	7-23	8-23	9-23
	1-22	2-22	3-22	4-22	5-22	6-22	7-22	8-22	9-22
	1-21	2-21	3-21	4-21	5-21	6-21	7-21	8-21	9-21
	1-20	2-20	3-20	4-20	5-20	6-20	7-20	8-20	9-20
	1-19	2-19	3-19	4-19	5-19	6-19	7-19	8-19	9-19
	1-18	2-18	3-18	4-18	5-18	6-18	7-18	8-18	9-18
	1-17	2-17	3-17	4-17	5-17	6-17	7-17	8-17	9-17
SB2	1-16	2-16	3-16	4-16	5-16	6-16	7-16	8-16	9-16
	1-15	2-15	3-15	4-15	5-15	6-15	7-15	8-15	9-15
	1-14	2-14	3-14	4-14	5-14	6-14	7-14	8-14	9-14
	1-13	2-13	3-13	4-13	5-13	6-13	7-13	8-13	9-13
	1-12	2-12	3-12	4-12	5-12	6-12	7-12	8-12	9-12
	1-11	2-11	3-11	4-11	5-11	1-11	7-11	8-11	9-11
	1-10	2-10	3-10	4-10	5-10	6-10	7-10	8-10	9-10
	1-9	2-9	3-9	4-9	5-9	1-9	7-9	8-9	9-9
SB1	1-8	2-8	3-8	4-8	5-8	6-8	7-8	8-8	9-8
	1-7	2-7	3-7	4-7	5-7	6-7	7-7	8-7	9-7
	1-6	2-6	3-6	4-6	5-6	6-6	7-6	8-6	9-6
	1-5	2-5	3-5	4-5	5-5	1-5	7-5	8-5	9-5
	1-4	2-4	3-4	4-4	5-4	6-4	7-4	8-4	9-4
	1-3	2-3	3-3	4-3	5-3	6-3	7-3	8-3	9-3
	1-2	2-2	3-2	4-2	5-2	6-2	7-2	8-2	9-2
	1-1	2-1	3-1	4-1	5-1	6-1	7-1	8-1	9-1
Vecteur	1	2	3	4	5	6	7	8	9

Figure 14.7: Switchboard : CritEmer. Ensemble des coefficients sélectionnés par le critère CritEmer sur Switchboard (50 femmes).

SB3	1-24	2-24	3-24	4-24	5-24	6-24	7-24	8-24	9-24
	1-23	2-23	3-23	4-23	5-23	6-23	7-23	8-23	9-23
	1-22	2-22	3-22	4-22	5-22	6-22	7-22	8-22	9-22
	1-21	2-21	3-21	4-21	5-21	6-21	7-21	8-21	9-21
	1-20	2-20	3-20	4-20	5-20	6-20	7-20	8-20	9-20
	1-19	2-19	3-19	4-19	5-19	6-19	7-19	8-19	9-19
	1-18	2-18	3-18	4-18	5-18	6-18	7-18	8-18	9-18
	1-17	2-17	3-17	4-17	5-17	6-17	7-17	8-17	9-17
SB2	1-16	2-16	3-16	4-16	5-16	6-16	7-16	8-16	9-16
	1-15	2-15	3-15	4-15	5-15	6-15	7-15	8-15	9-15
	1-14	2-14	3-14	4-14	5-14	6-14	7-14	8-14	9-14
	1-13	2-13	3-13	4-13	5-13	6-13	7-13	8-13	9-13
	1-12	2-12	3-12	4-12	5-12	6-12	7-12	8-12	9-12
	1-11	2-11	3-11	4-11	5-11	1-11	7-11	8-11	9-11
	1-10	2-10	3-10	4-10	5-10	6-10	7-10	8-10	9-10
	1-9	2-9	3-9	4-9	5-9	1-9	7-9	8-9	9-9
SB1	1-8	2-8	3-8	4-8	5-8	6-8	7-8	8-8	9-8
	1-7	2-7	3-7	4-7	5-7	6-7	7-7	8-7	9-7
	1-6	2-6	3-6	4-6	5-6	6-6	7-6	8-6	9-6
	1-5	2-5	3-5	4-5	5-5	1-5	7-5	8-5	9-5
	1-4	2-4	3-4	4-4	5-4	6-4	7-4	8-4	9-4
	1-3	2-3	3-3	4-3	5-3	6-3	7-3	8-3	9-3
	1-2	2-2	3-2	4-2	5-2	6-2	7-2	8-2	9-2
	1-1	2-1	3-1	4-1	5-1	6-1	7-1	8-1	9-1
Vecteur	1	2	3	4	5	6	7	8	9

Figure 14.8: Switchboard : CritEmer. Ensemble des coefficients sélectionnés par le critère CritEmer sur Switchboard (50 hommes).

Annexe C : Bibliographie personnelle

Cette annexe fournit la liste des articles publiés durant la thèse.

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*, Editors J. Schroeder, J. Campbell, pages 172–197, 2000.

The ELISA consortium, The ELISA systems for the NIST 99 evaluation in speaker detection and tracking, *Digital Signal Processing (DSP), a review journal - Special issue on NIST 1999 speaker recognition workshop*, Editors J. Schroeder, J. Campbell, pages 143–153, 2000.

Meignier S., Bonastre J.-F., Fredouille C., Merlin T., "Modèle de Markov évolutif pour les tâches de suivi de locuteurs", XXIII Journée d'études sur la parole (JEP), 18-23 Juin, Aussois, France.

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

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

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

Besacier L., Bonastre J.-F., Fredouille C., Localization and selection of speaker specific information with statistical modelling, *Speech Communication*, pages 89–106, 2000.

Bonastre J.-F., Delacourt P., Fredouille C., Meignier S., Merlin T., Wellekens C. J., Différentes stratégies pour le suivi du locuteur, *Reconnaissance des Formes et Intelligence Artificielle (RFIA)*, pages 123–129, 2000, Paris (France).

Gravier G., Kharroubi J., Chollet G., Bimbot F., Blouet R., Seck M., Bonastre J.-F., Fredouille C., Merlin T., Pigeon S., Verlinde P., Cernocky J., Petrovska D., Nedic B., Magrin-Chagnolleau I., Durou G., The ELISA'99 speaker recognition and tracking, *Workshop on Automatic Identification Advanced Technologies (AutoId)*, Octobre 1999, Summit (USA).

Fredouille C., Hennebert J., Jaboulet C., Unsupervised incremental enrollment experiments - 99, *Rapport technique, Projet Européen PICASSO (Work-package 5)*, Octobre 1999, Ubilab/UBS, Zurich, Suisse.

Fredouille C., Bonastre J.-F., Merlin T., Similarity normalization method based on world model and a posteriori probability for speaker verification, *European Conference on Speech Communication and Technology (Eurospeech)*, pages 983–986, Septembre 1999, Budapest (Hongrie).

Merlin T., Bonastre J.-F., Fredouille C., "Non directly acoustic process for costless speaker recognition", *Proc. Workshop on intelligent communication technologies and applications, with emphasis on mobile communication*, 5-7 Mai 1999, Neuchâtel, Suisse.

Fredouille C., Bonastre J.-F., Merlin, T., Segmental normalization for robust speaker verification, *European Conference on Speech Communication and Technology (Eurospeech)*, pages 103–106, Mai 1999, Tampere (Finlande).

Fredouille C., Bonastre J.-F., Informations dynamiques et méthodes statistiques du second ordre pour l'identification du locuteur, *XXIIèmes Journées d'Etudes sur la Parole (JEP)*, pages 5–8, Juin 1998, Martigny (Suisse).

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

Besacier L., Bonastre J.-F., Fredouille C., Architecture en sous-bandes pour la reconnaissance automatique du locuteur en milieu bruit, *Reconnaissances des Formes et Intelligence Artificielle (RFIA)*, 20-22 Janvier 1998.

Bibliographie

[Afify et al., 1998] Afify M., Haton J.-P. Non-parametric segment models for automatic speaker identification. *Workshop on Speaker Recognition and its Commercial and Forensic Applications (RLA2C)*, pages 68–71, Avril 1998, Avignon (France).

[Aha et al., 1996] Aha D. W., Bankert R. L. A comparative evaluation of sequential feature selection algorithms. *Artificial Intelligence and Statistics*, 1996.

[Ariyaeeinia et al., 1997] Ariyaeeinia A. M., Sivakumaran P. Analysis and comparison of score normalisation methods for text-dependent speaker verification. *European Conference on Speech Communication and Technology (Eurospeech)*, Septembre 1997, Rhôdes (Grèce).

[Artières, 1995] Artières T. *Méthodes prédictives neuronales : application à l'identification du locuteur.* Thèse de doctorat, Université de Paris-Sud, U.F.R. Scientifique d'Orsay, 1995, Paris (France).

[Artières et al., 1991] Artières T., Bennani Y., Gallinari P., Montacié C. Connectionnist and conventional models for free-text talker identification tasks. *NEURONIMES*, 1991, Nîmes (France).

[Artières et al., 1993] Artières T., Gallinari P. Neural models for extracting speaker characteristics in speech modelization systems. *European Conference on Speech Communication and Technology (Eurospeech)*, pages 2263–2266, 1993, Berlin (Allemagne).

[Atal, 1974] Atal B. S. Effectiveness of LPC characteristics of the speech waves for A.S.I and A.S.V. *Journal of Acoustical Society of America (JASA)*, volume 55, 1974.

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

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

[Auckenthaler et al., 1997] Auckenthaler R., Mason J. S. Equalizing subband error rates in speaker recognition. *European Conference on Speech Communication and Technology (Eurospeech)*, pages 2303–2306, Septembre 1997, Rhôdes (Grèce).

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

[Bennani et al., 1991] Bennani Y., Gallinari P. On the use of TDNN-extracted features information in talker identification. *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 385–388, 1991, Toronto (Canada).

[Bennani et al., 1994] Bennani Y., Gallinari P. Connexionist approaches for automatic speaker recognition. *Workshop on Automatic Speaker Recognition, Identification, Verification*, pages 95–102, Avril 1994, Martigny (Suisse).

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

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

[Besacier, 1998] Besacier L. *Un modèle parallèle pour la reconnaissance automatique du locuteur.* Thèse de doctorat, Laboratoire Informatique d'Avignon (LIA), Université d'Avignon et des Pays de Vaucluse, 1998, Avignon (France).

[Besacier et al., 1997] Besacier L., Bonastre J.-F. *Subband approach for automatic speaker recognition : optimal division of the frequency domain.* Audio- and Video-based Biometric Person Authentication (AVBPA), Bigun, et.al.Eds., Springer LNCS 1206, 1997.

[Besacier et al., 1998a] Besacier L., Bonastre J.-F. Frame pruning for speaker recognition. *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1998a, Seatle (USA).

[Besacier et al., 1998b] Besacier L., Bonastre J.-F. Time and frequency pruning for speaker identification. *Workshop on Speaker Recognition and its Commercial and Forensic Applications (RLA2C)*, pages 106–110, Avril 1998b, Avignon (France).

[Besacier et al., 2000a] Besacier L., Bonastre J.-F., Fredouille C. Localization and selection of speaker specific information with statistical modelling. *Speech Communication*, volume 31, pages 89–106, 2000a.

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

[Bimbot et al., 1999] Bimbot F., Blomberg M., Boves L., Chollet G., Jaboulet C., Jacob B., Kharroubi J., Koolwaaij J., Lindberg J., Mariéthoz J., Mokbel C., Mokbel H. An overview of the Picasso project research activities in speaker verification for telephone applications. *European Conference on Speech Communication and Technology (Eurospeech)*, Septembre 1999, Budapest (Hongrie).

[Bimbot et al., 1997] Bimbot F., Genoud D. Likelihood ratio adjustment for the compensation of model mismatch in speaker verification. *European Conference on Speech Communication and Technology (Eurospeech)*, Septembre 1997, Rhôdes (Grèce).

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

[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, Août 1995.

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

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

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

[Bonastre, 1994] Bonastre J.-F. *Stratégie analytique orientée connaissances pour la caractérisation et l'identification du locuteur.* Thèse de doctorat, Université d'Avignon, 1994, Avignon (France).

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

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

[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, Décembre 1993.

[Bourlard et al., 1996] Bourlard H., Dupont S. A new ASR approach based on independent processing and combination of partial frequency bands. *International Conference on Spoken Language Processing (ICSLP)*, 1996, Philadelphia (USA).

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

[Calliope, 1989] Calliope. *La parole et son traitement automatique.* Collection technique et scientifique des télécommunications, Masson, 1989.

[Campbell et al., 1998] Campbell J. P., Reynolds D. A. Corpora for the Evaluation of Speaker Recognition Systems. *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1998, Seatle (USA).

[Carey et al., 1992] Carey M. J., Parris E. S. Speaker verification using connected words. *Proceedings of Institute of Acoustics*, volume 14(6), pages 95–100, 1992.

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

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

[Corsi, 1982] Corsi P. *Speaker recognition : a survey.* Automatic speech analysis and recognition, D. Reidel, Ed., 1982.

[De Veth et al., 1994] De Veth J., Bourlard H. Comparison of hidden Markov model techniques for automatic speaker verification. *Workshop on Automatic Speaker Recognition, Identification, Verification*, pages 11–14, Avril 1994, Martigny (Suisse).

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

[Demirekler et al., 1999] Demirekler M., Haydar A. Feature selection using genetics-based algorithm and its application to speaker identification. *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Mars 1999, Phoenix (USA).

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

[Doak, 1992] Doak J. An evaluation of feature selection methods and their application to computer security. Technical report cse-92-18, University of California, Departement of Computer Science, 1992.

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

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

[Dubreucq et al., 1994] Dubreucq V., Vloeberghs C. The use of the pitch to improve an HMM based speaker recognition method. *Workshop on Automatic Speaker Recognition, Identification, Verification*, pages 15–17, Avril 1994, Martigny (Suisse).

[Duchnowsky, 1993] Duchnowsky P. *A new structure for automatic speech recognition.* Thèse de doctorat, Massachussets Institute of Technology (MIT), 1993, Boston (USA).

[Duez, 1995] Duez D. On spontaneous french speech : aspects of the reduction and contextual assimilation of voiced plosives. *Journal of phonetics*, volume 25, 1995.

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

[ELISA, 2000] ELISA. The ELISA systems for the NIST 99 evaluation in speaker detection and tracking. *Digital Signal Processing (DSP), a review journal - Special issue on NIST 1999 speaker recognition workshop*, 10(1-3), 2000.

[Fant, 1973] Fant G. *Speech sounds and features.* MIT. press, Cambridge, 1973.

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

[Frederickson et al., 1994] Frederickson S. E., Tarassenko L. Radial basis functions for speaker identification. *Workshop on Automatic Speaker Recognition, Identification, Verification*, pages 107–110, Avril 1994, Martigny (Suisse).

[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, Avril 1998, Avignon (France).

[Fredouille et al., 1999] Fredouille C., Bonastre J.-F., Merlin T. Similarity normalization method based on world model and a posteriori probability for speaker verification. *European Conference on Speech Communication and Technology (Eurospeech)*, volume 2, pages 983–986, Septembre 1999, Budapest (Hongrie).

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

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

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

[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.

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

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

[Furui, 1997] Furui S. Recent advances in speaker recognition. *Audio, Video-based Biometric Person Authentication (AVBPA)*, pages 237–252, Mars 1997, Crans-Montana (Suisse).

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

[Genoud, 1999] Genoud D. *Reconnaissance et transformation de locuteurs.* Thèse de doctorat, Ecole polytechnique Fédérale de Lausanne (EPFL), 1999, Lausanne (Suisse).

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

[Gish et al., 1994] Gish H., Schmidt M. Text independent speaker identification. *IEEE Signal Processing Magazine*, pages 18–32, Octobre 1994.

[Goldberg, 1989] Goldberg D. E. *Genetic Algorithms in Search, Optimization, and Machine Learning.* Addison-Wesley, Reading, Mass, 1989.

[Gravier et al., 1998] Gravier G., Chollet G. Comparison of normalization techniques for speaker recognition. *Workshop on Speaker Recognition and its Commercial and Forensic Applications (RLA2C)*, pages 97–100, Avril 1998, Avignon (France).

[Gravier et al., 2000] Gravier G., Kharroubi J., Chollet G. On the use of prior knowledge in normalization schemes for speaker verification. *Digital Signal Processing (DSP), a review journal - Special issue on NIST 1999 speaker recognition workshop*, 10(1-3), 2000.

[Gravier et al., 1999] Gravier G., Kharroubi J., Chollet G., Bimbot F., Blouet R., Seck M., Bonastre J.-F., Fredouille C., Merlin T., Pigeon S., Verlinde P., Cernocky J., Petrovska D., Nedic B., Magrin-Chagnolleau I., Durou G. The ELISA'99 speaker recognition and tracking. *Workshop on Automatic Identification Advanced Technologies (AutoId)*, Octobre 1999, Summit (USA).

[Grenier, 1977] Grenier Y. *Identification du locuteur et adaptation au locuteur d'un système de reconnaissance phonétique.* Thèse de doctorat, Ecole Nationale Supérieure des Télécommunications (ENST), 1977, Paris (France).

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

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

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

[Hattori, 1994] Hattori H. Text-independent speaker verification using neural networks. *Workshop on Automatic Speaker Recognition, Identification, Verification*, pages 103–106, Avril 1994, Martigny (Suisse).

[Heck et al., 1997] Heck L. P., Weintraub M. Handset-dependent background models for robust text-independent speaker recognition. *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 1071–1074, 1997, Munich (Allemagne).

[Hermansky et al., 1994] Hermansky H., Morgan N. RASTA processing of speech. *IEEE Transaction on Speech and Audio Processing*, volume 2, pages 578–589, 1994.

[Hermansky et al., 1996] Hermansky H., Tibrewala S., Pavel M. Towards ASR on partially corrupted speech. *International Conference on Spoken Language Processing (ICSLP)*, 1996, Philadelphia (USA).

[Higgins et al., 1991] Higgins A. L., Bahler L., Porter J. Speaker verification using randomized phrase prompting. *Digital Signal Processing (DSP)*, volume 1, pages 89–106, 1991.

[Higgins et al., 1986] Higgins A. L., Wohlford R. E. A new method of text-independent speaker recognition. *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1986, Tokyo (Japan).

[Holland, 1975] Holland D. *Adaptation in Natural and Artificial Systems.* University of Michigan Press, Ann Harbor, 1975.

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

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

[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, Avril 1994, Martigny (Suisse).

[Isobe et al., 1999] Isobe T., Takhashi J.-I. A new cohort normalization using local acoustic information for speaker verification. *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Mars 1999, Phoenix (USA).

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

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

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

[Kharroubi et al., 2000] Kharroubi J., Chollet G. Utilisation de mots de passe personnalisés pour la vérification du locuteur. *XXIIIèmes Journées d'Etudes sur la Parole (JEP)*, pages 331–334, 2000, Aussois (France).

[Kittler et al., 1997] Kittler J., Li Y. P., Matas J., Ramos Sanchez M. U. Combining evidence in multi-modal personal identity recognition systems. *Audio, Video-based Biometric Person Authentication (AVBPA)*, pages 327–334, Mars 1997, Crans-Montana (Suisse).

[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, Avril 1998, Avignon (France).

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

[Le Floch et al., 1996] Le Floch J.-L., Montacié C., Caraty M.-J. Coopération et compétition de modèles en reconnaissance du locuteur. *XXIèmes Journées d'Etudes sur la Parole (JEP)*, pages 395–398, Juin 1996, Avignon (France).

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

[Li et al., 1995] Li H., Haton J.-P., Su J., Gong Y. Speaker recognition with temporal transition models. *European Conference on Speech Communication and Technology (Eurospeech)*, pages 617–620, Septembre 1995, Madrid (Espagne).

[Li et al., 1988] Li K. P., Porter J. E. Normalizations and selection of speech segments for speaker recognition scoring. *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 595–598, 1988.

[Lindberg et al., 1998] Lindberg J., Koolwaaij J., Hutter H.-P., Genoud D., Pierrot J.-B., Blomberg M., Bimbot F. Techniques for a priori decision threshold estimation in speaker verification. *Workshop on Speaker Recognition and its Commercial and Forensic Applications (RLA2C)*, pages 89–92, Avril 1998, Avignon (France).

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

[Liu et al., 1996] Liu C.-S., Wang H.-C., Lee C.-H. Speaker verification using normalized log-likelihood score. *IEEE Transaction on Speech and Audio Processing*, volume 4(1), pages 56–60, Janvier 1996.

[Liu et al., 1998] Liu H.-C., Huang J.-S. Pattern recognition using evolution algorithms with fast simulated annealing. Technical report, China Institute of Technology and Commerce, Department of Mechanical Engineering, 1998.

[Lummis et al., 1972] Lummis R. C., Rosenberg A. E. Test of an automatic speaker verification method with intensively trained professional mimics. *Journal of Acoustical Society of America (JASA)*, volume 51(1), page 131, 1972.

[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, Septembre 1999, Budapest (Hongrie).

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

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

[Martin et al., 1997] Martin A. F., Przybocki M. A. The DET curve in assessment of detection task performance. *European Conference on Speech Communication and Technology (Eurospeech)*, pages 1895–1898, Septembre 1997, Rhôdes (Grèce).

[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, 1989, Paris (France).

[Matrouf, 1997] Matrouf D. *Adaptation des modèles acoustiques pour la reconnaissance de la parole bruitée.* Thèse de doctorat, LIMSI, Université de Paris-Sud centre d'orsay, 1997, Paris (France).

[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, 1992, San Francisco (USA).

[Matsui et al., 1994a] Matsui T., Furui S. Similarity normalization method for speaker verification based on a Posteriori probability. *Workshop on Automatic Speaker Recognition, Identification, Verification*, pages 59–62, Avril 1994a, Martigny (Suisse).

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

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

[Mella, 1989] Mella O. Méthodologie d'étude de la pertinence des paramètres phonétiques et acoustiques pour la reconnaissance du locuteur. *Séminaire sur la variabilité et spécificité du locuteur : études et applications*, pages 196–199, Juin 1989, Marseille (France).

[Mokbel et al., 1998] Mokbel C., Jouvet D., Monné J., De Mori R. Robust speech recognition. *Spoken dialog with computers*, 1998.

[Mokbel et al., 1994] Mokbel C., Pachés-Leal P., Jouvet D., Monné J. Compensation of telephone line effects for robust speech recognition. *International Conference on Spoken Language Processing (ICSLP)*, pages 987–990, 1994, Yokohama (Japon).

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

[Montacié et al., 1992a] Montacié C., Le Floch J.-L. AR-Vector models for free-text speaker recognition. *International Conference on Spoken Language Processing (ICSLP)*, pages 611–614, Octobre 1992a, Banff (Canada).

[Montacié et al., 1993] Montacié C., Le Floch J.-L. Discriminant AR-Vector models for free-text speaker verification. *European Conference on Speech Communication and Technology (Eurospeech)*, pages 161–164, 1993, Berlin (Allemagne).

[Montacié et al., 1992b] Montacié C., Le Floch J.-L., Rodet X. Modèles Auto-regréssifs Vectoriels et reconnaissance du locuteur. *XIXèmes Journées d'Etudes sur la Parole (JEP)*, pages 439–442, 1992b, Bruxelles (Belgique).

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

[Nolan, 1983] Nolan F. J. *The phonetic bases of speaker recognition.* Cambridge, Cambrigde University Press, 1983.

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

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

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

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

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

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

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

[Pruzansky et al., 1964] Pruzansky S., Mathews M. V. Talker-recognition procedure based on analysis of variance. *Journal of Acoustical Society of America (JASA)*, volume 36(1), pages 2041–2047, 1964.

[Przybocki et al., 1998] Przybocki M. A., Martin A. F. NIST speaker recognition evaluation - 97. *Workshop on Speaker Recognition and its Commercial and Forensic Applications (RLA2C)*, pages 120–123, Avril 1998, Avignon (France).

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

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

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

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

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

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

[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.

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

[Rosenberg et al., 1992] Rosenberg A. E., DeLong J., Lee C.-H., Juang B.-H., Soong F. K. The use of cohort normalized scores for speaker verification. *International Conference on Spoken Language Processing (ICSLP)*, pages 599–602, Octobre 1992, Banff (Canada).

[Rosenberg et al., 1994] Rosenberg A. E., Lee C.-H., Soong F. K. Cepstral channel normalization techniques for HMM-based speaker verification. *International Conference on Spoken Language Processing (ICSLP)*, pages 1835–1838, 1994, Yokohama (Japon).

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

[Rosenberg et al., 1996] Rosenberg A. E., Parthasarathy. Speaker background models for connected digit password speaker verification. *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 81–84, 1996, Atlanta (USA).

[Rosenberg et al., 1998b] Rosenberg A. E., Siohan O., Parthasarathy S. Small group speaker identification with common password phrases. *Workshop on Speaker Recognition and its Commercial and Forensic Applications (RLA2C)*, pages 190–193, Avril 1998b, Avignon (France).

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

[Sambur, 1975] Sambur M. R. Selection of acoustic features for speaker identification. *IEEE Transactions Acoustics, Speech, and Signal Processing (ASSP)*, volume 23(2), pages 176–182, Avril 1975.

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

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

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

[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, 1992, Tampa (USA).

[Su et al., 1974] Su L.-S., Li K. P., Fu K. S. Identification of speaker by use of nasal coarticulation. *Journal of Acoustical Society of America (JASA)*, volume 56, pages 1867–1882, 1974.

[Traunmuller, 1984] Traunmuller H. Articulatory and perceptual factors controlling the age and sex conditionned variability in formant frequencies vowels. *Speech Communication*, volume 3, pages 49–61, 1984.

[Vafaie et al., 1993] Vafaie H., De Jong K. Robust feature selection algorithms. *Fifteenth International Conference on Tools for Artificial Intelligence*, pages 356–363, 1993, Boston (USA).

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

[Waibel et al., 1988] Waibel A., Hanazawa T., Hinton G., Shikano K., Lang K. Phoneme recognition : neural networks vs. Hidden Markov Models. *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 107–110, 1988.

[Wolf, 1972] Wolf J. J. Efficient acoustic parameters for speaker recognition. *Journal of Acoustical Society of America (JASA)*, volume 51(6.2), pages 2044–2054, 1972.

[Xu et al., 1989] Xu L., Mason J. S. Instantaneous and transitional perceptually-based features in speaker identification. *European Conference on Speech Communication and Technology (Eurospeech)*, pages 271–274, 1989, Paris (France).

[Yang et al., 1997] Yang J., Honavar V. Feature subset selection using a genetic algorithm. *IEEE Expert, Special issue on Feature transformation and subset selection*, 1997.

[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*, 1995, Berlin (Allemagne).