Résidu de prédiction linéaire et reconnaissance de locuteurs indépendante du texte

Ph. Thévenaz

IMPRIMATUR POUR LA THÈSE

Résidu de prédiction linéaire et reconnaissance de locuteurs indépendante du texte

de Monsieur Philippe Thévenaz

UNIVERSITÉ DE NEUCHÂTEL

FACULTÉ DES SCIENCES

La Faculté des sciences de l'Université de Neuchâtel sur le rapport des membres du jury,

MM. H. Hügli, F. Pellandini,

F. Grosjean et M. Kunt (EPF-Lausanne)

autorise l'impression de la présente thèse.

Neuchâtel, le 5 août 1993

Le doyen :

A. Robert

Épigraphe

MAÎTRE DE PHILOSOPHIE. — Soit. Pour bien suivre votre pensée et traiter cette matière en philosophe, il faut commencer, selon l'ordre des choses, par une exacte connaissance de la nature des lettres et de la différente manière de les prononcer toutes. Et là-dessus j'ai à vous dire que les lettres sont divisées en voyelles, ainsi dites voyelles parce qu'elles expriment les voix; et en consonnes, ainsi appelées consonnes parce qu'elles sonnent avec les voyelles, et ne font que marquer les diverses articulations des voix. Il y a cinq voyelles ou voix: A, E, I, O, U.

MONSIEUR JOURDAIN. — J'entends tout cela.

MAÎTRE DE PHILOSOPHIE. — La voix A se forme en ouvrant fort la bouche: A.

MONSIEUR JOURDAIN. — A, A, oui.

MAÎTRE DE PHILOSOPHIE. — La voix E se forme en rapprochant la mâchoire d'en bas de celle d'en haut: A, E.

MONSIEUR JOURDAIN. — A, E; A, E. Ma foi, oui. Ah! que cela est beau!

MAÎTRE DE PHILOSOPHIE. — Et la voix I, en rapprochant encore davantage les mâchoires l'une de l'autre, et écartant les deux coins de la bouche vers les oreilles: A, E, I.

MONSIEUR JOURDAIN. — A, E, I, I, I, I, I. Cela est vrai. Vive la science!

Molière, le Bourgeois Gentilhomme, Acte II, Scène IV.

Table des matières

Résumé

Ce travail de thèse est une contribution à la résolution du problème de la reconnaissance automatique de locuteurs dans les conditions particulières où l'indépendance du texte est exigée. La tâche que nous y entreprenons consiste à vérifier la prétention d'identité d'un locuteur sur la base de sa voix seule.

La motivation de ce travail est la recherche d'un ensemble de méthodes de reconnaissance dont la combinaison soit caractérisée par un faible taux d'erreur. Comme il est nécessaire que leurs contributions soient indépendantes si l'on veut que cette combinaison soit fructueuse, nous examinons une collection de méthodes tant connues que nouvelles telles que l'on puisse supposer que leurs apports soient de nature différente. En particulier, nous avons fondé nos espoirs sur la combinaison de méthodes exploitant séparément les deux composantes majeures de l'analyse par prédiction linéaire, à savoir le filtre de synthèse d'une part et le résidu d'autre part; constatons que ce dernier a été délaissé jusqu'à aujourd'hui par les autres chercheurs.

Les méthodes les plus connues de reconnaissance de locuteurs indépendantes du texte sont celle du cepstre complexe moyen du filtre de synthèse, celle de l'erreur moyenne de quantification vectorielle du cepstre complexe du filtre de synthèse, celle de l'erreur moyenne de quantification vectorielle du cepstre complexe différentiel du filtre de synthèse et celle de la fréquence fondamentale moyenne. Une partie de ce travail de thèse est dévolue à la répétition de ces expériences classiques de reconnaissance.

Nous ajoutons à ces approches deux innovations principales. La première est une technique générale de reconnaissance, au même titre que l'erreur moyenne de quantification vectorielle qu'elle prétend compléter. Nous la baptisons conformité et nous la testons dans le contexte du cepstre complexe du filtre de synthèse; il en résulte un succès de reconnaissance meilleur que celui associé à la méthode du cepstre complexe moyen du filtre de synthèse. La seconde innovation porte sur le choix d'un vecteur caractéristique nouveau appelé cepstre

réel du résidu de l'analyse par prédiction linéaire. Utilisé conjointement avec une méthode inédite de reconnaissance de locuteurs indépendante du texte appelée proéminence, ce résidu s'avère d'une efficacité comparable à celle des méthodes basées sur la fréquence fondamentale tout en exigeant un nombre plus petit d'ajustements de paramètres de calcul.

Les méthodes nouvelles de reconnaissance de locuteurs indépendantes du texte testées sont donc celle de la conformité du cepstre complexe du filtre de synthèse, celle du cepstre réel moyen du résidu, celle de l'erreur moyenne de quantification vectorielle du cepstre réel du résidu et celle de la proéminence du cepstre réel du résidu.

Notre recherche de méthodes efficaces est guidée par l'expérimentation. Nous avons construit à cet effet deux bases de données propices à la reconnaissance de locuteurs dans des conditions où l'indépendance du texte est requise. La comparaison de toutes les méthodes de reconnaissance est alors facilitée par le fait que nous utilisons nos propres bases de données, dans les mêmes conditions d'expériences, et selon une méthodologie identique.

Enfin, nous sommes en mesure de montrer dans cette thèse comment réduire les taux d'erreur associés aux méthodes séparées de reconnaissance, à condition de les utiliser conjointement. Notamment, la combinaison d'un membre de la famille de celles qui exploitent le filtre de synthèse avec un membre de la famille de celles qui exploitent le résidu est particulièrement bénéfique car ces deux sources de vecteurs caractéristiques sont de nature différente. Cette indépendance est alors garante d'un succès accru pour la combinaison de plusieurs méthodes de reconnaissance de locuteurs indépendantes du texte qui, séparément, les utilisent avec succès.

Abstract

We address in this thesis the problem of automatic speaker recognition. In particular, we select the task of speaker verification as our working paradigm, in which we examine the agreement between an identity claim and a voice sample. Furthermore, our contribution stands in a text independence context.

This work is motivated by the search of recognition techniques that can be advantageously combined. As it is necessary to consider techniques that offer independent contributions in order to achieve a high combination efficiency, we examine several ones, old and new as well, such that we can assume that their contributions differ one another. Especially, we emphasise combining techniques which make a separate use of the two principal components from linear prediction analysis, namely the synthesis filter on the one hand and the residue on the other hand; incidentally, the latter has been overlooked until now by most other researchers.

The most well known techniques for text-independent speaker recognition are the average complex cepstrum of the synthesis filter, the average vector quantization error of the synthesis filter complex cepstrum, the average vector quantization error of the synthesis filter differential complex cepstrum and the average fundamental frequency. One part of this thesis work is devoted to the remake of these classical recognition experiments.

We add to these techniques two major innovations. The first one is a general recognition technique, in the same sense as the average vector quantization error is; furthermore, both are complementary. We call our new technique conformity and try it in the context of the synthesis filter complex cepstrum; the resulting error rate is lower than that associated with the technique of the average complex cepstrum of the synthesis filter. The second innovation consists in a new characteristic vector named real cepstrum of the linear prediction analysis residue. Used jointly with a novel text-independent speaker recognition technique named prominence, the residue proves to be as efficient for speaker

recognition as techniques based upon fundamental frequency, while requiring less fine-tuning of parameters.

The new text-independent speaker recognition techniques we test in this thesis are then the conformity of the synthesis filter complex cepstrum, the average real cepstrum of the residue, the average vector quantization error of the residue real cepstrum and the prominence of the residue real cepstrum.

Our quest for efficient speaker recognition techniques is guided through experimentation. To this end, we built two databases fitting the task of text-independent speaker recognition. With the aid of these very databases we realised a fair comparison of all speaker recognition techniques by using the same methodology in similar experimental conditions.

Finally, we are in a position to show how to reduce the error rates associated with several separate recognition techniques, provided one uses them jointly. Especially, the combination of a member from the family of techniques making use of the synthesis filter together with a member from the family of those making use of the residue is particularly propitious because the signal sources for characteristic vectors are of a different nature. This independence then guarantees an enhanced success in the combination of techniques that would use them efficiently.

Zusammenfassung

Diese Dissertation soll ein Beitrag zur Lösung des Problems des textunabhängigen automatischen Erkennens einer Stimme sein. Ziel ist es, daß die aufgegebene Identität des Sprechers schon aufgrund einer Stimmprobe verifiziert werden kann.

Diese Arbeit wurde durch die Forschung nach einer Erkennungstechnik, die sich später aus der Kombination von verschiedenen Komponenten zusammensetzt, motiviert. Um eine hohe Leistungsfähigkeit zu erreichen, untersuchten wir verschiedene alte, wie auch neue Techniken, so daß wir annehmen können, daß ihre Eigenschaften voneinander unabhängig sind. Im speziellen stützen wir uns auf die Kombination von Methoden die je einzeln auf die beiden Hauptfaktoren der linearen prädiktiven Analyse, namentlich einerseits dem Synthesefilter und dem Residuum andererseits, wirken, wobei letzteres bis heute von den meisten Forschern ausgelassen wurde.

Die wohl am besten bekannten Techniken des textunabhängigen Erkennens einer Stimme sind die des Mittelwertes des komplexen Cepstrums des Synthesefilters, die des akkumulierten Fehlers bei der Codierung mit Vektorquantisierung des komplexen Cepstrums des Synthesefilters, die Technik des akkumulierten Fehlers bei der Codierung mit Vektorquantisierung des komplexen Differentialcepstrums des Synthesefilters und die des Mittelwertes der Sprachgrundfrequenz. Ein Teil dieser Dissertationsarbeit widmet sich der Wiederholung der Erkenntnisse aus diesen klassischen Forschungen.

Zu diesen Erkenntnissen setzen wir zwei grundlegende Neuerungen. Die erste ist eine allgemeine Technik der Anerkennung, im gleichen Sinne wie der akkumulierte Fehler bei der Codierung mit Vektorquantisierung; die beide wirken vervollständigend. Wir nennen sie Anpassung und testen sie im Kontext des komplexen Cepstrums des Synthesefilters. Die Fehlerrate liegt tiefer als die bei der Verbindung mit der Methode des Mittelwertes des komplexen Cepstrums. Die zweite Erneuerung ist ein neues charakteristisches Vektor, genannt

Cepstrum des Residuums der linearen prädiktiven Analyse. Gebraucht zusammen mit der neuen Methode der textunabhängigen Sprechererkennung, die Prominenz genannt ist, erweist sich das Residuum als ebenso effizient wie die Methode des Mittelwertes der Sprachgrundfrequenz, wobei weit weniger feine Anpassungen der Parameter notwendig sind.

Die in dieser Dissertation von uns getestete neue textunabhängige Sprechererkennungstechnik sind dann die Anpassung des komplexen Cepstrums des Synthesefilters, das Mittelwert des reellen Cepstrums des Residuums, der akkumulierte Fehler bei der Codierung mit Vektorquantisierung des reellen Cepstrums des Residuums und die Prominenz des reellen Cepstrums des Residuums.

Unsere Forschungen nach einer effizienten Methode war von Experimenten begleitet. Zu diesem Zweck konstruierten wir zwei Databasen die auf die textunabhängige Sprechererkennung abgestimmt sind. Der Vergleich der verschiedenen Methoden ist also dadurch vereinfacht, da wir unsere eigene Datenbasis verwenden können, unter den gleichen Bedingungen und der gleichen Methodik.

Schließlich können wir aufzeigen wie die Fehlerraten, die die einzelnen Methoden aufweisen, durch Zusammenfügen verringert werden können. Demnach ist die Kombination eines Teiles der Familie, die den Synthesefilter ausbautet, und eines Teiles der Familie, die das Residuum ausbautet, äußerst günstig, da die beiden charakteristischen Vektoren von verschiedener Natur sind. Diese Unabhängigkeit ist daher Garantie zum Erfolg einer Kombination aus verschiedenen Methoden der textunabhängigen Sprechererkennung, die doch einzeln mit großem Erfolg verwendet werden können.

1 Introduction

«Sésame, ouvre-toi!»

Quelques mots, et déjà apparaît l'univers de la fable et du merveilleux. Ali-Baba et les quarante voleurs n'est-il pas un des contes les plus connus parmi les mille et uns que Shéhérazade récita au calife, ses nuits de veille? Le mot de passe qu'il contient, sans qui la trame de l'histoire n'existerait pas, est certainement le mot de passe secret le plus public de toute la planète... «Zesam öffne dich! Apriti sesamo! Open sesame! Abrete sesamo! Zhimakaimenba! Sesam open U!»

Le sens véhiculé par un mot de passe n'est pas apparenté à son contenu; qu'il soit incohérent ou non importe peu. Ce qui importe, c'est qu'il révèle l'identité de celui qui le prononce, c'est qu'il certifie l'identité du porteur du message vis-à-vis de l'auditeur. Par exemple, de nos jours encore l'acquisition, la connaissance et la transmission de mots de passes est une activité sans laquelle la société secrète des francs-maçons perdrait sa raison d'être, eu égard au mythe fondateur de Hiram Abiff.

Quittons les temps anciens et abordons l'époque moderne; nous pouvons constater que le mot de passe est aussi amené à y jouer un rôle en se substituant aux gardiens actuels de l'identité tels que clefs, signatures, papiers officiels, badges, cartes diverses à code magnétique ou à puce, etc. En effet, si le texte même du mot de passe n'offre aucune sécurité, pourtant la voix de celui qui le prononce ne peut être ni transmise, ni oubliée; de plus, la parole est un moyen de communication naturel de l'homme et ne demande pas de sa part un effort particulier. L'existence d'un dispositif permettant de reconnaître la voix d'un locuteur est donc envisageable; l'imagerie moderne a incorporé semblable dispositif dans le film cinématographique visionnaire du réalisateur Stanley Kubrick «2001, A Space Odyssey», où le héros franchit les contrôles douaniers sur présentation de sa voix [68KubS]. Ce film datant de près de vingt-cinq ans est un des premiers à traiter le thème de la science-fiction avec une ampleur pareille, et une volonté certaine de réalisme (intrigue mise à part).

La science-fiction et les histoires fabuleuses deviennent, parfois, réalité. C'est à regretter quand c'est le Méchant Dragon qui prend chair, ou à applaudir quand c'est le Bon Génie qui s'incarne. Ce travail de thèse a pour ambition l'amélioration de l'état des connaissances permettant de faciliter l'identification de personnes. Nous espérons sincèrement que notre contribution représente un coup de brosse supplémentaire sur la lanterne magique d'Aladin, et non pas un croc de plus dans une gueule déjà trop béante.

1.1 Généralités

Le titre de ce mémoire de thèse mentionne la prédiction linéaire: il s'agit d'une technique classique d'analyse de parole que nous allons utiliser et dont nous parlerons à l'envi tout au long de ce mémoire; le résidu est un des éléments de cette analyse. Il mentionne aussi la reconnaissance de locuteurs indépendante du texte: il s'agit là d'une tâche dont la réalisation automatique est difficile.

1.1.1 But

Le but de cette thèse est de proposer une solution au problème qui consiste à déterminer l'identité d'un locuteur sur la base de sa voix uniquement. Nous ignorerons donc les autres caractéristiques intrinsèques de l'individu telles que les empreintes digitales, rétiniennes, dentaires, génétiques, le poids, la taille, la pigmentation oculaire, capillaire, cutanée, etc. Par contre, nous nous permettrons de profiter de ses caractéristiques extrinsèques telles que par exemple un numéro d'identification personnel, moins comme aide à la reconnaissance que comme hypothèse à valider ou infirmer; lorsque cette information sera disponible nous parlerons d'une tâche de vérification de locuteur, par opposition à une tâche d'identification de locuteur, qui n'en fait pas usage.

1.1.2 Techniques

Les techniques dont nous ferons usage seront de nature strictement automatique. Notre approche exclut toute intervention d'un jugement humain; nous prétendons ainsi à une reconnaissance objective. De plus, nous ne nous intéresserons pas au contenu du message transporté par la parole; seul son aspect acoustique interviendra. De ce point de vue, notre approche fait partie des techniques de reconnaissance de locuteurs indépendantes du texte.

1.1.3 Moyens

Les outils que nous exploiterons font partie du domaine du traitement de signal; simultanément, les expériences que nous allons entreprendre utiliseront l'approximation de ces outils qu'offre un calculateur numérique. Nous déciderons pourtant de ne pas tenir compte des limitations introduites par la nature finie des ressources disponibles (essentiellement temps de calcul et capacité de stockage); ce dernier point sera toutefois tempéré par le bon sens, eu égard à l'état actuel de la technologie disponible pour un coût supportable. En outre, bien que la cadence d'échantillonnage soit modeste, choisie de sorte à refléter les conditions d'acquisition rencontrées le plus souvent en pratique, nous négligerons les effets des limitations introduites par ce procédé d'approximation (non seulement diminution de la bande passante inhérente au procédé d'échantillonnage, mais encore et surtout quantification).

1.2 Synopsis

Ce mémoire de thèse est découpé en neuf parties principales faisant toutes l'objet d'un chapitre séparé; cette synopsis a pour but de les présenter brièvement.

- Dans ce premier chapitre, nous introduisons le problème et nous abordons en langage naturel les méthodes propres à sa solution.
- Dans le deuxième chapitre, nous traitons de l'aspect formel de certaines des techniques mises en œuvre. On y trouve le modèle de vérification étudié ainsi que la formalisation de deux tâches importantes de la reconnaissance, constituées l'une par la construction de vecteurs représentatifs et l'autre par le processus de comparaison. Nous terminons ce chapitre avec la formalisation des taux d'erreur associés aux diverse tâches de reconnaissance.
- Dans le troisième chapitre, nous traitons des vecteurs caractéristiques sur lesquels nous fondons la reconnaissance de locuteurs. En particulier, nous exposons les détails de la technique de l'analyse par prédiction linéaire telle que nous l'avons réalisée, suivie de la technique de l'extraction de la fréquence fondamentale de sons laryngés.
- Dans le quatrième chapitre, nous éclairons quelques aspects de l'état de l'art de la reconnaissance d'individus. Plus particulièrement, nous abordons leur reconnaissance sur une base vocale à la fois d'un point de vue humain et d'un point de vue automatique; dans ce dernier cas nous examinons plus en détail les comptes rendus de deux expériences classiques de reconnaissance de locuteurs indépendante du texte.

- Dans le cinquième chapitre, nous décrivons les deux bases de données que nous avons été amené à construire pour la réalisation de cette thèse; en outre, nous y exposons nos choix quant à la façon de les exploiter.
- Dans le sixième chapitre, nous discutons le détail des conditions et des résultats des expériences que nous avons calquées sur certaines de celles décrites au troisième chapitre, mais que nous avons menées sur nos propres bases de données.
- Dans le septième chapitre, nous présentons notre contribution majeure qui s'articule en deux éléments. Il s'agit d'une part de la méthode dite de la conformité qui complète une méthode existante appelée erreur moyenne de quantification vectorielle; d'autre part, il s'agit d'un vecteur caractéristique inédit en reconnaissance de locuteurs nommé cepstre réel du résidu de l'analyse par prédiction linéaire. L'efficacité de ces innovations est confrontée au feu de l'expérience, puis nous montrons encore comment combiner ces méthodes pour améliorer la qualité de la reconnaissance.
- Dans le huitième chapitre, nous offrons au lecteur quelques spéculations de notre cru sur les éléments pertinents que nous n'avons pas pu examiner en pratique.
- Enfin, nous achevons ce mémoire de thèse par le neuvième et dernier chapitre où nous présentons nos conclusions.

1.3 Exposé du problème

Ce paragraphe aborde la transition entre la reconnaissance humaine et la reconnaissance automatique. Nous commencerons par montrer que la recherche d'une solution au problème de la reconnaissance automatique de locuteurs n'est pas vaine, puis nous esquisserons les deux méthodes principales que l'on rencontre dans la littérature.

1.3.1 Capacité de différentiation

Le premier point à établir est équivalent au théorème d'existence du mathématicien qui vient d'introduire un nouvel objet. Cette question existentielle peut se formuler ainsi: est-il possible de reconnaître un locuteur grâce à sa voix seule? La réponse est bien plus facile à donner aujourd'hui qu'autrefois. En effet, par le passé seule une volonté délibérée d'expérimentation et un protocole rigoureux aurait pu trancher cette question pour un esprit suspicieux, tandis que de nos jours cet esprit suspicieux n'existe certainement plus, tant est que chacun a, un jour ou l'autre, empoigné l'appareil nommé téléphone et reconnu son cor-

respondant avant même qu'il ne se présente formellement. Les émissions radiodiffusées offrent un autre exemple contemporain où chacun reconnaît la voix de son annonceur préféré, ou haï.

1.3.2 Performances humaines

La possibilité de reconnaître un locuteur sur la seule base de sa parole est donc avérée, comme il est vrai que certains athlètes parviennent à courir cent mètres en moins de dix secondes. Toutefois, ce dernier cas peut être considéré comme un exploit, ou à tout le moins une affaire de spécialistes entraînés. Au contraire, la reconnaissance de locuteurs paraît être un fait banal; on ne trouvera que peu de personnes niant être capables de reconnaître qui que ce soit au téléphone, même parmi leurs proches. De plus, l'attribution au locuteur d'une identité par l'auditeur est souvent immédiate, et ne paraît pas requérir d'habileté particulière. Or, si vous demandez à quelqu'un qu'il vous décrive une voix avec une précision suffisante pour vous permettre de la reconnaître, alors vous vous apercevrez que cet exercice est fort malaisé. Une description grossière est possible grâce à certains qualificatifs tels que voix rauque, claire, acide, douce, cassée, pleine, métallique, chaude, monotone, chantante, grave, aiguë, chevrotante, posée, faible, puissante, légère, profonde, chuintante, zézayante, flûtée, chuchotée, fluette, grasseyante, nasillarde, mielleuse, susurrante, enrhumée et quelques autres [64VoiW]. Cependant, cette description est sujette à interprétation; deux auditeurs ne s'accommoderont pas toujours de la même frontière entre une voix puissante et une voix faible, par exemple. En outre, une description de voix est souvent accompagnée de considérations sur des aspects qui ne sont pas strictement acoustiques, tels par exemple le vocabulaire choisi par le locuteur, ou sa provenance trahie par l'accent. En conclusion, si l'homme est apte à reconnaître une voix, il ne sait pas pour autant la décrire avec une efficacité suffisante pour caractériser la voix reconnue vis-à-vis d'un tiers. Il en résulte l'impossibilité de simuler par un dispositif automatique la façon précise dont un auditeur reconnaît un locuteur, les détails de cette reconnaissance restant inaccessibles et par conséquent trop mystérieux.

1.3.3 Degrés de dépendance du texte

Nous avons introduit le concept de locuteur et celui d'indépendance du texte. En réalité, si un locuteur est un individu qui émet de la parole, alors par définition il perd sa qualité de locuteur s'il se met à chanter ou à siffler; nous en concluons que l'indépendance du texte recouvre quelque chose de plus précis qu'une liberté d'expression orale totale. Essayons de délimiter les frontières des

divers degrés de dépendance que pourrait revêtir un dispositif de reconnaissance de locuteur.

Mot de passe: le texte prononcé par le locuteur est invariable. La machine a connaissance de ce texte et de sa prononciation particulière par l'individu dont elle examine l'identité.

Texte imposé: le texte prononcé par le locuteur est dicté par la machine; par contre, cette dernière n'a pas forcément connaissance de sa prononciation particulière par l'individu dont elle examine l'identité [91HigA1]. Elle s'attend toutefois à y découvrir des éléments remarquables en des points précis, comme par exemple une suite donnée de sons laryngés.

Vocabulaire restreint: le locuteur choisit les mots du texte qu'il prononce dans un ensemble fini, connu de la machine; par contre, cette dernière n'a pas forcément connaissance de la séquence de ces choix. Ce cas peut être ramené à celui du texte imposé pour autant que l'on admette que le dispositif traite le signal de parole d'abord sous l'angle de la reconnaissance de texte avant de le traiter sous l'angle de la reconnaissance de locuteur.s

Indépendance du texte: le locuteur a toute liberté quant au choix des mots qu'il prononce. On considère néanmoins d'une part que le locuteur montre une certaine coopération dans son comportement, et d'autre part que l'effet des perturbations liées à l'environnement du locuteur est négligeable.

Indépendance totale: l'unicité du locuteur est ici mise en question car on tolère maintenant le fait que plusieurs personnes parlent en séquences courtes, ou même simultanément. On accepte aussi les manifestations d'émotion telles que les cris, les rires, les pleurs. On s'accommode encore de la pollution du matériel sonore par des bruits d'origines diverses telles que bruits de circulation, musique, clapotis de vagues ou autres grondements de cascades. On supporte les manifestations non verbales du locuteur telles que sifflements, claquements de langue, râles, chants, bâillements, toux, raclements de gorge, soupirs, éternuements, éructations, dents qui s'entrechoquent. Enfin, il n'est plus forcément tenu pour acquis que le milieu de transmission est l'air, mais on est prêt à considérer les effets de l'atmosphère riche en hélium d'un plongeur des grandes profondeurs, par exemple.

1.3.4 Approches classiques

Le plus souvent, les dispositifs faisant usage d'un mot de passe comparent avec la production actuelle une production de référence, associée à un locuteur connu, et décident de leur similarité.

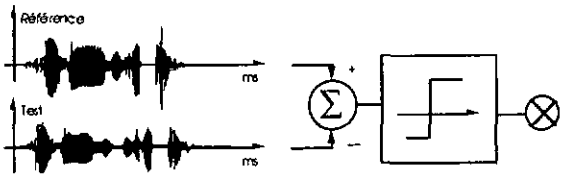

Figure 1.3.a Comparaison de deux mots de passe

La figure 1.3.a montre une esquisse d'un processus de reconnaissance de locuteur par comparaison de mots de passe. Deux locutions sont mises en correspondance et leur écart est comparé à un seuil de décision. L'échelle de temps est à court terme, la durée d'un mot de passe valant volontiers de une à deux secondes en pratique.

Le plus souvent, les dispositifs faisant usage de l'indépendance du texte comparent des statistiques à long terme estimées indépendamment sur la production actuelle et sur la production de référence. On ne compare plus directement des locutions; par contre, on se fie plus volontiers à l'écart entre deux représentations de densité de probabilité, par exemple le vecteur moyen [82SchR, 85WolJ].

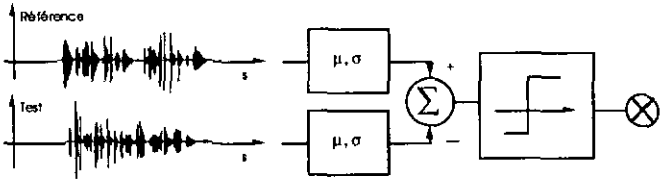

Figure 1.3.b Comparaison de deux locutions

La figure 1.3.b montre une esquisse d'un processus de reconnaissance de locuteurs indépendante du texte. L'écart entre les représentations statistiques de deux locutions est comparée à un seuil de décision. L'échelle de temps est le plus souvent à long terme dans un tel système, la durée de la locution de test

valant en pratique une dizaine de seconde au moins pour assurer une robustesse suffisante de l'estimation statistique. [81FurS2] présente une discussion comparative plus étendue de ces deux types de processus.

1.4 Nature de la tâche

Nous venons de décrire deux façons de comparer des locutions; elles ont en commun de fournir une mesure quantitative de leur dissemblance. Or, il est possible d'exploiter cette mesure de plusieurs manières; les deux approches les plus fréquentes font l'objet de ce paragraphe.

1.4.1 Identification

Considérons la situation où nous cherchons à joindre quelqu'un par téléphone à son domicile. Même sans connaître sa famille, nous savons a priori que la personne qui répondra à notre appel sera vraisemblablement le père, la mère, un des fils ou une des filles; parmi ces gens se trouve le correspondant attendu. Même si la personne qui répond à notre appel ne s'annonce pas de façon formelle, nous saurons néanmoins découvrir son identité grâce à cette information a priori. Nous sommes dans la situation d'une tâche d'identification, où nous cherchons à assigner l'identité la plus probable à la voix examinée, étant donné un ensemble de références connues. Dans notre cas, les références sont l'archétype d'une voix de mère de famille, l'archétype d'une voix de père de famille, etc.

Deux cas se présentent: soit une identité est nécessairement assignée au locuteur, soit on considère la possibilité que le locuteur n'ait pas d'identité connue. Dans le premier cas nous parlerons d'identification 1 à n, ou identification sans rejet, et nous associerons au locuteur l'identité dont la probabilité d'adéquation avec la voix examinée est la plus élevée, quelle que soit la valeur de cette probabilité. Dans le second cas nous parlerons d'identification 1 à $n+1$, ou identification avec rejet, et nous refuserons d'associer une identité connue au locuteur si nous constatons par exemple que la probabilité la plus élevée est encore trop basse, ou alors qu'elle ne se démarque pas suffisamment des autres possibilités.

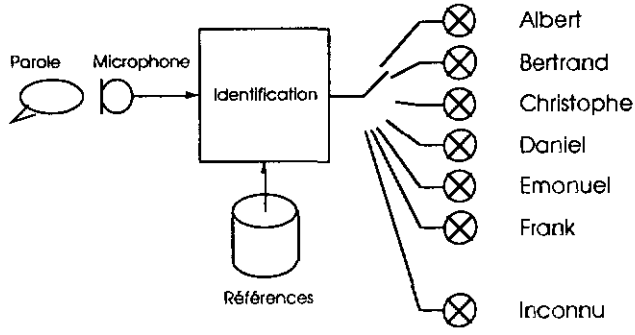

Figure 1.4.a Tâche d'identification avec rejet

La figure 1.4.a montre un processus d'identification qui analyse une locution en disposant d'une base de données où sont stockées les références des locuteurs et qui active une des lampes formant la sortie du processus. A chaque choix s'associe une identité; l'identification 1 à $n+1$ peut activer une sortie étiquetée "Inconnu", tandis que l'identification 1 à n n'active jamais cette sortie.

1.4.2 Vérification

Considérons la situation où nous avons joint quelqu'un par téléphone à son domicile. La personne qui vient de répondre à notre appel s'est annoncée de façon formelle «Allô, ici la Reine d'Angleterre». Laissons de côté le cas où cette identité et la voix qui y correspond nous seraient inconnues, et où nous serions donc incompétent pour prendre une décision; notre travail consiste alors simplement à examiner la correspondance entre la voix entendue et l'identité annoncée. De ces deux termes, c'est l'identité que nous choisirons comme référence; c'est à elle que nous rapporterons la voix entendue plutôt que le contraire. Donc, eu égard à l'identité annoncée, si nous nous attendons à une autre voix que celle que nous avons entendue alors nous pouvons en déduire que nous sommes en relation avec un imposteur. Au contraire, nous n'avons pas affaire à un imposteur quand la voix entendue correspond avec la voix que nous savons associer à l'identité annoncée.

Nous venons de décrire une tâche de vérification, qui peut conduire à deux décisions différentes pour une présentation simultanée de parole et de prétention d'identité: dans le premier cas, la tâche de vérification décide que la voix et l'identité ont été engendrées par la même personne; on dit alors que le couple

(*voix, identité*) est homogène. Dans le second cas, la tâche de vérification décide que voix et identité n'ont pas la même origine; ont dit que le couple (*voix, identité*) est hétérogène.

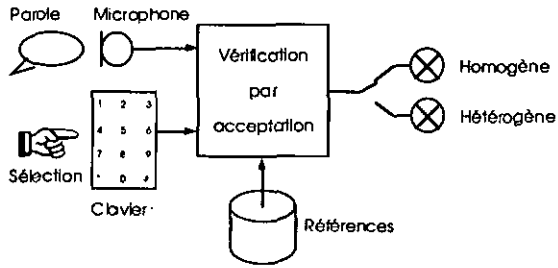

Figure 1.4.b Tâche de vérification

La figure 1.4.b montre un processus de vérification qui analyse une locution en disposant d'une base de données où sont stockées les références et qui active la sortie correspondant à la décision de rejet ou d'acceptation de l'homogénéité entre le signal acoustique et la prétention d'identité, représentée sur la figure par un code tapé sur un clavier.

Du point de vue interne du fonctionnement du processus de vérification, deux cas sont à distinguer. Le premier se contente d'examiner la correspondance entre l'identité prétendue et la voix présentée sur la base de la seule référence associée à l'identité en question. On parle alors de vérification par acceptation. Le second exploite plusieurs références disponibles et s'assure qu'elles permettraient toutes de rejeter le locuteur, à l'exception de l'identité prétendue. On parle alors de vérification par rejet, que l'on ne rencontre que très rarement dans la pratique.

1.4.3 Difficulté et taille de la population

Considérons une tâche d'identification 1 à n où la base de données associée ne contient qu'un seul et unique locuteur, et où l'on sous-entend qu'aucun inconnu n'accède au système. Il s'ensuit qu'aucune erreur n'est possible dans l'identification du locuteur! Enrichissons maintenant cette base de données en y introduisant une à une la référence vocale de chaque être humain habitant la planète. Il est clair que les possibilités de confusion entre deux personnes vont croître à chaque ajout.

Considérons une tâche d'identification 1 à $n+1$ où la base de données associée ne contient qu'un seul et unique locuteur; on sous-entend maintenant que d'innombrables locuteurs inconnus peuvent accéder au système. Il s'ensuit que, contrairement au cas précédent, des erreurs d'identification peuvent déjà se produire! Enrichissons maintenant cette base de données en y introduisant une à une la référence vocale de chaque être humain habitant la planète. Il est clair que les possibilités de confusion entre deux personnes vont croître à chaque ajout [91BasC], alors que le réservoir de locuteurs encore inconnus reste plein, de par l'hypothèse de variabilité infinie.

Considérons une tâche de vérification par acceptation. Que la base de données associée ne contienne qu'un seul et unique locuteur ou qu'elle soit beaucoup plus étoffée ne change rien à la façon de prendre la décision d'homogénéité ou d'hétérogénéité, de par le principe même de la vérification par acceptation qui ne fait usage que d'une seule référence. Il s'ensuit que la probabilité d'erreur, en tâche de vérification par acceptation, est individuellement dépendante du classificateur considéré, et de lui seul. L'estimation de cette probabilité se réalise en dénombrant les erreurs de vérification commises relativement au cardinal de l'ensemble des locuteurs testés. Par conséquent, contrairement aux autres tâches de reconnaissance citées, l'introduction de locuteurs de test supplémentaires n'a pas pour effet l'augmentation de la difficulté de la reconnaissance mais plutôt celui de l'augmentation de la robustesse des estimations individuelles de probabilité d'erreur. Cependant, l'introduction de locuteurs de référence supplémentaires nécessite de prendre en compte un paramètre nouveau: l'efficacité ou l'inefficacité individuelle d'une méthode de reconnaissance en fonction du locuteur de référence considéré. Or, on peut considérer que la probabilité générale d'échec d'une méthode donnée existe indépendamment de sa réalisation pour des locuteurs particuliers; elle peut être estimée en calculant la moyenne des probabilités d'échec individuels. Par conséquent, l'introduction de locuteurs de référence supplémentaires n'a pas non plus pour effet l'augmentation de la difficulté de la reconnaissance mais bien plutôt celui de l'augmentation de la robustesse de l'estimation générale de probabilité d'erreur.

1.5 Notre approche

Ce travail de thèse propose une solution au problème de la reconnaissance de locuteurs. Il prétend répondre au critère de l'indépendance du texte plutôt que de la dépendance du texte, et se concentre sur une tâche de vérification plutôt

que d'identification. Enfin, il se fonde sur l'expérimentation plus qu'il ne prétend être une contribution de théoricien.

1.5.1 Motivations

Plusieurs raisons nous permettent de justifier les choix que nous venons d'énoncer. Premièrement, quant à l'indépendance du texte, mentionnons le confort qu'elle apporte au locuteur qui n'a plus l'obligation non seulement de se souvenir d'un mot de passe, mais encore de la façon précise de le prononcer. Son corollaire, le fait que la coopération du locuteur ne soit plus requise, est par contre une caractéristique ambiguë, ressentie par certains comme un avantage et par d'autres comme une source d'abus potentiels. D'un point de vue moins passionnel, nous constatons que la reconnaissance de locuteurs dépendante du texte est plus mûre que ne l'est l'indépendante qui mérite par conséquent plus d'attention que la première. Enfin, il est bien plus facile d'étendre une méthode bâtie sur l'indépendance du texte et l'en rendre dépendante que le contraire; parallèlement, son domaine d'application recouvre des problèmes qu'une méthode dépendante du texte serait impuissante à résoudre.

Deuxièmement, quant à la tâche de vérification par acceptation, l'insensibilité de son efficacité face au nombre de locuteurs pris en compte est notre justification principale de l'abandon de la tâche de vérification par rejet et de la tâche d'identification, puisqu'elle permet d'extrapoler les performances d'un système grandeur nature à partir d'une base de données réduite. Les deux dernières tâches citées n'autorisant pas cette extrapolation, nous les avons considérées comme moins enrichissantes en termes d'apport de connaissances nouvelles et nous avons préféré renoncer à les utiliser.

Troisièmement, quant à l'expérimentation, nous croyons que la maturité des connaissances théoriques décrivant les caractéristiques précises qui font l'individualité d'un locuteur est insuffisante. En effet, la théorie seule ne permet pas de prédire la probabilité d'erreur d'une méthode de reconnaissance donnée. Dans l'état actuel, seule l'expérience permet d'estimer cette probabilité; c'est donc par ce biais que nous avons choisi de montrer la valeur des méthodes que nous proposons. En particulier, nous verrons plus loin que nous avons examiné un aspect considéré comme fondamental puisqu'il touche à la nature même du vecteur caractéristique extrait du signal de parole; de ce point de vue, nous espérons que cette contribution sera profitable au raffinement d'un éventuel modèle théorique futur.

1.5.2 Critère de qualité

Ce travail de thèse se concentre essentiellement sur une tâche de vérification, dont les seules erreurs possibles sont d'une part prendre une décision d'homogénéité entre un échantillon de voix et une identité alors qu'en réalité l'auteur de la locution est d'identité différente de celle à laquelle il a prétendu, et d'autre part accuser d'imposture un locuteur qui a pourtant présenté sa propre voix et sa propre identité au vérificateur. On parle respectivement de fausse acceptation et de faux rejet.

Le taux de fausse acceptation se déduit de l'observation des résultats obtenus par une tâche de vérification confrontée exclusivement à des entrées telles que l'identité prétendue ne corresponde pas à celle du locuteur considéré; le rapport entre le nombre de décision d'homogénéité des entrées et le nombre total d'essais détermine ce taux. Le taux de faux rejet se déduit de l'observation des résultats obtenus par une tâche de vérification confrontée exclusivement à des entrées telles que l'identité prétendue soit justement celle du locuteur; le rapport entre le nombre de décision d'hétérogénéité des entrées et le nombre total d'essais détermine ce taux.

1.5.3 Contribution

Nous allons consacrer le corps de cette thèse à l'exposé des éléments théoriques et pratiques permettant la mise en œuvre d'un système de reconnaissance de locuteur indépendante du texte. Notre originalité principale tient en l'exploration des potentialités du résidu de l'analyse par prédiction linéaire, un vecteur caractéristique inédit dans ce domaine. Nous avons aussi proposé une méthode nouvelle d'exploitation d'un vecteur caractéristique connu. Cette méthode complète celle de l'erreur moyenne de quantification vectorielle; nous l'avons baptisée méthode de la conformité. En outre, nous avons répété certaines expériences classiques et nous avons construit une nouvelle base de données comportant de nombreux locuteurs.

2 Outils de reconnaissance

L'objet de ce chapitre est de présenter certains des outils dont nous faisons usage dans cette thèse. En premier lieu, nous montrons un modèle de la vérification (la tâche de reconnaissance que nous avons choisie), puis nous introduisons deux façons d'établir les vecteurs représentatifs d'un locuteur; la première utilise une technique de valeur moyenne et la seconde fait usage de la technique de classification des nuées dynamiques que nous utiliserons souvent dans la suite de cette thèse. Nous abordons ensuite une description plus détaillée des mesures classiques et nouvelles de dissemblance. Enfin, nous définissons les taux d'erreur sur lesquels nous fondons l'interprétation des résultats.

2.1 Modèle de vérification

Nous avons exprimé et justifié au paragraphe 1.5 notre volonté d'expérimenter. Ce choix nous conduit à considérer principalement deux phases dans l'investigation des méthodes de reconnaissance que nous proposons; il s'agit premièrement de la phase d'apprentissage et secondement de la phase de test. Tandis que celle-ci permet d'évaluer la qualité des classificateurs obtenus, celle-là dessert le travail de leur construction. Chacune nécessite une certaine quantité de données que l'on extrait respectivement d'un ensemble d'apprentissage E_a et d'un ensemble de test E_t. Le premier sert à construire les références de la base de données représentée à la figure 1.4.b; ces références sont composées de deux parties distinctes: d'une part un vecteur représentatif du locuteur et d'autre part un seuil de décision associé. Le second sert de réserve où puiser les éléments permettant d'estimer la probabilité d'erreur liée à la méthode examinée.

2.1.1 Séparation des phases de traitement

Notons tout d'abord que réaliser les opérations d'apprentissage et de reconnaissance de façon séparée dans le temps n'est qu'une commodité, mais n'est nullement indispensable. Par exemple, nous verrons plus loin qu'un auditeur à qui l'on présente deux échantillons de parole est capable de distinguer le cas où deux locuteurs différents ont parlé du cas où il n'y en avait qu'un seul, même

s'il ne connaît aucun d'eux. Dans cet exemple, l'écoute des locuteurs, la construction des vecteurs représentatifs, l'établissement des seuils de décision ainsi que la reconnaissance elle-même sont des processus quasi simultanés.

En traitement automatique cependant, il est préférable de considérer une phase de reconnaissance bien séparée de l'apprentissage; quant à ce dernier, il est de plus souhaitable de distinguer de la phase de construction du vecteur représentatif celle de l'établissement des seuils. La raison en est que la reconnaissance nécessite la comparaison d'une locution de test avec un vecteur représentatif. Or, il est rarissime que celui-ci soit constitué directement par le signal temporel de parole, car aucune technique fonctionnant sur cette base n'a jamais été reconnue comme efficace; il est bien plus fréquent de faire usage d'une transformation de la locution de test et de celle d'apprentissage et de mener la comparaison dans cet espace transformé. Bien entendu, le vecteur représentatif obtenu est conservé et n'a pas à être recalculé à chaque nouvelle comparaison. Il arrive en outre bien souvent qu'un tel vecteur nécessite une capacité de stockage très inférieure à celle du signal dont il provient, car presque toujours il décrit sous une forme condensée et plus ou moins précise la distribution du signal de parole pour un locuteur donné. Il s'ensuit que le gain associé à la conservation de la locution de référence sous une forme transformée cumule une réduction de place et une réduction de temps de calcul, et justifie ainsi ce découpage en trois phases de la tâche de vérification.

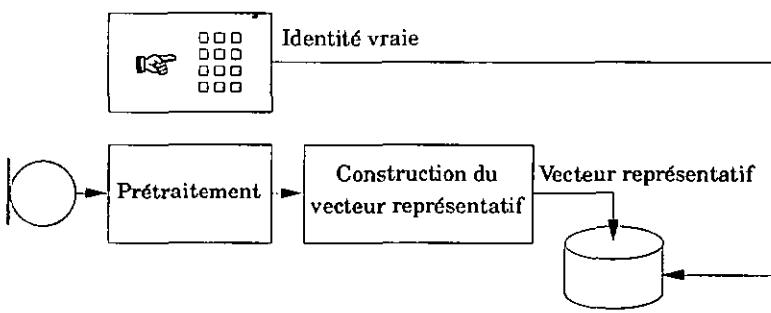

Figure 2.1.a Construction du vecteur représentatif

2.1.2 Construction du vecteur représentatif

La figure 2.1.a présente la phase de construction du vecteur représentatif. On y retrouve en entrée les deux éléments nécessaires: une émission vocale ainsi que l'identité du locuteur qui l'a produite. Cette dernière sert ici simplement à sélectionner l'élément à établir dans la base de données, tandis que la voix du locuteur subit tout d'abord un prétraitement de sorte à en extraire les vecteurs caractéristiques (cette étape sera décrite plus en détail au chapitre suivant). En phase d'apprentissage, ces vecteurs caractéristiques servent alors à construire l'élément vecteur représentatif de la référence du locuteur.

2.1.3 Établissement des seuils de décision

La tâche de vérification nécessite la connaissance de seuils de décision; or, cette connaissance s'acquiert dans la phase d'apprentissage d'un dispositif automatique de reconnaissance de locuteurs. La figure 2.1.b présente une vue fortement simplifiée de la phase d'établissement de ces seuils où l'on distingue trois entrées. L'une est l'identité prétendue; elle sert à sélectionner au sein de la référence le vecteur représentatif, que l'on suppose déjà connu, et le seuil concerné que l'on cherche à établir. L'autre est un échantillon de parole qui est prétraité puis comparé avec le vecteur représentatif sélectionné. Enfin, la troisième et dernière entrée est l'identité vraie du locuteur.

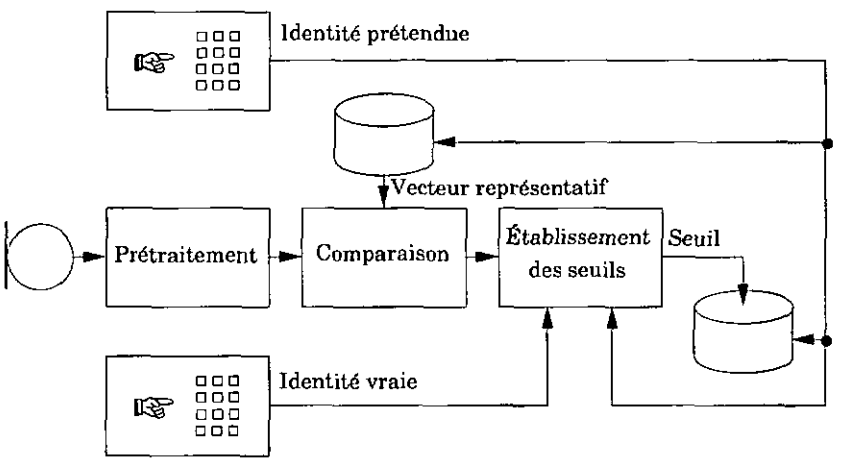

Figure 2.1.b Établissement des seuils de décision

Au cœur de cette figure, le module d'établissement des seuils de décision d'une tâche de vérification reçoit du module de comparaison une mesure de la dissemblance entre le vecteur représentatif et les vecteurs caractéristiques de la locution en cours, ainsi qu'une paire d'identités dont il peut tester l'égalité. La simplification de principe de la figure 2.1.b découle du fait qu'en réalité, pour une identité prétendue donnée, ce module n'est capable d'établir un seuil que sur la base de locutions d'au moins deux auteurs différents, dont celui possédant l'identité prétendue. Ce nombre est un minimum absolu; en pratique, il vaut mieux qu'il y soit très supérieur. Nous décrirons en détail au chapitre traitant des bases de données le mécanisme, occulté sur la figure 2.1.b, de report de décision et de sélection de multiples locutions.

2.1.4 Tâche de vérification

En phase de reconnaissance d'une tâche de vérification, la décision d'homogénéité ou d'hétérogénéité du couple (*voix*, *identité*) se base d'une part sur le seuil sélectionné par l'identité prétendue et d'autre part sur la mesure de dissemblance issue de la comparaison entre la locution transformée par le prétraitement et le vecteur représentatif sélectionné par l'identité prétendue.

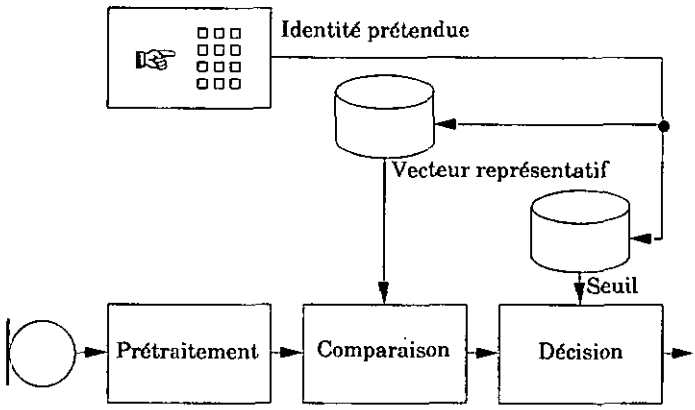

Figure 2.1.c Vérificateur

2.2 Vecteurs représentatifs

Comme nous l'avons vu au paragraphe 2.1.1, la justification de l'existence d'un vecteur représentatif est sa capacité de décrire de manière condensée la distribution des vecteurs caractéristiques pour le locuteur de référence. Dans ce paragraphe, nous examinons deux procédés particuliers pour la construction de ces vecteurs représentatifs. L'un est une représentation très rudimentaire de leur distribution; il s'agit simplement de la valeur moyenne. L'autre y est plus fidèle; il s'agit d'une liste de représentants bien choisis, nommée dictionnaire. Nous montrons ici comment établir cette liste en se basant sur un ensemble discret d'échantillons de la densité de probabilité des vecteurs caractéristiques; l'algorithme décrit est celui des nuées dynamiques.

2.2.1 Valeur moyenne

Soient $\mathbf{x}$ les coordonnées d'un vecteur caractéristique dans un espace U à N dimensions .

2.2.1
$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{pmatrix} \qquad \mathbf{x} \in U$$

Soit $p(\mathbf{x})$ la densité de probabilité multidimensionnelle des vecteurs caractéristiques dans l'espace à classifier U

2.2.2
$$\int_{\mathbf{x} \in U} p(\mathbf{x}) \mathrm{d}\mathbf{x} = 1 \qquad \wedge \qquad p(\mathbf{x}) \geq 0 \qquad \forall \mathbf{x} \in U$$

Soit X un ensemble discret de P vecteurs caractéristiques

2.2.3
$$X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_P\}_P \qquad \mathbf{x}_p \in U$$

La valeur moyenne que l'on cherche à estimer est

2.2.4
$$\bar{\mathbf{x}} = \int_{\mathbf{x} \in U} \mathbf{x} \cdot p(\mathbf{x}) \mathrm{d}\mathbf{x}$$

L'estimation de la valeur moyenne est donnée par

2.2.5 $$\langle \mathbf{x} \rangle = \frac{1}{P} \cdot \sum_{\forall \mathbf{x} \in X} \mathbf{x}$$

Si l'on admet que la distribution des éléments au sein de l'ensemble X reflète fidèlement la distribution $p(\mathbf{x})$, alors

2.2.6 $$\langle \mathbf{x} \rangle \cong \bar{\mathbf{x}}$$

2.2.2 Dictionnaire

La densité de probabilité d'un vecteur caractéristique n'est pas accessible en pratique; seule sa forme discrète l'est. Par exemple, on peut la représenter sous la forme d'un ensemble fini de cellules disjointes, dont l'union couvre l'espace entier, représentées chacune explicitement par un noyau et implicitement par le choix d'une métrique dans l'espace concerné [85Mak]]. Au sens de cette métrique, le lieu des points où la distance est identique par rapport à l'un et l'autre des deux plus proches noyaux forme la membrane des cellules, que l'on nomme cellules de Voronoi ou parfois régions de Dirichlet. Tous les points de l'espace couvert par une cellule donnée appartiennent à la même classe; en d'autres termes, chaque noyau représente la classe des points qui lui sont le plus proches. L'ensemble des noyaux est nommé dictionnaire.

Le but de ce paragraphe est de montrer comment procède une des techniques permettant de passer d'un ensemble de cardinal élevé d'échantillons représentatifs de la densité de probabilité à l'ensemble de cardinal fixe et généralement moindre des représentants hautement pertinents que sont ses noyaux.

A) Nuées dynamiques

L'algorithme des nuées dynamiques, ou l'une de ses variantes, est celui que l'on rencontre le plus souvent comme solution au problème de classification par partitions, qui consiste à déterminer les régions de Dirichlet et leur noyau [80LinY]. La propriété majeure qui le distingue d'autres algorithmes de classification est un nombre de classes K à imposer a priori. Il fonctionne par améliorations successives d'une solution initiale fournie sous la forme d'un ensemble de noyaux dont le cardinal fixe le nombre de classes. L'algorithme nécessite encore une population d'échantillons représentative de la distribution des éléments dans l'espace considéré. Cette population est classée en accord avec les noyaux initiaux; chaque ensemble d'échantillons de même classe sert alors de moule pour générer un nouveau noyau, plus représentatif de la classe en ques-

tion que le noyau initial. On remplace alors les noyaux initiaux par les noyaux nouvellement obtenus et l'algorithme est prêt à être réitéré. Quelques conditions d'arrêt possibles sont par exemple l'épuisement d'un nombre prescrit d'itérations, l'obtention d'une valeur suffisante d'une mesure de la qualité des classes obtenues, ou encore le fait que les noyaux engendrent les classes mêmes dont ils sont issus, ce que nous nommerons convergence.

B) Développement

Soit Y une partition de U donnée par un ensemble de K noyaux $\mathbf{y}_k$

2.2.7 $$Y = \left\{ \mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_K \mid \mathbf{y}_k \in U \quad \forall k \in [1,K] \right\}$$

Soit d la dissemblance entre un échantillon $\mathbf{x}$ et un noyau $\mathbf{y}$. Dans ce cas, le noyau $\mathbf{y}$ le plus proche voisin de l'échantillon $\mathbf{x}$ est donné par la fonction d'affectation q

2.2.8 $$\mathbf{y} = q(\mathbf{x}) = \underset{\forall \mathbf{y}_k \in Y}{\text{ArgMin}}\, d(\mathbf{x}, \mathbf{y}_k) \qquad \forall \mathbf{x} \in U$$

La cellule couverte par la classe de noyau $\mathbf{y}_k$ est

2.2.9 $$C_k = \left\{ \mathbf{x} \mid (\mathbf{x} \in U) \wedge (q(\mathbf{x}) = \mathbf{y}_k) \right\} \qquad \forall k \in [1,K]$$

L'ensemble X_k des échantillons d'une cellule C_k est

2.2.10 $$X_k = \left\{ \mathbf{x} \mid \mathbf{x} \in (X \cap C_k) \right\} \qquad \forall k \in [1,K]$$

Par construction, la propriété de partition est satisfaite

2.2.11 $$U = \bigcup_{k=1}^{K} C_k \qquad \wedge \qquad X = \bigcup_{k=1}^{K} X_k$$

Le problème qu'essaie de résoudre l'algorithme des nuées dynamiques est de déterminer une partition Y telle que la distorsion quadratique ε entre les échantillons et les noyaux soit minimale

2.2.12 $$\varepsilon = \sum_{\forall \mathbf{x} \in X} d^2(\mathbf{x}, q(\mathbf{x}))$$

Soit g la fonction de représentation qui permet de déterminer le noyau optimal $\mathbf{z}_k$ minimisant la distorsion entre lui-même et les échantillons X_k d'une cellule C_k

2.2.13 $$\mathbf{z}_k = g(X_k) = \underset{\forall \mathbf{y} \in U}{\text{ArgMin}} \sum_{\forall \mathbf{x} \in X_k} d^2(\mathbf{x}, \mathbf{y}) \qquad \forall k \in [1, \mathrm{K}]$$

Démontrons que si d est une distance euclidienne, alors le noyau optimal $\mathbf{z}_k$ est le centre de gravité de l'ensemble X_k. Commençons par annuler la dérivée de l'expression de la distorsion par rapport à chaque composante du noyau

2.2.14 $$\frac{\partial}{\partial y_i} \sum_{\forall \mathbf{x} \in X_k} d^2(\mathbf{x}, \mathbf{y}) = \sum_{\forall \mathbf{x} \in X_k} \sum_{j=1}^{N} \frac{\partial}{\partial y_i} (y_j - x_j)^2 = 2 \cdot \sum_{\forall \mathbf{x} \in X_k} (y_i - x_i) = 0 \qquad \forall i \in [1, \mathrm{N}]$$

Il s'ensuit

2.2.15 $$\mathbf{z}_k = \frac{1}{\text{Card}(X_k)} \cdot \sum_{\forall \mathbf{x} \in X_k} \mathbf{x} \qquad \forall k \in [1, \mathrm{K}]$$

Voici enfin l'algorithme lui-même où Y_0 est une partition initiale fournie par une méthode quelconque. La fonction d'affectation q est implicitement utilisée par la fonction de représentation g, car X_k y participe, et sa détermination nécessite $q(\mathbf{x})$ selon les définitions 2.2.10 et 2.2.9; cette fonction d'affectation se calculera en utilisant la partition courante Y

2.2.16 $$\begin{cases} \begin{cases} m \leftarrow 1 \\ Y \leftarrow Y_0 \end{cases} \\ \mu(m, Y, X_1, X_2, \ldots, X_\mathrm{K}) \end{cases} \begin{cases} \mathbf{z}_k \leftarrow g(X_k) \qquad \forall k \in [1, \mathrm{K}] \\ Y \leftarrow \{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_\mathrm{K}\} \\ m \leftarrow m + 1 \end{cases}$$

La condition d'arrêt μ a été volontairement laissée imprécise de sorte à permettre diverses approches. La solution la plus simple consiste à comparer le compteur m à un seuil arbitraire; l'expérience montre qu'il est judicieux d'imposer un facteur de proportionnalité entre la dimension N de l'espace de travail et la valeur du seuil. Une autre condition d'arrêt pourra exploiter la distorsion quadratique ε de la relation 2.2.12. Enfin, on pourra encore imposer d'attendre la convergence, qui apparaît nécessairement après un nombre fini d'itérations.

C) Utilité

Le caractère continu de la distribution $p(x)$ la rend inobservable dans la pratique, au contraire de la population X qui en est une approximation discrète; ce n'est pourtant que si l'on admet la réalité de l'existence de la distribution continue que l'on peut espérer rencontrer une autre population $X' \neq X$ qui puisse être représentée avec une fidélité suffisante par le même noyau que celui construit à l'aide de X, le lien entre ces deux populations étant réalisé justement par cette inaccessible densité de probabilité donnée à l'expression 2.2.2. En ce sens, l'utilité de l'algorithme des nuées dynamiques est de réduire le cardinal d'une population d'échantillons et simultanément d'en conserver la représentativité.

D) Convergence

La démonstration du caractère fini du nombre d'itérations nécessaires procède en deux étapes que nous ne développerons pas de manière formelle. La première montre que, à noyaux fixes, le classement selon la technique du plus proche voisin minimise la distorsion; la seconde montre que, à classes fixes, les noyaux générés par la fonction de représentation minimisent aussi la distorsion. Donc, si l'on utilise alternativement fonction d'affectation et fonction de représentation, cette distorsion ne peut que décroître; or, puisqu'elle est positive, elle possède une borne inférieure. Il s'ensuit qu'elle converge. Toute suite convergente sur un ensemble fini atteignant sa limite, nous concluons que la situation où les noyaux engendrent les classes mêmes dont ils sont issus est atteinte après un nombre fini d'opérations [82DidE].

Il reste à examiner la qualité de la solution obtenue. En effet, nous savons qu'elle satisfait la minimisation de la distorsion ε, mais rien ne permet d'affirmer que ce minimum est global. Au contraire, l'expérience montre qu'en réalité le minimum n'est que local, car il arrive souvent que l'algorithme fournisse des partitions différentes pour des conditions initiales Y_0 et Y_1 différentes, même si X reste inchangé. Dans un pareil cas, on retiendra bien sûr la partition pour laquelle la distorsion observée est la plus petite.

E) Exemple

Les figures qui suivent proposent une représentation visuelle du procédé de l'algorithme tel qu'il s'applique à un espace à deux dimensions.

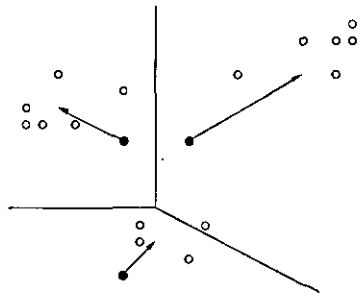

Figure 2.2.a Nuées dynamiques, situation initiale

La figure 2.2.a montre la situation initiale, avec en cercles pleins les trois noyaux considérés et en cercles vides les échantillons de l'ensemble d'apprentissage. Les droites de la figure montrent les frontières des cellules, tandis que les flèches visualisent l'effet de la fonction de représentation sur chaque noyaux. Constatons que, pour l'instant, un des échantillons de la partie inférieure de la figure paraît être mal classé.

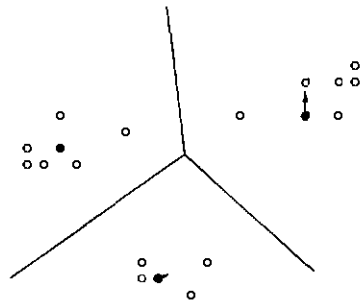

Figure 2.2.b Nuées dynamiques, situation intermédiaire

Après avoir ajusté une première fois la position des noyaux, la forme des frontières donnée par la fonction d'affectation s'est modifiée déjà suffisamment pour qu'un échantillon change de classe. La figure 2.2.b montre en outre que le noyau de droite n'a plus qu'un petit déplacement à réaliser, que minuscule est la distance entre la position actuelle du noyau inférieur et celle du centre de gravité des échantillons dont il représente la classe, et qu'enfin le noyau de

gauche ne bouge déjà plus. La figure 2.2.c montre le résultat obtenu à la convergence, en mettant en évidence les classes générées.

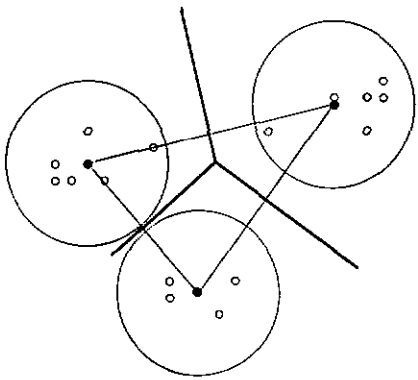

Figure 2.2.c Nuées dynamiques, situation finale

Ce petit exemple permet de se rendre compte d'un certain nombre de problèmes, dont le moindre n'est pas le choix du nombre de classes; en effet, l'algorithme des nuées dynamiques ne cherche pas à adapter ce nombre en fonction des données qu'il connaît. Au contraire, c'est au commanditaire d'une classification de le choisir a priori, d'une façon qu'il jugera compatible avec la densité de probabilité. Notons que l'introduction d'un quatrième noyau dans la situation initiale de la figure 2.2.a serait malvenue.

2.3 Comparaisons

L'objet de ce paragraphe est de présenter différentes techniques de comparaison des vecteurs caractéristiques de test et du vecteur représentatif. Les trois premières sont classiques et mesurent l'écart entre deux vecteurs représentatifs de même dimension que les vecteurs caractéristiques. La quatrième mesure l'écart entre un ensemble de vecteurs caractéristiques de test d'une part et un vecteur représentatif de référence, lui-même constitué d'une liste de vecteurs; dans ce cas, le principe de comparaison fait appel à la quantification vectorielle, de même que pour la cinquième technique de comparaison, nommée conformité. Enfin, la description de la proéminence clôt ce paragraphe.

2.3.1 Distance euclidienne

Soient $\mathbf{x}$ les coordonnées d'un vecteur caractéristique dans un espace U à N dimensions

2.3.1
$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{pmatrix} \qquad \mathbf{x} \in U$$

Si $\mathbf{y}$ est un autre vecteur caractéristique plongé dans U, alors on appelle distance au sens de la métrique L^n la somme

2.3.2
$$d_n(\mathbf{x},\mathbf{y}) = \left(\sum_{k=1}^{N} |y_k - x_k|^n \right)^{\frac{1}{n}} \qquad \forall \mathbf{x},\mathbf{y} \in U \quad \wedge \quad n > 0$$

En particulier, la condition $n = 2$ correspond à la distance euclidienne. On peut montrer qu'elle satisfait les propriétés d'une métrique

2.3.3
$$\forall \mathbf{x},\mathbf{y},\mathbf{z} \in U \qquad \begin{cases} d_2(\mathbf{x},\mathbf{y}) = d_2(\mathbf{y},\mathbf{x}) \\ d_2(\mathbf{x},\mathbf{y}) \geq 0 \\ d_2(\mathbf{x},\mathbf{y}) = 0 \quad \Leftrightarrow \quad \mathbf{x} = \mathbf{y} \\ d_2(\mathbf{x},\mathbf{y}) \leq d_2(\mathbf{x},\mathbf{z}) + d_2(\mathbf{z},\mathbf{y}) \end{cases}$$

La première condition de l'expression 2.3.3 exprime la contrainte de symétrie, les deux suivantes celle de positivité et la quatrième celle de l'inégalité triangulaire.

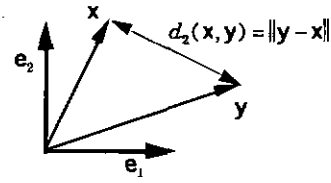

Figure 2.3.a Distance euclidienne

Pour se convaincre par un exemple de la force de la condition imposée par l'inégalité triangulaire, il suffit simplement d'appliquer une quantification aux distances euclidiennes; on constate alors que la distance euclidienne quantifiée

n'est déjà plus une métrique car elle ne satisfait plus l'inégalité triangulaire dans certains cas. Nous avions cependant décrit au paragraphe 1.1.3 notre choix d'utiliser un calculateur numérique tout en acceptant implicitement les effets dus aux approximations introduites. La perte des propriétés d'une métrique pour la distance euclidienne quantifiée en fait partie; néanmoins, nous ignorerons ce fait dans la suite de cette thèse.

■ 2.3.2 Distance euclidienne pondérée

Soit **w** un vecteur plongé dans l'espace U. Par définition, la distance euclidienne pondérée est

2.3.4 $$d_{2,\text{pond}}(\mathbf{x},\mathbf{y}) = \sqrt{\sum_{k=1}^{N} w_k^2 \cdot (y_k - x_k)^2} \qquad \forall \mathbf{x},\mathbf{y} \in U \ \wedge \ \mathbf{w} \in U$$

L'interprétation géométrique de cette mesure de distance est liée à la distance euclidienne de la façon suivante: cette dernière est une mesure géométrique de l'éloignement des extrémités des deux vecteurs participant à la détermination de leur distance, tandis que la métrique euclidienne pondérée commence par étirer les axes du repère associé à l'espace dans lequel sont plongés ces vecteurs. Ce n'est qu'après cet étirement que la mesure géométrique peut procéder; le repère n'est plus nécessairement orthonormé mais reste orthogonal.

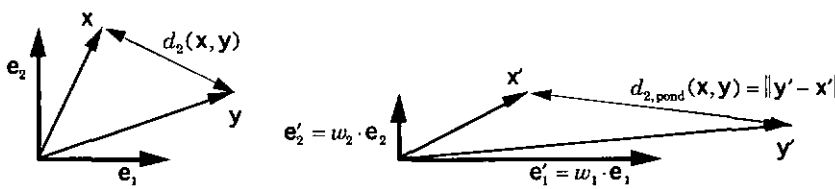

Figure 2.3.b Distance euclidienne pondérée

Le choix des pondérations w_k^2 est souvent lié à l'aspect que l'on désire mettre en valeur. Pour une tâche de reconnaissance, il s'agit du pouvoir discriminant individuellement observable pour chacune des N dimensions. Malheureusement, sa mesure fait intervenir une prise de décision discrète et de ce fait ne permet pas de le maximiser par un mécanisme d'annulation de dérivée, puisque son caractère analytique est perdu. Il s'ensuit que l'on préfère, dans la pratique, se livrer à une détermination de ces pondérations selon un critère

moins ambitieux, mais plus facilement applicable que nous allons justifier puis décrire.

Observons la variance de la composante x_k du vecteur caractéristique $\mathbf{x}$. Si cette variance est grande relativement à celle des autres composantes, alors nous pouvons nous attendre, dans un calcul de distance, à ce qu'une contribution d'amplitude donnée, selon cette composante, soit moins pertinente qu'une contribution d'amplitude identique observée pour une composante de variance moindre. Il suit de ces considérations qu'il n'est pas absurde d'imposer aux coefficients de pondération une valeur inverse de celle de la variance

2.3.5 $$w_k^2 = \frac{1}{\sigma_k^2} \qquad \forall k \in [1,N]$$

2.3.3 Distance de Mahalanobis

La distance de Mahalanobis généralise le principe de la distance euclidienne pondérée parce qu'elle prend encore en compte la combinaison des composantes entre elles. Si $\mathbf{W}$ est une matrice de pondération, alors la distance de Mahalanobis entre deux vecteurs $\mathbf{x}$ et $\mathbf{y}$ de l'espace U vaut

2.3.6 $$d_{2,\text{Maha}}(\mathbf{x},\mathbf{y}) = \sqrt{(\mathbf{y}-\mathbf{x})^T \cdot \mathbf{W} \cdot (\mathbf{y}-\mathbf{x})} \qquad \forall \mathbf{x},\mathbf{y} \in U$$

Pour des raisons semblables à celles évoquées au paragraphe précédent, on choisit volontiers la matrice de pondération comme égale à l'inverse de la matrice de covariance $\mathbf{C}$

2.3.7 $$\mathbf{W} = \mathbf{C}^{-1}$$

Dans cette expression, chaque composante C_{ij} de la matrice $\mathbf{C}$ correspond à la covariance entre les composantes i et j. Géométriquement, la métrique de Mahalanobis non seulement étire les axes mais encore leur fait subir une rotation avant de procéder à la mesure géométrique; le repère n'est plus nécessairement ni orthonormé ni même orthogonal.

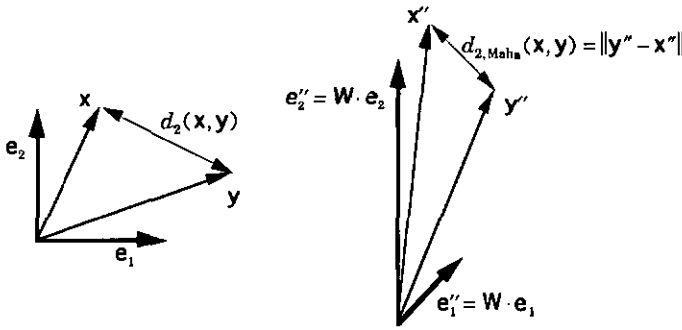

Figure 2.3.c Distance de Mahalanobis

2.3.4 Erreur moyenne de quantification vectorielle

La quatrième mesure de distance que nous présentons ici est nommée erreur moyenne de quantification vectorielle; une de ses particularités est de considérer la distance non pas entre un vecteur de test $\mathbf{x}^{(i)}$ et un vecteur de référence $\mathbf{y}^{(k)}$, comme précédemment, mais entre un ensemble à taille quelconque de vecteurs de test $X^{(i)}$ et un ensemble donné de noyaux représentatifs $Y^{(k)}$. Cette asymétrie fait irrémédiablement perdre le statut de métrique à cette mesure de distance.

A) Quantification vectorielle

La quantification vectorielle des échantillons $\mathbf{x}$ de X par les noyaux $\mathbf{y}$ de Y consiste à appliquer la fonction d'affectation q de l'expression 2.2.8 à tous les éléments de $X \subseteq U$

2.3.8 $$q: U \to Y \mid \mathbf{x} \mapsto \mathbf{y} \qquad \forall \mathbf{x} \in X$$

B) Erreur moyenne

Il résulte du processus de substitution de $\mathbf{x}$ par $q(\mathbf{x})$ une certaine erreur de quantification. Globalement, la valeur de dissemblance issue de la comparaison de l'ensemble des vecteurs caractéristiques de test X avec le vecteur représentatif Y, formé lui-même d'un ensemble de noyaux, est donnée par l'erreur moyenne de quantification

2.3.9 $$d_{VQ}(X^{(i)}, Y^{(k)}) = \frac{1}{\text{Card}(X^{(i)})} \cdot \left(\sum_{\forall \mathbf{x} \in X^{(i)}} d(\mathbf{x}, q^{(k)}(\mathbf{x})) \right) \qquad Y^{(k)}, \forall X^{(i)} \subset U$$

Remarquons que le choix de la mesure de distance d, qui permet de déterminer le noyau le plus proche au moyen de l'expression 2.2.8, est indépendant du choix de la mesure d de distance selon 2.3.9, responsable de l'intégration des contributions individuelles d'erreur dans une mesure globale. Cette indépendance découle du fait qu'il s'agit en premier lieu de réaliser le processus de quantification vectorielle, puis seulement de construire l'erreur moyenne de quantification vectorielle.

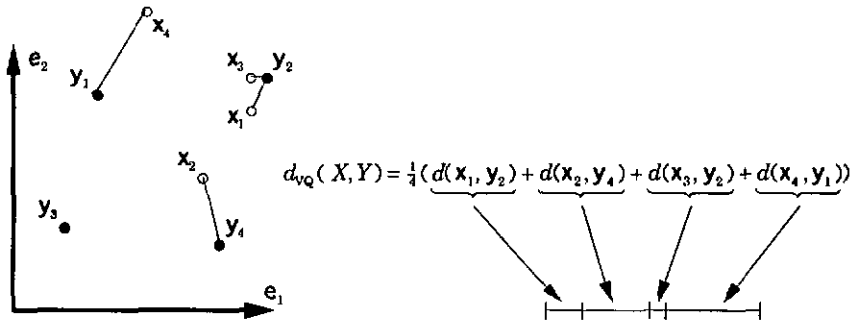

Figure 2.3.d Erreur moyenne de quantification vectorielle

2.3.5 Conformité

La méthode de la conformité a pour but de permettre la vérification de l'hypothèse selon laquelle un ensemble d'échantillons serait une réalisation d'une variable aléatoire de distribution connue. Son principe est de commencer par extraire une série de vecteurs caractéristiques d'une locution, puis de les classer en accord avec une partition convenue de l'espace des paramètres. Il est alors possible de comparer le nombre d'éléments observés dans chaque classe à celui que fournirait la distribution attendue; par exemple, si une mesure globale de l'écart entre réalisation et variable aléatoire est trop grand, alors il y a lieu de déclarer les échantillons comme non confonnes.

Cette méthode complète avantageusement celle de l'erreur moyenne de quantification vectorielle parce que cette dernière ne tient aucun compte de l'information contenue dans la suite des codes identifiant les noyaux des classes, alors que la méthode de la confonnité y porte un intérêt exclusif. En effet, nous avons vu que l'identité des classes n'intervient pas dans le calcul de l'erreur moyenne de quantification vectorielle; en revanche, la méthode de la confor-

mité ignore l'erreur de quantification $d(\mathbf{x},q(\mathbf{x}))$,dont il est fait usage à l'expression 2.3.9, au profit de la représentativité de chaque noyau.

Cette approche se trouve à l'état d'ébauche dans [83LiK]. Cependant, dans ce dernier cas la représentativité sert exclusivement à la sélection d'un sous-ensemble d'un dictionnaire universel lors de la phase de création de références; elle ne participe pas au calcul de distance entre une référence et une locution de test.

Soit Y une partition de l'ensemble des vecteurs caractéristiques intermédiaires, dans l'espace U, donnée par un ensemble de $\mathbf{K}$ noyaux, et nommée dictionnaire universel

2.3.10 $$Y = \left\{ \mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_\mathrm{K} \mid \mathbf{y}_k \in U \quad \forall k \in [1,\mathrm{K}] \right\}$$

Soit X une locution représentée sous la forme d'un ensemble de $\mathbf{P}$ vecteurs caractéristiques intermédiaires

2.3.11 $$X = \left\{ \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_\mathrm{P} \mid \mathbf{x}_p \in U \quad \forall p \in [1,\mathrm{P}] \right\}$$

Soit q la fonction d'affectation d'un vecteur à sa version quantifiée donnée à l'expression 2.2.8, et soit encore $\delta(x)$ la fonction impulsion unité scalaire

2.3.12 $$\delta(x) = \begin{cases} 1 & x = 0 \\ 0 & x \neq 0 \end{cases} \qquad \forall x$$

Soit $\delta(\mathbf{x})$ la fonction impulsion unité vectorielle

2.3.13 $$\delta(\mathbf{x}) = \prod_{n=1}^{N} \delta(x_n) \qquad \forall \mathbf{x} \in U$$

Le vecteur caractéristique intermédiaire $\mathbf{h}$ de la méthode de la conformité représente la fréquence d'apparition de chacun des noyaux associés aux vecteurs caractéristiques de la locution considérée

2.3.14 $$\mathbf{h} = \begin{pmatrix} h_1 \\ h_2 \\ \vdots \\ h_K \end{pmatrix} \qquad h_k = \frac{1}{P} \cdot \sum_{\forall x \in X} \delta(q(\mathbf{x}) - \mathbf{y}_k) \qquad \forall k \in [1, K]$$

Enfin, la mesure de dissemblance issue de la comparaison d'un ensemble $X^{(i)}$ de vecteurs caractéristiques de test avec un vecteur représentatif $\mathbf{h}^{(k)}$ est donnée par

2.3.15 $$d_{Conf}(X^{(i)}, \mathbf{h}^{(k)}) = d(\mathbf{h}^{(i)}, \mathbf{h}^{(k)})$$

De même que pour l'erreur moyenne de quantification vectorielle, le choix de la distance d utilisée à l'expression 2.2.8 pour la détermination du plus proche voisin est indépendant du choix de la mesure de distance d donnée en 2.3.15. Cette indépendance découle du fait qu'il s'agit en premier lieu d'établir le vecteur représentatif, puis seulement de calculer une distance.

La figure 2.3.e illustre le fonctionnement de la mesure de conformité. On y voit un dictionnaire universel constitué de 4 noyaux, destiné à la quantification vectorielle des deux ensembles $X^{(i)}$ et $X^{(k)}$ de test et de référence respectivement qui contiennent ici chacun 11 éléments. Les valeurs des composantes j des vecteurs représentatifs $\mathbf{h}^{(i)}$ et $\mathbf{h}^{(k)}$ sont données graphiquement sur la figure.

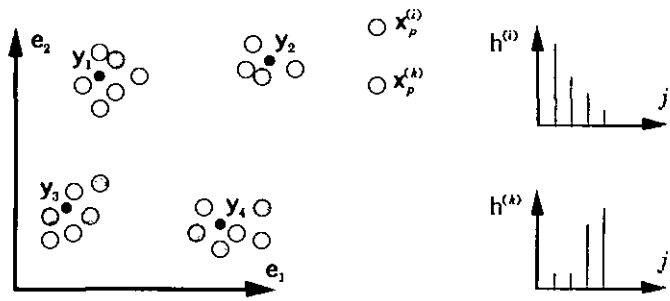

Figure 2.3.e Comparaison par conformité

2.3.6 Proéminence

Nous présentons ici la sixième et dernière technique de comparaison d'un ensemble de vecteurs caractéristiques avec un vecteur représentatif. Dans les

mêmes notations que précédemment, soit $\langle x_j^{(k)} \rangle$ la composante j de la valeur moyenne de l'ensemble des vecteurs caractéristiques déterminant le vecteur représentatif du locuteur de référence (k)

2.3.16 $$\langle x_j^{(k)} \rangle = \frac{1}{P^{(k)}} \cdot \sum_{\forall x \in X^{(k)}} x_j \qquad \forall j \in [1,N]$$

Soit $\varepsilon(x)$ la fonction de Heaviside

2.3.17 $$\varepsilon(x) = \begin{cases} 0 & x \le 0 \\ 1 & x > 0 \end{cases} \qquad \forall x$$

Soit $\mathbf{e}^{(i,k)}(p)$ l'écart positif entre $\mathbf{x}_p^{(i)}$ le vecteur caractéristique p du locuteur de test (i) et la moyenne temporelle des vecteurs caractéristiques associés au locuteur de référence (k)

2.3.18 $$e_j^{(i,k)}(p) = \varepsilon\left(x_j^{(i)}(p) - \langle x_j^{(k)} \rangle\right) \cdot \left(x_j^{(i)}(p) - \langle x_j^{(k)} \rangle\right) \qquad \forall j \in [1,N] \quad \forall p \in [1,P]$$

Soit $w_j^{(i,k)}$ la fraction d'éléments strictement positifs de la composante j de $\mathbf{e}^{(i,k)}(p)$, observés sur toute une locution

2.3.19 $$w_j^{(i,k)} = \frac{1}{P^{(i)}} \cdot \sum_{\forall x \in X^{(i)}} \varepsilon\left(x_j^{(i)} - \langle x_j^{(k)} \rangle\right) \qquad \forall j \in [1,N]$$

Soit $\sigma_j^{(k)}$ l'estimation de la racine carrée de la variance des écarts strictement positifs pour le locuteur de référence (k)

2.3.20 $$\sigma_j^{(k)} = \sqrt{\frac{\sum_{\forall x \in X^{(k)}} \left(e_j^{(k,k)}\right)^2 - \frac{1}{P^{(k)} \cdot w_j^{(k,k)}} \cdot \left(\sum_{\forall x \in X^{(k)}} e_j^{(k,k)}\right)^2}{P^{(k)} \cdot w_j^{(k,k)} - 1}} \qquad \forall j \in [1,N]$$

Normalisons l'écart par la racine de la variance et diminuons l'importance des grandes valeurs par une non-linéarité bien choisie

2.3.21 $$p_j^{(i,k)} = \frac{1}{P^{(i)} \cdot w_j^{(i,k)}} \cdot \sum_{\forall x \in X^{(i)}} \mathrm{Ln}\left(1 + \frac{e_j^{(i,k)}}{\sigma_j^{(k)}}\right) \qquad \forall j \in [1,N]$$

Finalement, la distance d_{Pro} entre un vecteur représentatif de référence $\left(\langle \mathbf{x}^{(k)} \rangle, \sigma^{(k)}\right)^T$ et un ensemble de vecteurs caractéristiques de test $X^{(i)}$ s'obtient par une pondération indépendante des composantes avant l'application d'une métrique euclidienne

2.3.25
$$d_{Pro}\left(\left(\langle \mathbf{x}^{(k)} \rangle, \sigma^{(k)}\right)^T, X^{(i)}\right) = \sqrt{\sum_{j=1}^{N} \left(\frac{p_j^{(i,k)} \cdot w_j^{(i,k)} - p_j^{(k,k)} \cdot w_j^{(k,k)}}{w_j^{(i,k)} + w_j^{(k,k)}} \right)^2}$$

La figure 2.3.f illustre le fonctionnement de la comparaison par proéminence. On y voit les composantes $\langle x_j^{(k)} \rangle$ du vecteur moyen de référence ainsi que les composantes $x_j^{(i)}$ d'un vecteur de test. Seules les parties ombrées de cette figure contribuent au calcul de distance en comparaison par proéminence; elles doivent encore être soumises à une pondération où intervient $\sigma_j^{(k)}$ la racine carrée de la variance des écarts positifs du vecteur de référence.

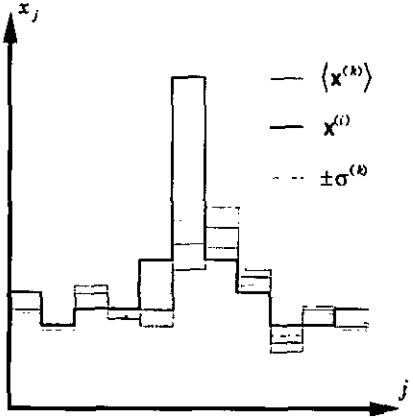

Figure 2.3.f Comparaison par proéminence

2.4 Taux d'erreur

Le but de ce paragraphe est de définir formellement les taux d'erreurs de reconnaissance. Nous commencerons par donner le cadre général, puis nous discuterons cas par cas le type d'erreur rencontrée en rapport avec le type de tâche réalisée.

■ 2.4.1 Généralités

Soit O^* l'ensemble des identités ω_i attribuées à tous les locuteurs

2.4.1 $$O^* = \{\omega_i\} \qquad \forall i > 0$$

Soit Ω^* le sous-ensemble des P locuteurs connus

2.4.2 $$\Omega^* = \{\omega_1, \omega_2, \ldots, \omega_P\}_P$$

Soit encore L l'ensemble de toutes les locutions produites, chaque locuteur (i) ayant pu produire plusieurs locutions repérées par j

2.4.3 $$L = \{\lambda_j^{(i)}\} \qquad \forall i > 0 \quad \wedge \quad \forall j$$

Enfin, soit Λ le sous-ensemble des locutions produites uniquement par les locuteurs connus

2.4.4 $$\Lambda = \{\lambda_j^{(i)}\} \qquad \forall i \in [1, P] \quad \wedge \quad \forall j$$

L'identité d'un locuteur non reconnu sera traduite par le symbole ω_0, dont la signification n'est pas une identité particulière mais représente la classe des locuteurs du sous-ensemble $O^* \setminus \Omega^*$. L'ensemble des identités connues augmenté de l'identité ω_0 attribuée à la classe des inconnus est

2.4.5 $$\Omega = \Omega^* \cup \{\omega_0\}$$

■ 2.4.2 Identification sans rejet

La tâche d'identification sans rejet $\mathbf{I}^*$, ou identification 1 à n, est une application du sous-ensemble Λ des locutions sur le sous-ensemble Ω^* des identités connues

2.4.6 $$\mathbf{I}^* : \Lambda \to \Omega^* \,\big|\, \lambda_j^{(i)} \mapsto \omega$$

De deux choses l'une: ou bien le locuteur identifié est celui qui a émis la locution et l'on parle dans ce cas d'identification correcte ($\omega = \omega_i$), ou bien la tâche d'identification s'est trompée et l'on parle de confusion ($\omega \neq \omega_i$). On peut estimer le taux de confusion en menant N expériences d'identification et en consi-

dérant le rapport entre le nombre C de confusions observées et le nombre N = C+I de tests où I est le nombre d'identifications correctes. Le tableau de la figure 2.4.a montre comment se répartissent ces valeurs.

$\mathbf{I}^*(\lambda_j^{(\omega)}) = \omega$	$\omega = \omega_i$	$\omega \neq \omega_i$
	I	C

Figure 2.4.a Cas d'identification sans rejet

Finalement, l'estimation du taux de confusion ρ_c est

2.4.7 $$\rho_c = \frac{C}{N}$$

2.4.3 Identification avec rejet

La tâche d'identification avec rejet $\mathbf{I}$, ou identification 1 à $n+1$, est une application de l'ensemble L des locutions sur l'ensemble Ω des identités connues augmenté de l'identité ω_0 attribuée à la classe des inconnus

2.4.8 $$\mathbf{I}: L \to \Omega \,\big|\, \lambda_j^{(\omega)} \mapsto \omega$$

Si l'on mène un certain nombre N d'expériences, on peut construire le tableau de la figure 2.4.b, qui accumule le nombre d'occurrences des cas possibles.

$\mathbf{I}(\lambda_j^{(\omega)}) = \omega$	$\omega = \omega_i$	$\omega = \omega_0$	$\omega_i \neq \omega \neq \omega_0$
$\omega_i \in \Omega^*$	I	D	C
$\omega_i \in O^* \setminus \Omega^*$	0	R	M

Figure 2.4.b Cas d'identification avec rejet

Dans cette figure, on découvre I le nombre d'identifications correctes, D le nombre de cas de dédain d'un locuteur connu, C le nombre de cas où l'on confond des locuteurs connus, R le nombre de cas d'assignation correcte de l'identité inconnue à l'auteur de la locution présentée, et enfin M le nombre de cas de méprise où l'on croit avoir affaire à un locuteur connu alors qu'il s'agit d'un inconnu. Bien entendu, nous avons

2.4.9 $$N = I + D + C + R + M$$

Nous nommerons taux de méprise ρ_m la fraction des cas observés où un locuteur connu a été choisi bien qu'en réalité il dût être déclaré inconnu

2.4.10 $$\rho_m = \frac{M}{R+M}$$

Nous nommerons taux de dédain ρ_d la fraction des cas observés où un locuteur, bien que connu, a été délaissé par la tâche d'identification

2.4.11 $$\rho_d = \frac{D}{I+D+C}$$

Tout en restant cohérent avec l'expression 2.4.7, nous redéfinirons le taux de confusion ρ_c par

2.4.12 $$\rho_c = \frac{C}{I+D+C}$$

Il est utile de considérer ici le cas particulier où l'ensemble des locuteurs connus Ω^* est un singleton. Dans ce cas, toute confusion devient impossible; la figure 2.4.c reflète cette situation.

$\mathbf{I}(\lambda_j^{(\omega)}) = \omega$	$\omega_i = \omega = \omega_i$	$\omega = \omega_0$	$\omega_i \neq \omega_1 = \omega \neq \omega_0$
$\omega_i = \omega_1$	I	D	0
$\omega_i \neq \omega_1$	0	R	M

Figure 2.4.c Cas d'identification avec rejet d'un locuteur unique

2.4.4 Vérification

La tâche de vérification $\mathbf{v}$ est une application de l'ensemble L de toutes les locutions et du sous-ensemble Ω^* des seules identités connues sur l'ensemble D des deux éléments de décision $\{h\}$ et $\{\bar{h}\}$ qui représentent respectivement une déclaration d'homogénéité, ou d'hétérogénéité, entre l'identité et la voix présentées

2.4.13 $$D = \{h, \bar{h}\}$$

Notons que nous excluons les identités du sous-ensemble $\{\omega_0\} \cup O^* \setminus \Omega^*$ sans pour autant renoncer aux locutions du sous-ensemble $L \setminus \Lambda$

2.4.14 $$\mathbf{v}: (L \times \Omega^*) \to D \,\Big|\, (\lambda_j^{(i)}, \omega_k) \mapsto \kappa$$

Si l'on mène un certain nombre N d'expériences, on peut construire le tableau de la figure 2.4.d, qui accumule le nombre d'occurrences des cas possibles.

$\mathbf{v}(\lambda_j^{(i)}, \omega_k) = \kappa$	$\kappa = h$	$\kappa = \overline{h}$
$\omega_k = \omega_i$	A	$\overline{\mathrm{R}}$
$\omega_k \neq \omega_i$	$\overline{\mathrm{A}}$	R

Figure 2.4.d Cas de vérification

Dans cette figure, on découvre A le nombre d'acceptations correctes de la prétention d'homogénéité entre voix et identité, $\overline{\mathrm{A}}$ le nombre d'acceptations incorrectes, $\overline{\mathrm{R}}$ le nombre de rejets à tort et R le nombre de rejets corrects. Bien entendu, on a

2.4.15 $$\mathrm{N} = \mathrm{A} + \overline{\mathrm{A}} + \overline{\mathrm{R}} + \mathrm{R}$$

Nous nommerons taux de fausse acceptation ρ_a la fraction des cas observés où une décision d'homogénéité a été prise alors que l'identité et la voix présentées ne correspondaient pas

2.4.16 $$\rho_a = \frac{\overline{\mathrm{A}}}{\overline{\mathrm{A}} + \mathrm{R}}$$

Nous nommerons taux de faux rejet ρ_r la fraction des cas observés où voix et identité ont été considérées à tort comme hétérogènes par la tâche de vérification

2.4.17 $$\rho_r = \frac{\overline{\mathrm{R}}}{\mathrm{A} + \overline{\mathrm{R}}}$$

2.5 Décision en tâche de vérification

Dans une tâche de vérification, la décision $\mathbf{D}$ de rejet $\{\overline{h}\}$ ou d'acceptation $\{h\}$ de l'homogénéité du couple (*voix*, *identité*) est fondée sur la comparaison d'une

des dissemblances d du paragraphe 2.3 à un seuil arbitraire μ. Formellement, nous avons

2.5.1

$$\mathbf{D}(d,\mu) = \begin{cases} \{h\} & d < \mu \\ \{\overline{h}\} & d \geq \mu \end{cases}$$

Il existe plusieurs façons de choisir le seuil de décision. Ces choix se distinguent par le compromis qu'ils imposent entre deux situations extrêmes. L'une impose un seuil μ tellement petit que la seule décision observée est l'hétérogénéité $\{\overline{h}\}$, quelle que soit la dissemblance présentée. Dans ce cas, le taux de fausse acceptation ρ_a du paragraphe 2.4.4 est nul puisque l'hypothèse d'homogénéité n'est jamais acceptée; le taux de faux rejet ρ_r est alors unitaire. L'autre situation impose un seuil μ tellement grand que la seule décision observée est l'homogénéité $\{h\}$, quelle que soit la dissemblance présentée. Dans ce cas, le taux de fausse acceptation est unitaire puisque aucun imposteur n'est démasqué, et le taux de faux rejet est nul puisque le nombre de décisions d'hétérogénéité est nul. Entre ces extrêmes, les compromis que l'on rencontre le plus fréquemment dans la littérature quant au choix du seuil sont

- un seuil de décision propre à engendrer un taux de fausse acceptation égal au taux de faux rejet $(\mu|(\rho_a = \rho_r))$; la valeur correspondante de ces taux est alors appelée taux d'erreur équitable. Cette approche se rencontre fréquemment parmi les chercheurs car elle permet de comparer deux méthodes de reconnaissance sur la base d'un critère scalaire unique;
- un seuil de décision propre à engendrer un taux donné de fausse acceptation $(\mu|\rho_a)$. Le système peut alors être caractérisé par le taux de faux rejet ρ_r résultant. Cette mesure se rencontre fréquemment lorsque l'on cherche à appliquer à une situation réelle un système de vérification du locuteur, le cahier des charges demandant alors très souvent de limiter impérativement les fausses acceptations à un taux donné, au prix éventuel d'une exigence moins sévère en termes de faux rejets;
- un seuil de décision propre à engendrer un taux moyen minimal de faux rejet et de fausse acceptation $(\mu|\partial\frac{1}{2}(\rho_a + \rho_r)/\partial\mu = 0)$. Bien que livrant lui aussi un taux unique, égal à cette valeur minimale, il ne paraît pas être aussi pratique que le taux d'erreur équitable. La raison en est peut-être que si la valeur de ce taux est unique, pourtant la situation où plusieurs seuils y correspondent simultanément n'est pas exclue. De plus, il peut aussi être considéré comme trop optimiste car, même dans le pire des cas, sa valeur ne saurait dépasser 50%; à

cet égard, on peut montrer que sa borne supérieure est donnée par le taux d'erreur équitable.

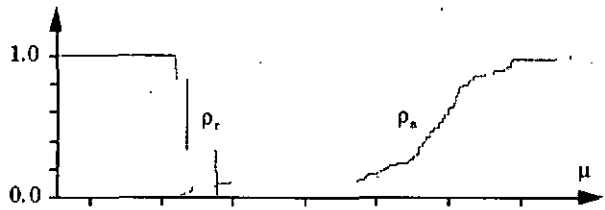

Figure 2.5.a Exemple d'un diagramme des taux d'erreur

Ces grandeurs peuvent toutes être décrites dans le diagramme de la figure 2.5.a, dont l'abscisse est le seuil de décision d'une tâche de vérification et dont l'ordonnée est un taux d'échec. En reportant sur ce diagramme simultanément les taux de fausse acceptation ρ_a et de faux rejet ρ_r en fonction du seuil de décision μ, avec éventuellement leur moyenne, on livre toutes les informations nécessaires à la détermination des mesures d'imperfection citées. Il est cependant nécessaire de se rendre compte que cette approche n'est pas toujours possible; par exemple dans une expérience de vérification exécutée par un auditeur humain, le seuil de décision d'homogénéité ne peut pas être manipulé à loisir.

3 Vecteurs caractéristiques

Le prétraitement a pour but la transformation du signal acoustique de parole. Par définition, on dit qu'une locution dans l'espace transformé est représentée par une suite de vecteurs caractéristiques. L'objet de ce chapitre est de présenter certaines des transformations efficaces pour la reconnaissance de locuteurs; en particulier, nous y développons l'analyse par prédiction linéaire qui permet de transformer une locution en une suite temporelle de vecteurs caractéristiques aptes à reconstruire exactement la locution [75MakJ]. Le modèle de production de parole introduit par cette analyse est généralement considéré comme pertinent.

Le traitement de la parole est un domaine à facettes multiples dont l'ouvrage [87ShaD], parmi bien d'autres [70FanG, 72FlaJ, 75FanG, 76MarJ, 78RabL, 83FerM, 84JayN, 89ProJ, 90WheC], permet d'en survoler les aspects majeurs; nous voulons en donner ici les éléments essentiels à cette thèse. Après l'analyse par prédiction linéaire nous aborderons deux représentations classiques, mais partielles, de certains des paramètres extraits; la première code l'enveloppe du cepstre complexe à court terme du signal de parole [81FurS1] tandis que la seconde découle directement de la précédente puisque le signal y est représenté par la pente temporelle des vecteurs caractéristiques précédents. La représentation suivante est particulière à ce travail de thèse et code le cepstre réel à court terme du résidu du signal de parole. La méthode de l'extraction de la fréquence fondamentale clôt ce chapitre.

3.1 Généralités

Un vecteur caractéristique ne peut susciter l'attention que s'il satisfait certains critères intéressants pour la reconnaissance de locuteurs indépendante du texte [72WolJ]. Au nombre de ceux-ci on trouve

- le naturel: le vecteur caractéristique doit apparaître fréquemment dans la parole et ne pas engendrer de contraintes pour le locuteur. On renoncera par exemple à ne se satisfaire que de voix chantée;

- la robustesse: le vecteur caractéristique doit être insensible aux perturbations du signal de parole occasionnées par exemple par le canal de transmission ou par du bruit de fond;
- la facilité de mesure: si le vecteur caractéristique nécessite un traitement très compliqué pour l'extraire du signal de parole, alors son intérêt pratique peut être mis en doute. De même, si l'acquisition du signal fait appel à des techniques mal maîtrisées ou contraignantes, comme par exemple enregistrer le seul signal de parole véhiculé par la conduction osseuse, à l'exclusion de tout autre, alors son utilité pratique est mineure;
- l'infaillibilité: le vecteur caractéristique ne doit pas pouvoir être modifié par un effort conscient du locuteur. Il doit être tel qu'un imposteur ne puisse réussir une tentative d'imitation;
- la pérennité: le vecteur caractéristique (ou sa distribution) doit rester stable au cours du temps pour un locuteur donné; en particulier, il ne doit pas être affecté par des éléments perturbant le locuteur tels que par exemple sa santé ou son état émotif. S'il s'avère que son évolution le fait varier plus que de l'écart qui le distingue de ses pairs, alors son efficacité doit être mise en doute;
- enfin et surtout l'individualité: le vecteur caractéristique doit permettre de distinguer les locuteurs entre eux. Il est nécessaire que la mesure de cette dissemblance atteigne des valeurs élevées en comparaison de celles obtenues par la comparaison de deux locutions du même locuteur.

3.1.1 Notations

Nous avons parlé jusqu'à maintenant de locutions sans préciser la nature de leur représentation primaire; cependant, les moyens que nous comptons utiliser, comme annoncé au paragraphe 1.1.3, nous imposent une représentation échantillonnée, quantifiée et limitée du signal. Nous avons choisi une cadence d'échantillonnage compatible avec la bande passante d'un canal téléphonique; la précision de la quantification nous a été imposée par le matériel disponible, en particulier par les convertisseurs analogiques-numériques utilisés; la durée du signal a pu varier de cas en cas. Les détails de ces conditions seront donnés au chapitre traitant des bases de données. Le signal de parole temporel continu non quantifié $s(t)$ n'est donc accessible que sous sa représentation discrète, quantifiée et limitée

$$\hat{s}(n) \qquad \forall n \in [N_-, N_+] \tag{3.1.1}$$

Dans cette notation, le tilde diacritique représente le mécanisme de quantification et n est un nombre entier du domaine $[N_-,N_+]$. Par commodité, nous décidons d'ores et déjà de renoncer à préciser la nature quantifiée du signal, à spécifier l'instant origine, et enfin à tenir compte de la durée limitée du signal. La notation devient

3.1.2 $$s(n) \qquad \forall n$$

Ici, n est un nombre entier naturel dont l'origine sera placée de manière adéquate dans le signal.

3.2 Prédiction linéaire

L'analyse de parole par prédiction linéaire est une technique très volontiers utilisée [76Mar]. Son succès est lié d'une part à l'adéquation entre le modèle qu'elle propose et la réalité de l'émission des sons par l'être humain, et d'autre part à un coût opératoire suffisamment faible pour permettre son application en temps réel à des fins de codage ou de reconnaissance de la parole, par exemple. Qui plus est, les paramètres qu'elle fournit possèdent une interprétation physique pertinente.

3.2.1 Modèle de génération d'un signal

Considérons le générateur de signal présenté à la figure 3.2.a.

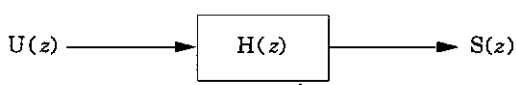

Figure 3.2.a Générateur de signal

Dans cette figure, le signal généré est $S(z)$; $H(z)$ est un filtre dit de synthèse et $U(z)$ est une excitation. Par convention nous avons

3.2.1 $$H(z) = \frac{1+\sum_{l=1}^{q} b_l \cdot z^{-l}}{1-\sum_{k=1}^{p} a_k \cdot z^{-k}} = \frac{B(z)}{A(z)}$$

Le dénominateur $A(z)$ est un polynôme en z^{-1}, appelé filtre inverse, dont le premier tenne est unitaire, et dont le signe apposé aux coefficients est négatif par convention. Les paramètres (p,q) décident du nombre de pôles et de zéros du filtre de synthèse. Nous imposerons $q = 0$, ce qui rend le filtre $H(z)$ du type tout-pôle, ou méromorphe; il suffit alors de jouer sur l'excitation $U(z)$ pour synthétiser à loisir un signal arbitraire $S(z)$ donné, quel que soit $H(z) = 1/A(z)$, les pôles manquants et les zéros de $S(z)$ étant fournis par l'excitation. Nous avons

$$U(z) = H^{-1}(z) \cdot S(z) = A(z) \cdot S(z) \qquad 3.2.2$$

On voit que les coefficients a_k sont arbitraires; nous pouvons donc choisir librement un critère sur l'excitation $U(z)$ tel qu'il impose, s'il est satisfait, un filtre de synthèse univoque pour un signal à synthétiser donné.

3.2.2 Critère d'analyse par prédiction linéaire

Dans la séparation de $S(z)$ en excitation $U(z)$ et filtre de synthèse $H(z)$, l'idée est de modéliser deux comportements séparés du signal: la structure générale (l'enveloppe) de $S(z)$ est décrite par $H(z)$ avec un nombre restreint de coefficients pertinents, tandis que les détails sont cachés dans $U(z)$. Cette vision des choses conduit à tenter l'exploitation du critère qui consiste à minimiser l'énergie du signal d'excitation. Nous verrons que ce critère 1) spécifie $H(z)$ de façon univoque, 2) est de complexité opératoire acceptable et 3) est bien adapté à l'analyse d'un signal de parole, le terme analyse étant mis ici pour décrire l'opération qui consiste à caractériser les paramètres du synthétiseur de la figure 3.2.a qui satisferait le critère choisi et qui reconstruirait de façon exacte un signal $S(z)$ donné.

En passant du domaine échantillonné au domaine temporel nous pouvons écrire, à partir de l'équation 3.2.2

$$u(n) = s(n) - \sum_{k=1}^{p} a_k \cdot s(n-k) \qquad \forall n \qquad 3.2.3$$

Cette expression met en évidence une des propriétés du critère d'analyse retenu: s'il devient possible de tellement minimiser l'énergie de l'excitation que l'on puisse même l'annuler complètement, alors dans ce cas (mais dans ce cas seulement) il devient possible de prédire exactement tout échantillon $s(n)$ du signal à partir d'une combinaison linéaire des p échantillons précédents. Dans

une pareille situation, on peut en outre constater que le filtre de synthèse $H(z)$ est nécessairement instable, puisqu'il génère un signal de sortie non forcément nul pour une entrée rigoureusement nulle.

■ 3.2.3 Minimisation de l'énergie de l'excitation

Développons l'expression de E l'énergie de l'excitation ou, de façon équivalente, de l'erreur quadratique de prédiction progressive

3.2.4
$$\begin{aligned} \mathrm{E} &= \sum_{n=-\infty}^{\infty} u^2(n) = \sum_{n=-\infty}^{\infty} \left(s(n) - \sum_{k=1}^{p} a_k \cdot s(n-k) \right)^2 \\ &= \sum_{n=-\infty}^{\infty} \left(s^2(n) + \left(\sum_{k=1}^{p} a_k \cdot s(n-k) \right)^2 - 2 \cdot s(n) \cdot \sum_{k=1}^{p} a_k \cdot s(n-k) \right) \end{aligned}$$

La minimisation de cette énergie se fait selon les p coefficients a_i du filtre de synthèse

3.2.5
$$\begin{aligned} &\frac{\partial \mathrm{E}}{\partial a_i} = 0 \qquad \forall i \in [1,p] \qquad \Rightarrow \\ &\sum_{n=-\infty}^{\infty} 2 \left(\sum_{k=1}^{p} a_k \cdot s(n-k) \right) \cdot s(n-i) = \sum_{n=-\infty}^{\infty} 2 \cdot s(n) \cdot s(n-i) \qquad \forall i \in [1,p] \qquad \Leftrightarrow \\ &\sum_{k=1}^{p} a_k \cdot \sum_{n=-\infty}^{\infty} s(n-i) \cdot s(n-k) = \sum_{n=-\infty}^{\infty} s(n) \cdot s(n-i) \qquad \forall i \in [1,p] \end{aligned}$$

Par définition, la fonction $\varphi(i)$ d'autocorrélation d'un signal réel et à énergie finie $s(n)$ est

3.2.6
$$\varphi(i) = \sum_{n=-\infty}^{\infty} s(n) \cdot s(n-i) \qquad \forall i$$

En identifiant les termes il vient

3.2.7
$$\sum_{k=1}^{p} a_k \cdot \varphi(k-i) = \varphi(i) \qquad \forall i \in [1,p]$$

■ 3.2.4 Approximation

Dans la pratique, il est impossible de connaître la fonction d'autocorrélation $\varphi(i)$ d'un signal naturel tel que de la parole, par exemple, en raison du nombre infini d'échantillons nécessaires; tout au plus peut on se livrer à une estimation à court terme des valeurs prises par l'équation 3.2.6. Pour cela, il est nécessaire

de savoir se contenter de connaître $s(n)$ à travers une fenêtre de durée finie. Deux possibilités s'offrent à nous: ou bien nous acceptons les effets de la distorsion spectrale introduite par la convolution directe d'une fenêtre avec le signal, ou bien nous décidons de ne prendre en compte ces effets que de manière détournée, par l'application d'une fenêtre au calcul de l'énergie E plutôt que par la pondération de $s(n)$. Ces deux choix sont connus dans la littérature sous le nom de technique d'autocorrélation pour la première, et technique de la covariance pour la seconde.

3.2.5 Technique de l'autocorrélation

Contemplons $s(n)$ à travers une fenêtre de N échantillons au plus. En raison de la réciprocité de la multiplication dans le domaine temporel et de la convolution dans le domaine des fréquences, il est souhaitable que le spectre d'amplitude de cette fenêtre se présente sous la forme d'un pic central étroit flanqué de lobes secondaires aussi peu marqués que possible. La largeur du pic imposera la résolution en fréquence de l'estimation, tandis que de l'amplitude relative des lobes secondaires et du pic central dépendront les imprécisions traduites par les phénomènes de recouvrement. L'offre, en matière de fenêtres de pondération, est abondante [78RabL, 85MaxJ, 84CouF, 84KunM, 87EllD]. En voici les familles principales:

- fenêtre polynomiale, obtenue par convolution successive $(m-1)$ d'une fenêtre rectangulaire de durée N/m avec elle-même. L'atténuation des lobes secondaires peut devenir considérable, au prix de l'élargissement du pic central. La fenêtre rectangulaire $(m=1)$ et la fenêtre de Bartlett $(m=2)$ font partie de cette famille;
- fenêtre de Blackman généralisée, dont les fenêtres de Hanning, de Hamming et de Blackman sont des cas particuliers. Pour une atténuation des lobes secondaires identique à celle obtenue par une fenêtre polynomiale, l'élargissement du pic central est généralement moindre;
- fenêtre de Kaiser, qui permet de spécifier explicitement le compromis entre l'atténuation des lobes secondaires et la largeur du pic central.

Nommons $w(n)$ le facteur de pondération associé à une fenêtre de longueur N. Le signal que nous désirons analyser, c'est-à-dire dont nous voulons découvrir les paramètres propres à sa synthèse, devient

3.2.8
$$\begin{cases} x(n) = w(n) \cdot s(n) & \forall n \in [0,N[\\ x(n) = 0 & \forall n \notin [0,N[\end{cases}$$

En substituant $x(n)$ à $s(n)$ dans les équations 3.2.4 et 3.2.5, nous sommes amené à estimer la fonction d'autocorrélation 3.2.6 par

3.2.9 $$R(i) = \sum_{n=i}^{N-1} x(n) \cdot x(n-i) = \sum_{n=-\infty}^{\infty} w(n) \cdot s(n) \cdot w(n-i) \cdot s(n-i) \qquad \forall i \in [1, p]$$

En faisant usage de cette nouvelle définition et en utilisant la notation matricielle, le système 3.2.7 devient

3.2.10 $$\mathbf{R}\,\mathbf{a} = \mathbf{r}$$

On reconnaît là l'équation de Yule-Walker, où la matrice caractéristique est du type Toeplitz symétrique, dont la particularité est de posséder des coefficients constants pour chaque diagonale. La solution de cette équation est particulièrement efficace, en terme de nombre d'opération, si l'on fait usage de l'algorithme de Levinson-Durbin, ou de celui apparenté de Schur-Leroux-Gueguen. En conclusion, nous avons retenu la technique de l'autocorrélation pour mener l'analyse par prédiction linéaire.

3.2.6 Existence d'une solution

Constatons d'emblée que le système linéaire 3.2.7 ne possède pas forcément de solution: il suffit de considérer le cas dégénéré d'un signal $S(z)$ identiquement nul pour s'en convaincre, car alors la fonction d'autocorrélation est identiquement nulle, et la matrice associée du système aussi. De même, des cas moins dégénérés peuvent conduire à une matrice dont le rang est inférieur à l'ordre d'analyse p; toutefois, on observe en pratique que les signaux naturels de parole sont de nature à rendre possible l'inversion de $\mathbf{R}$. Confrontés au cas contraire, nous proposons la démarche suivante: réduire l'ordre d'analyse jusqu'à obtenir un système pour lequel il existe une solution. Dans le pire des cas, tous les coefficients a_k sont nuls, à l'exception de a_0 qui reste unitaire en vertu de 3.2.1.

3.2.7 Stabilité du filtre de synthèse

Le filtre de synthèse étant un filtre récursif, la question de sa stabilité est pertinente; cependant, du fait que nous avons défini l'analyse comme une recherche des paramètres de synthèse aptes à un codage exact découle que les implications d'une éventuelle instabilité de $H(z)$ sont mineures, vu qu'elle sera compensée par le comportement de $U(z)$. Toutefois, ceci ne serait vrai que si la

précision des calculateurs était infinie; or, dans la réalité, cette précision est finie. La sagesse veut donc que l'on renonce à utiliser un filtre instable, par exemple par diminution de l'ordre d'analyse jusqu'à l'apparition de la stabilité. Dans le pire des cas, tous les coefficients a_k sont nuls, à l'exception de a_0 qui reste unitaire en vertu de 3.2.1.

Une approche plus classique est de forcer la stabilité par une action sur les paramètres internes, nommés coefficients de réflexion, que l'on rencontre dans les algorithmes de Levinson-Durbin et de Schur-Leroux-Gueguen. Toutefois, si l'on suit cette voie alors le critère exprimé au paragraphe 3.2.2 n'est plus toujours vérifié. Une autre approche encore est d'adjoindre à l'erreur de prédiction progressive de l'expression 3.2.4 une erreur de prédiction rétrograde; il devient alors possible d'assurer la stabilité du filtre engendré par la minimisation de la combinaison de ces deux erreurs. Selon la nature de la combinaison, l'algorithme résultant est dit de Burg ou d'Itakura; on parle aussi parfois de covariance-treillis pour le premier des deux.

3.2.8 Modèle autorégressif classique

Jusqu'à maintenant, nous avons développé l'analyse par prédiction linéaire selon un schéma se rapportant au codage exact. Nous allons désormais passer à un schéma se rapportant au domaine de la synthèse de parole. Considérons pour cela la figure 3.2.b, où seul le modèle de l'excitation est modifié par rapport à celui rencontré à la figure 3.2.a.

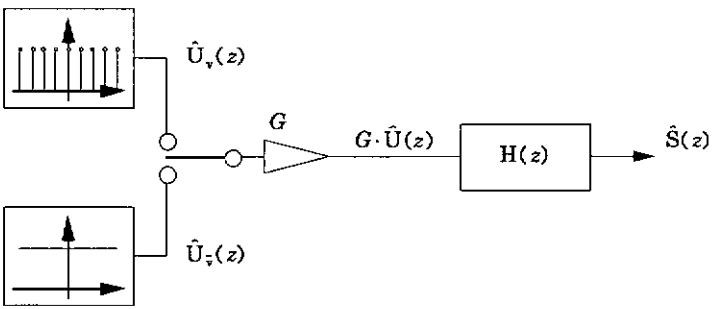

Figure 3.2.b Synthétiseur de signal

Dans cette figure, on reconnaît un sélecteur qui permet de choisir entre deux sources de signaux. La première génère un train d'impulsions espacées d'une

période $1/F_0$, tandis que la seconde génère un bruit blanc stationnaire de moyenne nulle et de variance unité. L'excitation est obtenue par la multiplication de l'une de ces deux sources par un facteur de gain scalaire G. Qualitativement, ce modèle cherche à copier le mécanisme d'émission de la voix humaine: les impulsions glottales du mode d'émission des sons laryngés sont associées aux impulsions du premier générateur $\hat{U}_v(z)$, l'excitation turbulente du mode d'émission des sons dévoisés est associée au second générateur $\hat{U}_{\bar{v}}(z)$, autant la réponse spectrale du conduit vocal que la forme d'onde des impulsions vocales sont associées au filtre de synthèse $H(z)$, et enfin l'intensité sonore est reflétée par la valeur du gain G [70FanG, 72FlaJ].

■ 3.2.9 Fidélité de la synthèse

Quantitativement, si nous analysons un signal à imiter $S(z)$ de sorte à déterminer $(H(z),U(z))$, alors nous verrons que le signal synthétique $\hat{S}(z)$ obtenu par le filtrage de $G\cdot\hat{U}(z)$ par $H(z)$ est proche du signal original, pour autant qu'il s'agisse de parole, et au sens d'une métrique à définir. Nous allons justifier formellement cette affirmation dans le cas d'un son dévoisé, puis y donner une justification empirique dans le cas d'un son laryngé. Considérons tout d'abord la réponse impulsionnelle $h(n)$ du filtre de synthèse. Par hypothèse, ce filtre récursif fournit une sortie identiquement nulle au moins jusqu'à l'instant $n=0$ où il est nourri d'une impulsion unité, puis continue en roue libre. La substitution dans l'expression 3.2.3 de l'excitation $u(n)$ par une impulsion unitaire de Dirac $\delta(n)$ fournit

3.2.11
$$h(n)=\begin{cases}0 & n<0\\ 1 & n=0\\ \sum_{k=1}^{n} a_k\cdot h(n-k) & \forall n\in[1,p]\\ \sum_{k=1}^{p} a_k\cdot h(n-k) & n>p\end{cases}$$

Tout signal $\hat{s}(n)$ issu du synthétiseur présenté à la figure 3.2.b peut être considéré comme égal à la convolution de la réponse impulsionnelle $h(n)$ du filtre de synthèse avec l'excitation $G\cdot\hat{u}(n)$. Ceci se traduit par

3.2.12
$$\hat{s}(n)=h(n)*G\cdot\hat{u}(n)=G\cdot\sum_{j=-\infty}^{\infty}h(j)\cdot\hat{u}(n-j)\qquad\forall n$$

■ 3.2.10 Synthèse de sons dévoisés

La corrélation croisée entre l'entrée normalisée et la sortie du filtre de synthèse est égale à sa réponse impulsionnelle multipliée par le gain, à condition que l'excitation soit un bruit blanc stationnaire de moyenne nulle et de variance unité, multiplié par le même gain. En introduisant 3.2.12 dans 3.2.6, il vient

$$\begin{aligned}\varphi_{\hat{u}\hat{s}}(i) &= \sum_{n=-\infty}^{\infty}\hat{u}(n)\cdot\hat{s}(n+i) = \sum_{n=-\infty}^{\infty}\hat{u}(n)\cdot G\cdot\sum_{j=-\infty}^{\infty}h(j)\cdot\hat{u}(n+i-j)\\ &= G\cdot\sum_{j=-\infty}^{\infty}h(j)\cdot\sum_{n=-\infty}^{\infty}\hat{u}(n)\cdot\hat{u}(n+i-j) = G\cdot\sum_{j=-\infty}^{\infty}h(j)\cdot\varphi_{\hat{u}\hat{u}}(i-j)\\ &= G\cdot\sum_{j=-\infty}^{\infty}h(j)\cdot\delta(i-j) = G\cdot h(i) \qquad \forall i\end{aligned} \tag{3.2.13}$$

La valeur $\varphi_{\hat{s}\hat{s}}$ de la fonction d'autocorrélation du signal $\hat{s}(n)$ synthétisé par l'application de 3.2.3 devient par conséquent

$$\begin{aligned}\varphi_{\hat{s}\hat{s}}(i) &= \sum_{n=-\infty}^{\infty}\hat{s}(n)\cdot\hat{s}(n+i) = \sum_{n=-\infty}^{\infty}\left(G\cdot\hat{u}(n)+\sum_{k=1}^{p}a_k\cdot\hat{s}(n-k)\right)\cdot\hat{s}(n+i)\\ &= G\cdot\sum_{n=-\infty}^{\infty}\hat{u}(n)\cdot\hat{s}(n+i)+\sum_{k=1}^{p}a_k\cdot\sum_{n=-\infty}^{\infty}\hat{s}(n-k)\cdot\hat{s}(n+i)\\ &= G\cdot\varphi_{\hat{u}\hat{s}}(i)+\sum_{k=1}^{p}a_k\cdot\varphi_{\hat{s}\hat{s}}(k+i) = G^2\cdot h(i)+\sum_{k=1}^{p}a_k\cdot\varphi_{\hat{s}\hat{s}}(k+i) \qquad \forall i\end{aligned} \tag{3.2.14}$$

Il est utile de se remémorer que la fonction d'autocorrélation d'un signal réel est paire. Par introduction de cette propriété dans 3.2.14, nous avons

$$\varphi_{\hat{s}\hat{s}}(i) = \varphi_{\hat{s}\hat{s}}(-i) = G^2\cdot h(-i)+\sum_{k=1}^{p}a_k\cdot\varphi_{\hat{s}\hat{s}}(k-i) \qquad \forall i \tag{3.2.15}$$

Considérons maintenant le signal $s(n)$ à imiter, et penchons-nous plus particulièrement sur la valeur à l'origine de sa fonction d'autocorrélation. Par identification avec l'expression 3.2.4, il vient

$$\begin{aligned}\varphi_{ss}(0) = \varphi(0) &= \sum_{n=-\infty}^{\infty}s(n)\cdot s(n)\\ &= \mathrm{E}-\sum_{n=-\infty}^{\infty}\left(\sum_{k=1}^{p}a_k\cdot s(n-k)\right)^2+\sum_{n=-\infty}^{\infty}2\cdot s(n)\cdot\sum_{k=1}^{p}a_k\cdot s(n-k)\\ &= \mathrm{E}-\sum_{n=-\infty}^{\infty}\sum_{i=1}^{p}\sum_{k=1}^{p}a_i\cdot a_k\cdot s(n-k)\cdot s(n-i)+2\cdot\sum_{k=1}^{p}a_k\cdot\sum_{n=-\infty}^{\infty}s(n)\cdot s(n-k)\end{aligned} \tag{3.2.16}$$

Après substitution des termes par leur définition 3.2.6 et en utilisant la relation 3.2.7, il vient

$$\begin{aligned}\varphi_{ss}(0) &= \mathrm{E} - \sum_{i=1}^{p}\sum_{k=1}^{p} a_i \cdot a_k \cdot \varphi_{ss}(k-i) + 2 \cdot \sum_{k=1}^{p} a_k \cdot \varphi_{ss}(k) \\ &= \mathrm{E} - \sum_{i=1}^{p} a_i \cdot \sum_{k=1}^{p} a_k \cdot \varphi_{ss}(k-i) + 2 \cdot \sum_{k=1}^{p} a_k \cdot \varphi_{ss}(k) \\ &= \mathrm{E} - \sum_{i=1}^{p} a_i \cdot \varphi_{ss}(i) + 2 \cdot \sum_{i=1}^{p} a_i \cdot \varphi_{ss}(i) = \mathrm{E} + \sum_{k=1}^{p} a_k \cdot \varphi_{ss}(k)\end{aligned} \qquad 3.2.17$$

Par comparaison de 3.2.17 avec 3.2.15, où il a été fait usage de 3.2.11, nous pouvons constater un intéressant parallèle

$$\begin{aligned}\varphi_{\hat{s}\hat{s}}(0) &= G^2 + \sum_{k=1}^{p} a_k \cdot \varphi_{\hat{s}\hat{s}}(k) \\ \varphi_{ss}(0) &= \mathrm{E} + \sum_{k=1}^{p} a_k \cdot \varphi_{ss}(k)\end{aligned} \qquad 3.2.18$$

Par comparaison de 3.2.7 avec 3.2.15, où nous avons introduit le fait que la réponse impulsionnelle est nulle pour un argument négatif, le parallèle se précise

$$\begin{aligned}\varphi_{ss}(i) &= \sum_{k=1}^{p} a_k \cdot \varphi_{ss}(k-i) \qquad \forall i \in [1, p] \\ \varphi_{\hat{s}\hat{s}}(i) &= \sum_{k=1}^{p} a_k \cdot \varphi_{\hat{s}\hat{s}}(k-i) \qquad i > 0\end{aligned} \qquad 3.2.19$$

Nous concluons des relations 3.2.18 et 3.2.19 que si l'excitation est un bruit blanc stationnaire de moyenne nulle et de variance unité, multiplié par un gain égal à la racine carrée de l'erreur quadratique de prédiction progressive, et si cette même erreur quadratique est minimale par rapport aux coefficients de prédiction, alors le signal original et le signal synthétique ont la même fonction d'autocorrélation jusqu'à l'ordre d'analyse

$$\varphi_{ss}(i) = \varphi_{\hat{s}\hat{s}}(i) \qquad \forall i \in [0, p] \qquad 3.2.20$$

Ceci implique que la densité spectrale d'énergie, égale à la transformée de Fourier de la fonction d'autocorrélation, est identique pour les deux signaux, à condition qu'on la contemple avec une résolution donnée par $p+1$. Le filtre de synthèse $\mathbf{H}(z)$ est donc le filtre méromorphe d'ordre p qui approche au mieux

le comportement de l'amplitude spectrale du signal analysé. Comme promis au paragraphe 3.2.2, la description de la structure générale (l'enveloppe) de $S(z)$ est maintenant décrite avec un nombre restreint de coefficients pertinents.

En résumé, nous avons imité de la parole dévoisée par un signal issu du passage, dans le filtre de synthèse, d'une excitation artificielle constituée par un bruit blanc qui possède par définition un spectre de puissance constante. L'opération de filtrage a vu ce spectre plat être multiplié par celui du filtre de synthèse; le spectre de puissance du signal de parole synthétique a donc la même apparence que le spectre d'énergie de $H(z)$, au facteur de gain près. Or, nous venons de voir que le spectre d'amplitude du filtre de synthèse approche de façon optimale celui du signal à imiter. Il s'ensuit que le modèle d'analyse par prédiction linéaire est bien adapté à l'analyse des signaux de parole, pour autant qu'ils correspondent à une émission dévoisée.

3.2.11 Synthèse de sons laryngés

Pour un son laryngé, le spectre d'amplitude de la parole est un spectre de raies harmoniques, la morphologie de chaque raie étant partiellement dépendante de la forme de l'onde glottale. Or, l'examen de la figure 3.2.b montre que, pour un son déterminé comme laryngé, le train d'impulsions associé à l'excitation ne fait rien d'autre que d'échantillonner le spectre du filtre de synthèse; il s'ensuit que le spectre du signal synthétique est aussi un spectre de raies harmoniques, où chaque raie est de largeur infinitésimale. Une autre justification empirique de la justesse du choix du critère de minimisation de l'énergie de l'excitation apparaît en considérant l'expression 3.2.3: si le son est laryngé, et si le filtre de synthèse est un bon modèle du conduit vocal, alors l'excitation doit apparaître sous la forme d'impulsions glottales espacées de la période du fondamental. Dans les conditions d'analyse rencontrées dans la pratique, la durée d'une impulsion est courte par rapport à cet espacement; il s'ensuit que le signal $u(n)$ reste longtemps de petite amplitude et ne prend que brièvement une grande amplitude. Il n'est donc pas absurde d'aller jusqu'à minimiser son énergie, même pour un son laryngé.

Pour juger de la qualité de l'approximation spectrale du signal naturel par le signal synthétique pour une section de son laryngé, nous pouvons considérer l'application directe du critère choisi. Si nous appliquons le théorème de Parseval à l'énergie définie en 3.2.4, et si nous considérons la relation 3.2.2, alors

3.2.21 $$E = \sum_{n=-\infty}^{\infty} u^2(n) = \frac{1}{2\pi} \cdot \int_{-\pi}^{\pi} \left|U(e^{j\omega})\right|^2 d\omega = \frac{1}{2\pi} \cdot \int_{-\pi}^{\pi} \frac{\left|S(e^{j\omega})\right|^2}{\left|H(e^{j\omega})\right|^2} d\omega$$

Minimiser l'erreur quadratique progressive revient donc à minimiser la moyenne, sur toutes les fréquences, du rapport entre le spectre d'énergie du signal et celui du filtre de synthèse, dont la croissance est limitée par la condition de normalisation $a_0 = 1$ rencontrée à la définition 3.2.1. Nous avons vu que le spectre d'amplitude de $H(z)$ tend à être similaire à celui de $S(z)$ pour un signal où l'excitation est bien modélisée par un bruit blanc; or, si le bon sens nous permet de supposer que c'est aussi le cas pour une excitation donnée par un train d'impulsions glottales (au moins pour l'ordre de grandeur du spectre), alors il existe des fréquences ω où $\left|H(e^{j\omega})\right|^2$ excède $\left|S(e^{j\omega})\right|^2$, et réciproquement. L'examen de 3.2.21 montre que les excès du spectre d'énergie du filtre de synthèse contribuent peu à l'erreur quadratique, tandis que la contribution est plus importante aux fréquences où $\left|H(e^{j\omega})\right|^2$ sous-estime $\left|S(e^{j\omega})\right|^2$. C'est donc cette dernière condition qui sera la plus touchée par l'opération de minimisation de E. Nous en concluons que si le filtre de synthèse peut, à lui seul, reproduire convenablement le comportement du spectre d'amplitude d'un signal de parole dévoisée, il est toutefois contraint de se contenter d'approcher l'enveloppe d'amplitude spectrale d'un signal de parole laryngée juste en dessous de ses pics harmoniques, les vallées étant peu représentées en raison de la grandeur de l'ordre d'analyse, choisi d'ordinaire justement de sorte à ce que le filtre de synthèse possède un spectre d'amplitude qui soit une version lisse de celui du signal.

■ 3.2.12 Gain du système

La valeur du gain G de la figure 3.2.b découle directement de l'identification des termes de l'expression 3.2.18. Pour la technique de l'autocorrélation on remplacera $\varphi(k)$ par son estimation à court terme $R(k)$, et par son estimation pondérée pour la technique de la covariance. Nous avons donc

3.2.22 $$G = \sqrt{\varphi(0) - \sum_{k=1}^{P} a_k \cdot \varphi(k-i)}$$

L'énergie de la source $u(n)$ est par définition unitaire. L'énergie totale du système est donc répartie à la fois dans l'excitation, par le biais du coefficient de gain G, et dans le filtre de synthèse, dont la normalisation porte seulement sur le coefficient particulier a_0 qui doit être unitaire. Par conséquent, l'énergie as-

sociée à la réponse impulsionnelle $h(n)$ n'est pas indépendante des coefficients d'autocorrélation extraits du signal analysé. En pratique, on observe toutefois pour un signal de parole que les variations d'énergie se reflètent plus volontiers dans le signal de gain que dans la mesure de l'énergie intrinsèque à la représentation normale du filtre de synthèse.

■ 3.2.13 Codage et décodage exact

Nous savons désormais, à l'issue du paragraphe 3.2, appliquer à de la parole l'analyse par prédiction linéaire, ce qui nous permet de construire le système de codage-décodage exact de la figure 3.2.c.

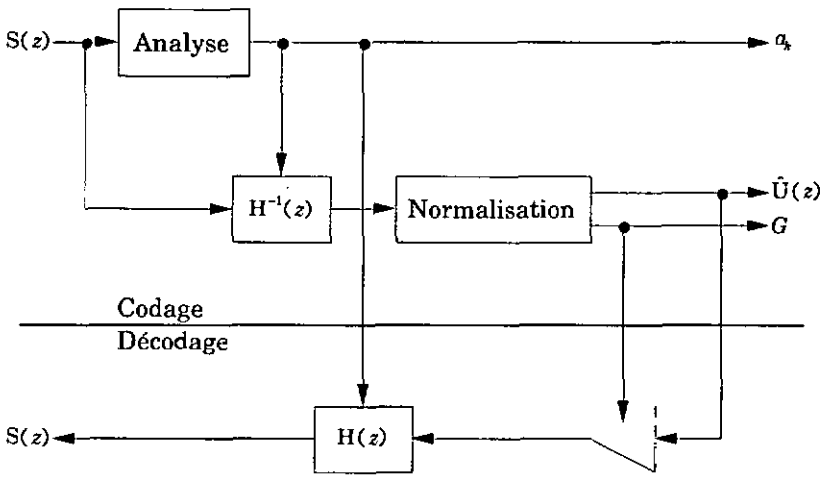

Figure 3.2.c Codage-décodage par prédiction linéaire

■ 3.2.14 Exemple

La partie de gauche de la figure 3.2.d montre un exemple d'analyse de parole naturelle dévoisée, extraite du son [ʃ] du mot «chat»=[ʃa]. On y distingue d'abord le signal temporel $s(n)\cdot w(n)$, suivi de son spectre d'amplitude $|S(e^{j\omega})|$. On distingue ensuite $|H(e^{j\omega})|$ le spectre d'amplitude du filtre de synthèse obtenu de l'analyse par prédiction linéaire de ce signal par la technique de l'autocorrélation, puis $|U(e^{j\omega})|$ le spectre d'amplitude de l'excitation au sens du codage exact énoncé au paragraphe 3.2.2. Enfin, on trouve $u(n)$, qui est l'excitation dans le domaine temporel. La partie de droite de cette même figure montre quels sont les résultats pour la partie laryngée [a] du mot «chat»=[ʃa].

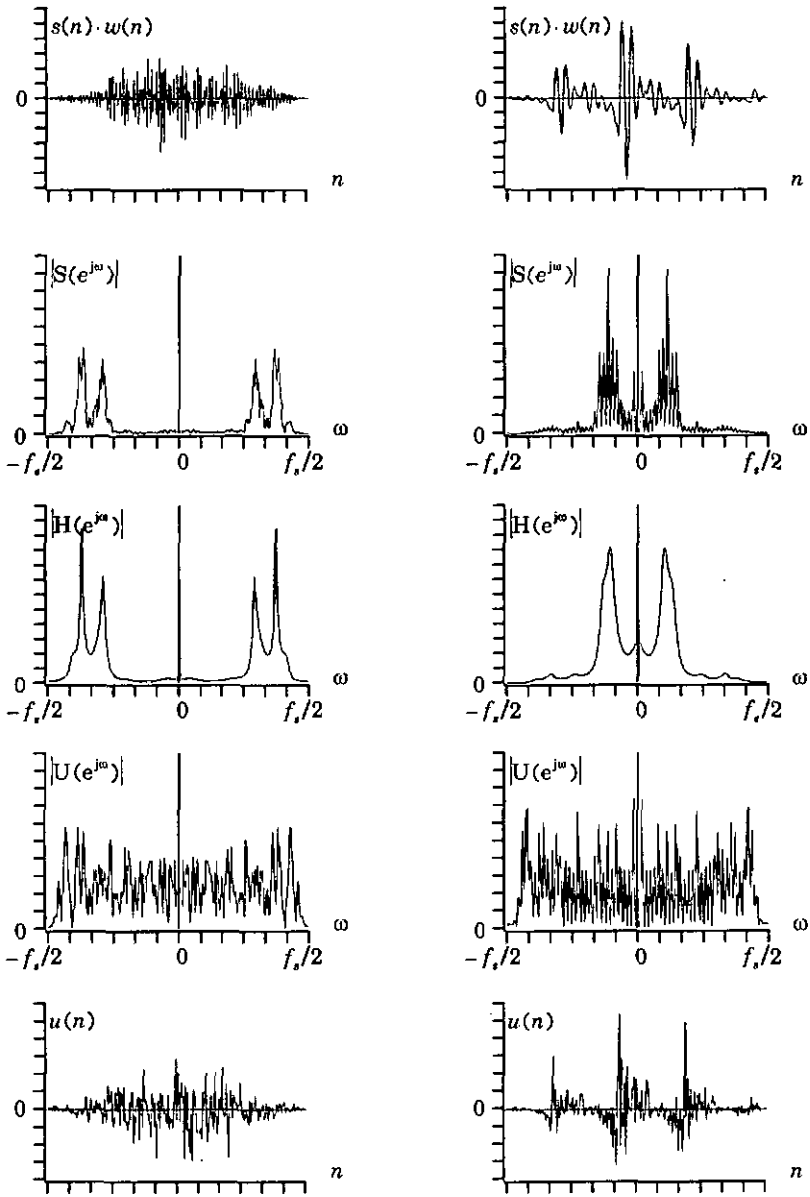

Figure 3.2.d Analyse d'un signal naturel

La localisation temporelle d'impulsions glottales est particulièrement aisée dans le signal d'excitation $u(n)$ de la partie droite de la figure 3.2.d. Cette observation conforte les considérations du paragraphe 3.2.11 tendant à justifier la validité du modèle de la figure 3.2.b pour un cas laryngé. Pour un cas dévoisé, la condition d'existence d'une excitation qui soit assimilable à un bruit blanc, posée au paragraphe 3.2.10, est partiellement vérifiée car l'enveloppe de $|U(e^{j\omega})|$ paraît plate pour la partie gauche de la figure 3.2.d.

Dans cet exemple, l'ordre d'analyse est $p = 14$, la technique de l'autocorrélation a été appliquée, la cadence d'échantillonnage est $f_e = 8000$ Hz et le support de la fenêtre triangulaire $w(n)$ choisie (fenêtre de Bartlett) vaut 0.030 s. Le choix de ces valeurs de paramètres est classique.

3.3 Analyse cepstrale

Le cepstre est une représentation des périodicités d'un signal, comme le spectre est une représentation du contenu en fréquences d'un signal. Sa particularité est de permettre l'analyse homomorphique, dont l'intérêt principal réside dans la capacité d'effectuer une opération de déconvolution [68OppA], sous réserve de la satisfaction d'un certain nombre de précautions trop souvent négligées [86VerW]. Le cepstre complexe d'un signal s'obtient par le calcul de la transformée en z inverse du logarithme du spectre du signal, le cepstre réel étant la transformée en z inverse du logarithme du spectre d'amplitude du signal. Évaluer ces transformées sur le cercle unité $|z|^2 = 1$ revient à calculer des transformations de Fourier; dans ce cas, si le signal en question est lui-même réel, alors son spectre d'amplitude est de symétrie paire et son spectre de phase est de symétrie impaire. Il s'ensuit que la partie réelle du logarithme de son spectre est de symétrie paire, de même que la partie imaginaire du logarithme de son spectre est de symétrie impaire. Nous en concluons que le cepstre d'un signal réel est réel. De même, le cepstre réel d'un signal réel est lui aussi réel; il ne représente toutefois que la contribution de la partie à phase minimale qui possède le même spectre d'amplitude que le signal. Par contre, le cepstre réel d'un signal complexe est complexe, car la symétrie du spectre d'amplitude est perdue.

3.3.1 Cepstre complexe d'un filtre de synthèse

L'opération logarithmique introduite fait usage d'une fonction non linéaire dont le développement en série de Taylor est infini. Intuitivement, ceci a pour

conséquence de rendre infini le support de la représentation cepstrale d'un signal, même si ce dernier est à support fini; cependant nous verrons que les p premiers coefficients suffisent à le caractériser, à condition qu'il satisfasse un modèle autorégressif. Si les valeurs $c(n)$ décrivent le cepstre complexe de la réponse impulsionnelle d'un filtre de synthèse par prédiction linéaire, alors nous avons les relations récursives suivantes

3.3.1 $$a_k \rightarrow c(n) \qquad \begin{cases} c(1) = a_1 & n = 1 \\ c(n) = a_n + \sum_{k=1}^{n-1} \frac{n-k}{n} \cdot a_k \cdot c(n-k) & \forall n \in]1, p] \\ c(n) = \sum_{k=1}^{p} \frac{n-k}{n} \cdot a_k \cdot c(n-k) & n > p \end{cases}$$

3.3.2 $$c(n) \rightarrow a_k \qquad \begin{cases} a_1 = c(1) & k = 1 \\ a_k = c(k) + \sum_{n=1}^{k-1} \frac{n-k}{k} \cdot a_n \cdot c(k-n) & \forall k \in]1, p] \end{cases}$$

L'utilisation de la représentation cepstrale complexe du filtre de synthèse étant fructueuse dans la pratique, nous allons donner au paragraphe suivant le développement qui permet d'obtenir les relations de récurrence 3.3.1. La récurrence 3.3.2 découle directement de la précédente.

■ 3.3.2 Calcul du cepstre complexe du filtre de synthèse

Soit $C(z)$ la transformée en z du cepstre complexe de la réponse impulsionnelle du filtre de synthèse. Il est licite de n'en considérer que les coefficients d'argument n positif, car par hypothèse le filtre de synthèse est stable, possède donc tous ses pôles et ses zéros à l'intérieur du cercle unité et est par conséquent à phase minimale. Dans ce cas, son cepstre complexe est réel et causal. Notre but est d'établir les valeurs $c(n)$ en fonction des coefficients a_k

3.3.3 $$C(z) = \ln(1/A(z)) = c(0) + \sum_{n=1}^{\infty} c(n) \cdot z^{-n}$$

Rappelons l'expression du filtre inverse donnée en 3.2.1 avec $q = 0$

3.3.4 $$A(z) = 1 - \sum_{k=1}^{p} a_k \cdot z^{-k}$$

La dérivée de cette expression selon la variable complexe z donne

3.3.5
$$\frac{\partial A(z)}{\partial z} = -\sum_{k=1}^{p} -k \cdot a_k \cdot z^{-(k+1)}$$

La dérivée selon z de $C(z)$ la transformée en z du cepstre donnée à l'expression 3.3.3 devient

3.3.6
$$\frac{\partial C(z)}{\partial z} = \sum_{n=1}^{\infty} -n \cdot c(n) \cdot z^{-(n+1)}$$
$$= \frac{\partial \ln(1/A(z))}{\partial z} = \frac{\frac{-1}{A^2(z)} \cdot \frac{\partial A(z)}{\partial z}}{1/A(z)} = -\frac{\partial A(z)}{\partial z} \cdot \frac{1}{A(z)}$$

Finalement, la dérivée du filtre inverse est mise sous la forme d'un polynôme en z où apparaissent les valeurs cherchées $c(n)$

3.3.7
$$\frac{\partial A(z)}{\partial z} = -A(z) \cdot \frac{\partial C(z)}{\partial z} = -\left(1 - \sum_{k=1}^{p} a_k \cdot z^{-k}\right) \cdot \sum_{n=1}^{\infty} -n \cdot c(n) \cdot z^{-(n+1)}$$
$$= \sum_{n=1}^{\infty} n \cdot c(n) \cdot z^{-(n+1)} - \sum_{n=1}^{\infty} \sum_{k=1}^{p} n \cdot a_k \cdot c(n) \cdot z^{-(n+k+1)}$$

Après avoir permuté l'ordre de sommation, changeons la variable de sommation $n \rightarrow i = n + k$

3.3.8
$$\frac{\partial A(z)}{\partial z} = \sum_{n=1}^{\infty} n \cdot c(n) \cdot z^{-(n+1)} - \sum_{k=1}^{p} \sum_{n=1}^{\infty} n \cdot a_k \cdot c(n) \cdot z^{-(n+k+1)}$$
$$= \sum_{n=1}^{\infty} n \cdot c(n) \cdot z^{-(n+1)} - \sum_{k=1}^{p} \sum_{i=k+1}^{\infty} (i-k) \cdot a_k \cdot c(i-k) \cdot z^{-(i-k+k+1)}$$

Séparons le dernier terme de la somme dont l'index est k et exprimons la somme dont l'index est i sous la forme de deux sommes partielles

3.3.9
$$\frac{\partial A(z)}{\partial z} = \sum_{n=1}^{\infty} n \cdot c(n) \cdot z^{-(n+1)} -$$
$$- \sum_{k=1}^{p-1} \left(\sum_{i=k+1}^{p} (i-k) \cdot a_k \cdot c(i-k) \cdot z^{-(i+1)} + \sum_{i=p+1}^{\infty} (i-k) \cdot a_k \cdot c(i-k) \cdot z^{-(i+1)} \right)$$
$$- \sum_{i=p+1}^{\infty} (i-p) \cdot a_p \cdot c(i-p) \cdot z^{-(i+1)}$$

Le développement permet la permutation de l'ordre de sommation d'un des termes, et par identification le dernier terme peut être incorporé dans une des sommes. Les relations algébriques correspondantes sont

3.3.10
$$\frac{\partial A(z)}{\partial z} = \sum_{n=1}^{\infty} n \cdot c(n) \cdot z^{-(n+1)} - \sum_{k=1}^{p-1} \sum_{i=k+1}^{p} (i-k) \cdot a_k \cdot c(i-k) \cdot z^{-(i+1)} - \\ - \sum_{i=p+1}^{\infty} \sum_{k=1}^{p-1} (i-k) \cdot a_k \cdot c(i-k) \cdot z^{-(i+1)} - \sum_{i=p+1}^{\infty} (i-p) \cdot a_p \cdot c(i-p) \cdot z^{-(i+1)}$$

3.3.11
$$\frac{\partial A(z)}{\partial z} = \sum_{n=1}^{\infty} n \cdot c(n) \cdot z^{-(n+1)} - \sum_{k=1}^{p-1} \sum_{i=k+1}^{p} (i-k) \cdot a_k \cdot c(i-k) \cdot z^{-(i+1)} - \\ - \sum_{i=p+1}^{\infty} \sum_{k=1}^{p} (i-k) \cdot a_k \cdot c(i-k) \cdot z^{-(i+1)}$$

Développons maintenant le terme central de la relation 3.3.11 en séparant les sommes en sommes partielles

3.3.12
$$\begin{aligned} f(p) &= \sum_{k=1}^{p-1} \sum_{i=k+1}^{p} (i-k) \cdot a_k \cdot c(i-k) \cdot z^{-(i+1)} \\ &= \sum_{k=1}^{p-2} \sum_{i=k+1}^{p-1} (i-k) \cdot a_k \cdot c(i-k) \cdot z^{-(i+1)} + \sum_{k=1}^{p-2} (p-k) \cdot a_k \cdot c(p-k) \cdot z^{-(p+1)} + \\ &\quad + a_{p-1} \cdot c_1 \cdot z^{-(p+1)} \\ &= \sum_{k=1}^{p-2} \sum_{i=k+1}^{p-1} (i-k) \cdot a_k \cdot c(i-k) \cdot z^{-(i+1)} + z^{-(p+1)} \cdot \sum_{k=1}^{p-1} (p-k) \cdot a_k \cdot c(p-k) \qquad p > 2 \end{aligned}$$

Le premier des deux termes résultants n'est autre que $f(p-1)$. Nous pouvons en déduire la construction itérative suivante

3.3.13
$$\begin{aligned} f(p) &= f(p-1) + z^{-(p+1)} \cdot \sum_{k=1}^{p-1} (p-k) \cdot a_k \cdot c(p-k) = f(p-1) + \Delta(p) \\ &= f(2) + \sum_{n=3}^{p-1} f(n) + f(p) - f(2) - \sum_{n=3}^{p-1} f(n) = f(2) + \sum_{n=3}^{p} f(n) - \sum_{n=2}^{p-1} f(n) \\ &= f(2) + \sum_{n=3}^{p} f(n) - \sum_{n=3}^{p} f(n-1) = f(2) + \sum_{n=3}^{p} (f(n) - f(n-1)) \\ &= f(2) + \sum_{n=3}^{p} \Delta(n) \qquad p > 2 \end{aligned}$$

En développant explicitement l'expression 3.3.13, il vient

3.3.14
$$f(p) = a_1 \cdot c(1) \cdot z^{-3} + \sum_{n=3}^{p} z^{-(n+1)} \cdot \sum_{k=1}^{n-1} (n-k) \cdot a_k \cdot c(n-k)$$
$$= \sum_{n=2}^{p} z^{-(n+1)} \cdot \sum_{k=1}^{n-1} (n-k) \cdot a_k \cdot c(n-k) \qquad p \geq 2$$

Par utilisation de la relation 3.3.5, par introduction de l'expression 3.3.14 de $f(p)$ dans l'équation 3.3.11 et par substitution symbolique des variables de sommation $i \rightarrow n$ et $k \rightarrow n$, nous avons

3.3.15
$$\sum_{n=1}^{p} z^{-(n+1)} \cdot n \cdot a_n = \sum_{n=1}^{\infty} z^{-(n+1)} \cdot n \cdot c(n) - \sum_{n=2}^{p} z^{-(n+1)} \cdot \sum_{k=1}^{n-1} (n-k) \cdot a_k \cdot c(n-k)$$
$$- \sum_{n=p+1}^{\infty} z^{-(n+1)} \cdot \sum_{k=1}^{p} (n-k) \cdot a_k \cdot c(n-k)$$

En réarrangeant quelque peu les termes, et en remplaçant la première somme du membre de droite par une somme des sommes partielles, il vient

3.3.16
$$\sum_{n=1}^{p} z^{-(n+1)} \cdot n \cdot a_n = \sum_{n=2}^{p} z^{-(n+1)} \cdot n \cdot c(n) - \sum_{n=2}^{p} z^{-(n+1)} \cdot \sum_{k=1}^{n-1} (n-k) \cdot a_k \cdot c(n-k) +$$
$$+ \sum_{n=p+1}^{\infty} z^{-(n+1)} \cdot n \cdot c(n) - \sum_{n=p+1}^{\infty} z^{-(n+1)} \cdot \sum_{k=1}^{p} (n-k) \cdot a_k \cdot c(n-k) + z^{-2} \cdot c(1)$$

En découpant le membre de gauche en deux sommes partielles et en les incorporant dans le membre de droite, nous pouvons maintenant fusionner les sommes de mêmes bornes

3.3.17
$$0 = \sum_{n=2}^{p} z^{-(n+1)} \cdot \left(n \cdot (c(n) - a_n) - \sum_{k=1}^{n-1} (n-k) \cdot a_k \cdot c(n-k) \right) +$$
$$+ \sum_{n=p+1}^{\infty} z^{-(n+1)} \cdot \left(n \cdot c(n) - \sum_{k=1}^{p} (n-k) \cdot a_k \cdot c(n-k) \right) + z^{-2} \cdot (c(1) - a_1)$$

Cette expression doit rester valide quelle que soit la valeur de la variable complexe z, à condition qu'elle appartienne à la région de convergence de la transformée en z du cepstre considéré. La récurrence 3.3.1 a été construite en considérant que les coefficients multiplicatifs des diverses puissance $n+1$ de $z^{(n+1)}$ ne peuvent être que nuls si cette contrainte doit être satisfaite.

3.4 Cepstre complexe moyen

Le cepstre complexe moyen du filtre de synthèse de l'analyse par prédiction linéaire est un vecteur caractéristique utile en reconnaissance de locuteurs. En accord avec les notations du paragraphe 2.2.1, il est donné composante par composante comme suit

3.4.1 $$\langle c(n)\rangle = \frac{1}{\mathrm{P}} \cdot \sum_{c \in X_n} c \qquad n > 0$$

Dans cette expression, X_n représente l'ensemble des P valeurs observées pour la composante n des cepstres complexes à court terme.

3.5 Cepstre complexe différentiel

La dérivée temporelle du cepstre à court terme du filtre de synthèse peut nous éclairer sur l'aspect dynamique de ce vecteur caractéristique. Or, nous sommes dans un espace échantillonné; ce fait nous empêche de considérer une pente instantanée. Nous nous contenterons donc de l'approximation donnée par

3.5.1 $$\frac{\partial c(n,t)}{\partial t} \cong \frac{\sum_{\tau=-\mathrm{T}}^{\mathrm{T}} \tau \cdot w(\tau) \cdot c(n, t+\tau)}{\sum_{\tau=-\mathrm{T}}^{\mathrm{T}} \tau^2 \cdot w(\tau)} \qquad n > 0 \quad \wedge \quad \forall t$$

Dans cette expression, $c(n,t)$ représente la composante d'ordre n du coefficient cepstral de la fenêtre repérée par t et $w(\tau)$ représente une pondération liée à la fenêtre d'estimation de la pente cepstrale temporelle, qui fait usage de $2 \cdot \mathrm{T} + 1$ fenêtres sur le signal. Le plus souvent, cette fenêtre est simplement rectangulaire

3.5.2 $$w(\tau) = \begin{cases} 0 & |\tau| > \mathrm{T} \\ 1 & |\tau| \leq \mathrm{T} \end{cases}$$

3.6 Résidu

Dans le système de la figure 3.2.c, la grandeur $\hat{U}(z)$ correspondant à l'excitation d'énergie unitaire est appelée résidu. Traditionnellement, ce signal est réputé ne contenir que peu d'information, cette affirmation s'appuyant sur le fait qu'une quantification très grossière de ce résidu autorise déjà une synthèse de parole intelligible pour autant que l'on tire parti d'une structure telle que celle de la figure 3.2.b. Ainsi, on allouera volontiers 1 bit pour représenter le mode de phonation du larynx, et quelques autres pour spécifier la périodicité du train d'impulsions chaque fois que cette spécification s'avérera nécessaire.

Une des originalités de cette thèse tient dans le refus d'un modèle aussi simple. Certes, un certain nombre de chercheurs se sont déjà attaqué au problème d'un codage du résidu offrant une meilleure qualité de synthèse pour un coût modeste en terme de capacité d'information nécessaire; nous songeons en particulier aux techniques de codage à impulsions multiples [82AtaB, 84BerM, 84SinS, 86AtaB, 86BerM, 86OzaK]. Cependant nous sommes les premiers, à notre connaissance, à tenter l'exploitation de ce résidu en tant que vecteur caractéristique potentiellement apte à permettre la reconnaissance de locuteurs.

3.6.1 Extension temporelle du résidu

Le résidu est l'excitation normalisée par le gain de prédiction. En vertu de l'expression 3.2.3, nous avons

3.6.1 $$\hat{u}(n) = \frac{1}{G} \cdot \left(s(n) - \sum_{k=1}^{p} a_k \cdot s(n-k) \right) \qquad \forall n$$

Le gain G est donné par 3.2.22. L'examen de l'expression 3.6.1. montre que le caractère fini ou infini de la durée du résidu dépend étroitement du signal dont il constitue la source. En particulier, la technique de la covariance suppose que $s(n)$ est à support infini; il en sera donc de même pour $\hat{u}(n)$. Par contre, la technique de l'autocorrélation coiffe le signal original d'une fenêtre à support fini; il s'ensuit que le résidu possédera dans ce cas un support de durée égale à la durée de la fenêtre augmentée de l'ordre d'analyse.

Dans ce qui suit, nous avons considéré que nous ne disposons que d'un signal découpé en fenêtres de durée fixe égale au nombre N d'échantillons, et dont

l'origine est à chaque fois déplacée de sorte à correspondre au premier échantillon

3.6.2 $$s(n) \qquad \forall n \in [0,N[$$

En outre, nous avons tronqué le résidu correspondant, ne conservant que les échantillons pour lesquels la valeur de ce signal intervient directement

3.6.3 $$\hat{u}(n) = \frac{1}{G} \cdot \left(s(n) - \sum_{k=1}^{\mathrm{Min}(p,n)} a_k \cdot s(n-k) \right) \qquad \forall n \in [0,N[$$

■ 3.6.2 Rejet de la phase du résidu

Le découpage du signal en fenêtres de durée fixe s'est fait de façon arbitraire, sans chercher à synchroniser la translation de l'origine mentionnée plus haut avec les événements associés aux impulsions glottales caractéristiques d'un son laryngé. Il s'ensuit que ces impulsions apparaissent dans le résidu sans entretenir de relation particulière avec l'origine. On pourrait chercher à les localiser et ainsi découvrir la phase linéaire à appliquer pour synchroniser résidu et impulsions glottales, mais cette approche nous a paru compliquée; en revanche, la voie que nous avons choisie est celle du nœud gordien. Nous nous sommes en effet contenté de supprimer toute contribution de phase en transformant le résidu $\hat{u}(n)$ en son spectre d'amplitude $\left|\hat{U}(k)\right|$

3.6.4 $$\left|\hat{U}(k)\right| = \left| \sum_{n=0}^{N-1} \hat{u}(n) \cdot e^{-j2\pi \cdot n \cdot k/N} \right| \qquad \forall k \in [0,N[$$

Le résidu $\hat{u}(n)$ étant un signal réel, son spectre d'amplitude $\left|\hat{U}(k)\right|$ présente une symétrie paire et ne possède que $N/2+1$ valeurs indépendantes d'amplitude si N est pair, respectivement $(N-1)/2+1$ valeurs si N est impair.

■ 3.6.3 Espace des périodicités

Sous certaines conditions, la minimisation de l'excitation (ou de l'erreur de prédiction progressive) introduite au paragraphe 3.2.3 a pour effet d'aplatir son spectre. Par conséquent, le spectre du résidu est plat lui aussi parce qu'il est identique au précédent à un facteur de normalisation G près; nous en tirons comme conclusion le fait que le spectre d'amplitude du résidu n'est pas d'un intérêt direct. Il faut donc le transformer. Deux possibilités principales se présentent: ou bien nous cherchons à établir le couple (*transformation, métrique*)

dont le pouvoir de discrimination soit maximal, ou bien nous choisissons arbitrairement un des deux éléments du couple et nous nous contentons d'adapter l'autre. Nous devons toutefois convenir à regret que le premier de ces deux cas est en fait trop lourd pour être réalisable dans le contexte d'une reconnaissance de locuteurs indépendante du texte. En outre, nous considérons comme plus sage de choisir une transformation d'abord et d'y adapter une métrique ensuite que de faire le contraire.

L'offre en matière de transformations est vaste. Pourtant, le choix que nous avons fait ne brille pas par son extravagance, puisque ce n'est autre que la très classique transformation inverse de Fourier que nous avons retenue, la justification principale étant justement le fait qu'elle soit bien connue. Là encore, si nous désirons pourtant nous en tenir à la manipulation des objets classiques du traitement de signal, alors deux choix interviennent: ou bien nous l'appliquons sans autre forme de procès et nous nous retrouvons avec la fonction d'autocorrélation du résidu, ou bien nous construisons le cepstre réel du résidu en faisant d'abord usage d'une fonction logarithmique, comme exposé au paragraphe 3.3.

Parmi tous ces choix arbitraires nous avons retenu celui du cepstre réel du résidu, sans motivation particulière. Toutefois, en guise de justification nous rappelons encore une fois que le critère ultime est le pouvoir de discrimination du couple (*transformation, métrique*), et que par conséquent le choix de l'un ou de l'autre des éléments du couple ne peut pas être objectif s'il est mené séparément. Notons cependant que le cepstre réel du résidu possède l'avantage de permettre une interprétation du vecteur caractéristique obtenu puisqu'il représente alors les périodicités du signal. Néanmoins, cet avantage n'est autre que subjectif et ne garantit en aucune façon son efficacité dans la tâche à laquelle il est destiné. Formellement, le cepstre réel du résidu est

3.6.5 $$v(n) = \frac{1}{N} \cdot \sum_{k=0}^{N-1} \mathrm{Ln}\left(\left|\hat{U}(k)\right|\right) \cdot e^{j 2\pi \cdot n \cdot k/N} \qquad \forall n \in [0, N/2]$$

Le spectre d'amplitude du résidu étant réel, positif et symétrique, son logarithme est réel et symétrique. Il s'ensuit que le cepstre réel du résidu l'est aussi; $N/2+1$ éléments suffisent à sa caractérisation si N est pair. L'abscisse de ce cepstre est gradué en unités de temps baptisées quéfrences pour l'occasion.

■ 3.6.4 Exemple

La figure 3.6.a est un complément à la figure 3.2.d et montre successivement le logarithme naturel du spectre d'amplitude du résidu puis son cepstre réel. La disposition conserve à gauche le son [ʃ] et à droite le son [a].

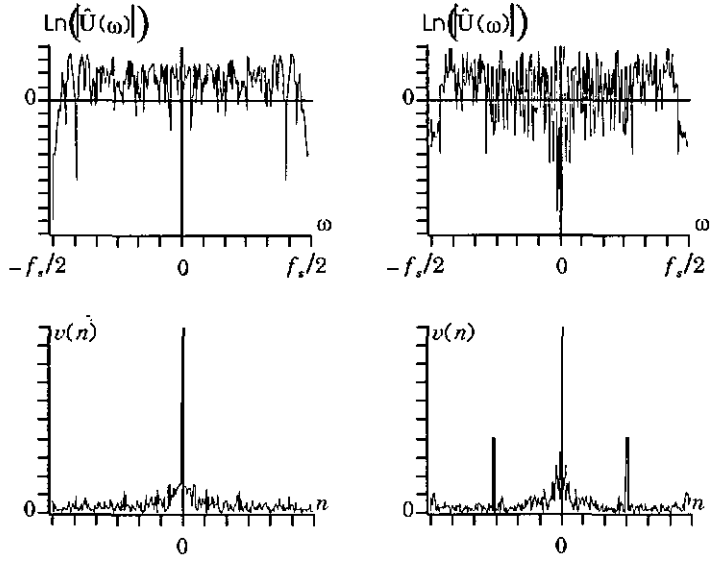

Figure 3.6.a Spectre et cepstre du résidu d'un signal naturel

Le cepstre réel du résidu code la contribution de chaque périodicité présente dans l'excitation. Or, pour un son laryngé, il apparaît presque toujours que la contribution d'une de ces périodicités domine de beaucoup toutes les autres, cette domination étant d'ailleurs bien souvent prise comme fondement d'une décision sur l'estimation du mode de vibration du larynx. En effet, si un son laryngé correspond en principe strictement à une émission sonore où les cordes vocales interviennent, en pratique il est rarissime que ces cordes vocales ne battent pas à une fréquence fondamentale bien définie, sauf état pathologique. Il s'ensuit que l'approximation qui identifie la présence d'une fréquence fondamentale avec l'émission d'un son laryngé est bien acceptée. On remarque ainsi sur la figure 3.6.a que le son laryngé [a] se distingue par la présence de deux

pics symétriques par rapport à l'origine et correspondant à la période fondamentale de l'excitation.

3.7 Cepstre réel moyen du résidu

Le cepstre réel moyen du résidu de l'analyse par prédiction linéaire est un vecteur caractéristique utile en reconnaissance de locuteurs. En accord avec les notations du paragraphe 2.2.1, il est donné composante par composante comme suit

3.7.1 $$\langle v(n) \rangle = \frac{1}{P} \cdot \sum_{v \in X_n} v \qquad \forall n \in [0, N/2]$$

Dans cette expression, X_n représente l'ensemble des P valeurs observées pour la composante n des cepstres réels à court terme.

3.8 Fréquence fondamentale

L'objet de ce chapitre est la description d'une méthode permettant d'extraire la fréquence fondamentale F_0 d'un signal de parole. Après une discussion de la distinction entre pitch et F_0, nous abordons une à une les étapes nécessaires à la détermination de cette dernière grandeur. Il s'agit d'abord de segmenter le signal en silence et parole, puis de qualifier les segments de parole par une fréquence fondamentale et enfin de ne retenir que les éléments correspondant à un son laryngé.

3.8.1 Pitch

La fréquence fondamentale d'un son laryngé est dans une certaine mesure caractéristique d'un locuteur. Cependant, nous ne la percevons d'ordinaire pas directement; elle n'est par exemple pas transmise par un canal téléphonique où une économie de bande passante intervient entre 0 Hz et 300 Hz sans que les auditeurs s'en aperçoivent. Nous sommes en revanche sensibles à l'espacement en fréquence des harmoniques de la fréquence fondamentale car il est clair que nous savons différencier entre une personne à la voix aiguë, de pitch plus élevé que celui d'une personne à la voix grave, même lors d'une conversation téléphonique. Le terme de jargon décrivant la hauteur perçue des sons est emprunté à l'anglais; c'est ce "pitch", qui correspond à la fréquence fondamentale

produisant, ou susceptible de produire un son à la hauteur perçue. La fréquence fondamentale F_0, mesurable, est donc utilisable dans le contexte de la reconnaissance de locuteurs.

3.8.2 Préaccentuation

L'opération de préaccentuation est nécessaire pour le bon fonctionnement de la segmentation en parole et en silence. Son but est simplement de renforcer l'énergie des hautes fréquences qu'il est important de détecter d'une part parce qu'elles signent les portions transitoires du signal de parole, particulièrement au début d'une séquence sonore, et d'autre part parce qu'elles permettent parfois de masquer certaines sources de bruit indésirables telles que le ronronnement à fréquence basse d'un éclairage à tube au néon. Si μ est le coefficient de préaccentuation, alors le signal préaccentué y devient

3.8.1
$$y(n) = \begin{cases} 0 & \forall n \notin]0,N[\\ s(n) - \mu \cdot s(n-1) & \forall n \in]0,N[\end{cases}$$

3.8.3 Segmentation en silence et parole

Associons une énergie $E(t)$ à chaque fenêtre t du signal de parole

3.8.2
$$E^2(t) = \sum_{n=0}^{N-1} \left(s(n,t) - \langle s(t) \rangle \right)^2 \qquad \forall t$$

Dans cette expression, $\langle s(t) \rangle$ est la valeur moyenne du signal de parole pour la fenêtre considérée

3.8.3
$$\langle s(t) \rangle = \frac{1}{N} \cdot \sum_{n=0}^{N-1} s(n,t) \qquad \forall t$$

L'énergie $E(t)$ est comparée à un seuil μ; grossièrement, si le seuil y est supérieur, alors la fenêtre est étiquetée comme silence. Sinon, elle est étiquetée comme parole. Dans le détail, deux seuils adaptatifs sont utilisés, de sorte à construire une hystérésis; en outre, l'adaptation des seuils diffère selon l'étiquette associée provisoirement au signal. Cette étiquette ne devient définitive que lorsque des conditions de durée suffisante sont remplies pour chaque état.

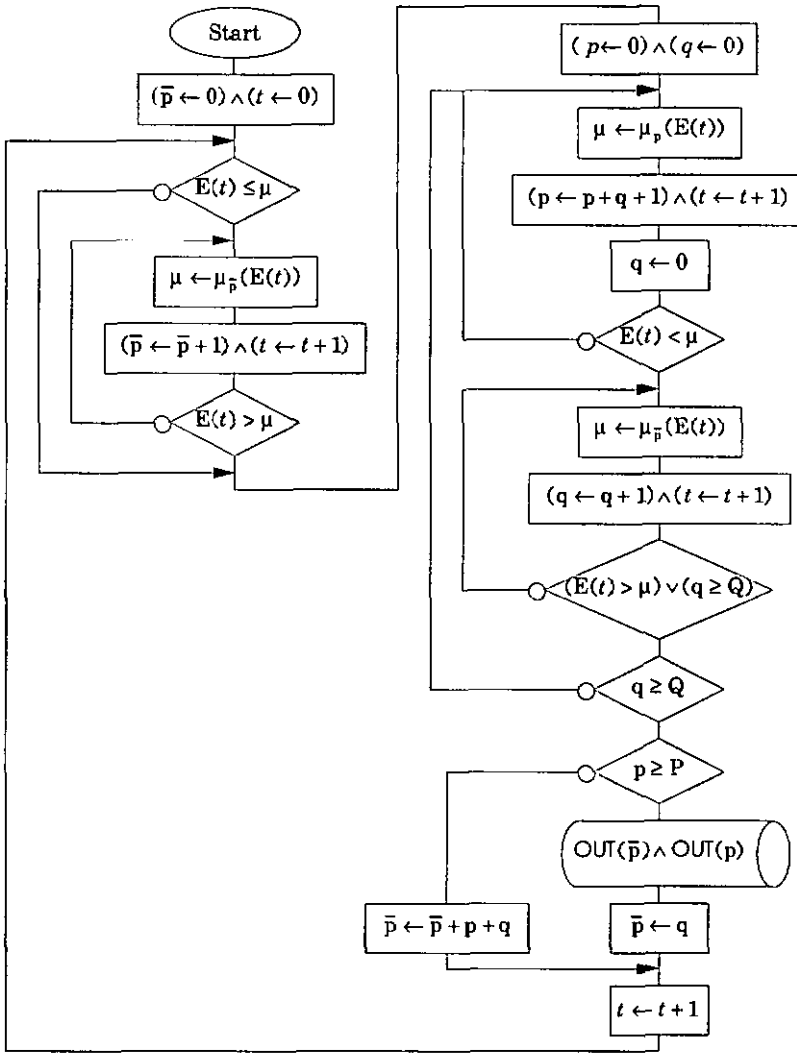

Figure 3.8.a Algorithme de segmentation silence et parole

La figure 3.8.a présente l'algorithme de segmentation dans ses grandes lignes (les conditions de fin de traitement ont été laissées de côté, ainsi que quelques menues subtilités). Dans cette figure, $\bar{p}$ accumule le nombre de fenêtres étiquetées comme silence. La variable p s'occupe des segments de parole; la variable

q dénombre les étiquettes provisoirement attribuées au silence. Le seuil actif μ, dont l'adaptation varie en fonction de l'énergie E, prend une valeur initiale estimée en considérant a priori que les premiers instants (0.05 s) d'un signal sont toujours du silence, l'énergie moyenne sur cette durée servant d'estimation. L'adaptation est réalisée par une technique de moyenne glissante, avec pour fenêtre une exponentielle qui décroît vers le passé. L'équation de mise à jour est

3.8.4 $$\mu(t+\Delta t) = (1-\lambda\cdot\Delta t)\cdot\mu(t)+\zeta\cdot\lambda\cdot\Delta t\cdot E(t) \qquad \forall t$$

Les valeurs du coefficient d'adaptation λ valent 5 s^{-1} pour le silence et 0.7 s^{-1} pour la parole. Les valeurs du coefficient de marge ζ valent 1.2 pour le silence et 1.3 pour la parole. Les durées minimales de silence Q et de parole P valent respectivement 0.05 s et 0.1 s.

La nécessité de cette segmentation entre silence et parole se justifie par le fait que la fréquence fondamentale n'est définie que pendant un signal de parole laryngée. Or, la simple existence d'une périodicité n'est pas un critère suffisant pour décider si la fenêtre en cours d'analyse est de la parole ou non, cette périodicité pouvant avoir d'autres causes que la voix du locuteur, par exemple le ronronnement d'un ventilateur. L'algorithme présenté, approximativement inspiré de [85MokA], permet de rejeter un grand nombre de cas où la voix du locuteur n'est pas à l'origine du signal analysé, en ne conservant que les fenêtres d'énergie suffisante.

3.8.4 Filtrage passe-bas

Nous avons préaccentué le signal de sorte à ce que le signal soit étiqueté correctement en silence ou parole. Malheureusement, l'accroissement des hautes fréquences est néfaste au bon fonctionnement du détecteur de périodicités. En outre, la gamme attendue des valeurs de fréquence fondamentale est restreinte par rapport à la gamme des fréquences disponibles dans le signal. Nous allons donc traiter ce dernier par un filtre passe-bas, de sorte à améliorer les conditions de travail de l'étape suivante.

La figure 3.8.b présente la structure en treillis du filtre utilisé. Un filtre elliptique du septième ordre a été retenu; on y dénombre 22 additionneurs et 8 coefficients multiplicatifs, dont les valeurs répertoriées à la figure 3.8.c génèrent un filtre passe-bas de fréquence de coupure égale à 1.0 kHz. Sa réponse spectrale d'amplitude est donnée à la figure 3.8.d.

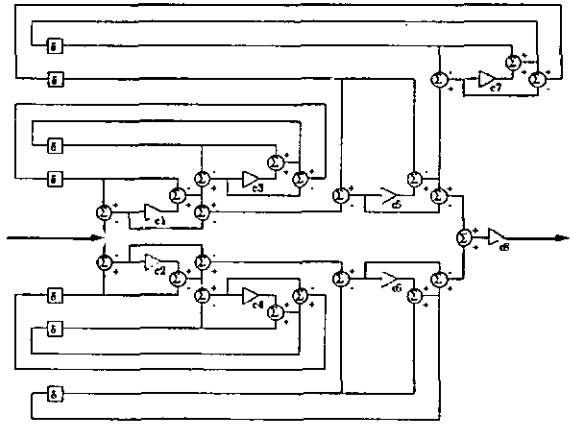

Figure 3.8.b Structure du filtre passe-bas

$c_1 = 0.42424830$	$c_2 = 0.23363519$
$c_3 = 0.12289445$	$c_4 = 0.21064269$
$c_5 = 0.07120495$	$c_6 = 0.30987329$
$c_7 = 0.25589200$	$c_8 = 0.50000000$

Figure 3.8.c Valeur des coefficients du filtre passe-bas

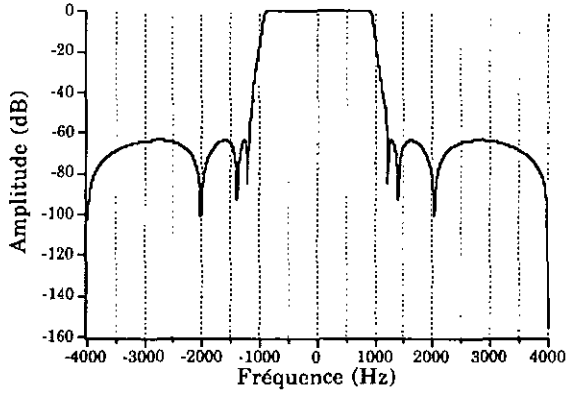

Figure 3.8.d Réponse spectrale du filtre passe-bas

■ 3.8.5 Extraction de la fréquence fondamentale

Nous sommes maintenant en mesure d'estimer une valeur de fréquence fondamentale pour chaque fenêtre étiquetée comme parole. Pour ce faire, suivons, du moins partiellement, la recette donnée par [83CoxR]. Commençons par soustraire au signal filtré passe-bas sa valeur moyenne puis recherchons le plus grand pic (en valeur absolue) du premier tiers du signal [0,N/3[, ainsi que ceux du deuxième [N/3,2·N/3[et du troisième tiers [2·N/3,N[. Un fraction du plus petit de ces trois pics (±0.3 minimum maximorum) servira à définir une bande d'intolérance. Les valeurs du signal à l'intérieur de cette bande seront annulées; les valeurs du signal qui excèdent les seuils positifs ou négatifs seront réduites de la valeur du seuil. Ainsi, seules subsisteront les parties proéminentes du signal (signal réduit).

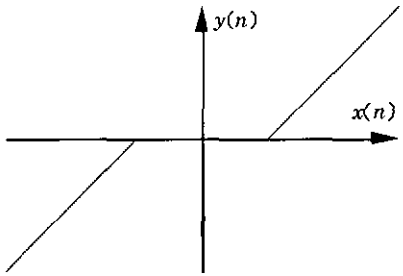

Figure 3.8.e Fonction de transfert du signal réduit

La figure 3.8.e montre la fonction de transfert entre le signal de parole filtré passe-bas $x(n)$ et le signal réduit $y(n)$. Nous calculons ensuite les coefficients d'autocorrélation à court terme du signal filtré passe-bas réduit, en concentrant nos efforts sur les seuls coefficients susceptibles de nous intéresser, soit ceux correspondant à une valeur de fréquence fondamentale comprise entre 64 Hz et 320 Hz

3.8.5 $$R(i) = \sum_{n=0}^{N-1-i} y(n) \cdot y(n+i) \qquad \forall i \in \{0\} \cup [I_{min}, I_{max}]$$

Nous devons encore normaliser ces coefficients en les divisant par R(0)

3.8.6 $$r(i) = \frac{R(i)}{R(0)} \qquad \forall i \in \{0\} \cup [I_{min}, I_{max}]$$

Nous devons ensuite les ajuster de sorte à pénaliser les phénomènes de doublement de la période fondamentale. Nous y parvenons en multipliant chaque coefficient normé $r(i)$ par un facteur correctif qui reste unitaire jusqu'à une période limite correspondant à 200 Hz, puis qui décroît de façon linéaire avec une pente valant 2.35 s^{-1}

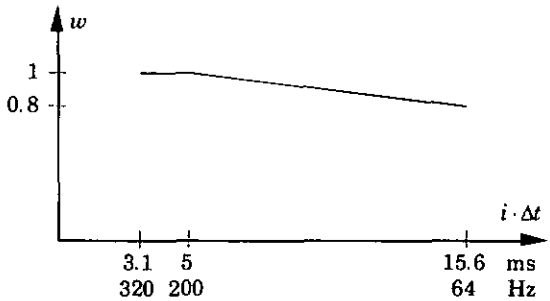

Figure 3.8.f Pénalisation du doublement de la période fondamentale

Enfin, nous pouvons associer la période fondamentale F_0 de la fenêtre analysée avec l'ordre du coefficient de plus grande valeur observée r

3.8.7 $$r = \underset{i=I_{min}}{\overset{I_{max}}{Mox}}\, r(i) \qquad \forall i \in [I_{min}, I_{max}]$$

Si Δt représente l'intervalle entre deux échantillons, alors

3.8.8 $$F_0 = \frac{1}{\Delta t \cdot \underset{i=I_{min}}{\overset{I_{max}}{ArgMox}}\, r(i)} \qquad \forall i \in [I_{min}, I_{max}]$$

3.8.6 Décision d'émission laryngée

Nous ne sommes pas encore au bout de nos peines. En effet, bien que disposant de fenêtres d'énergie suffisante et munies d'une estimation optimale de fréquence fondamentale, rien pourtant ne nous assure que les fenêtres en question recouvrent réellement un son laryngé, seul cas où la fréquence fondamentale est une description convenable de l'excitation. Nous devons donc encore segmenter la parole sur ce critère [86CamJ]. Nous avons choisi deux conditions qui doivent être réalisées simultanément pour que la décision d'émission laryngée soit prise; premièrement, nous imposons que le coefficient

de corrélation de plus grande valeur observée r possède une valeur minimale absolue. Secondement, nous imposons que cette valeur soit aussi minimale relative, selon le critère de l'équation 3.8.9

3.8.9 $$(r > 0.5) \wedge \left(r > \gamma \cdot \left(\frac{1}{F_0} - \frac{1}{F_{max}} \right) + r_0 \right)$$

Ce second critère semble favoriser le doublement de la période fondamentale; cependant les valeurs des paramètres sont telles que son effet est plutôt de pénaliser les périodes trop courtes. Dans l'équation 3.8.9, F_{max} est la fréquence maximale considérée (320 Hz), $r_0 = 0.8$ est l'abscisse du seuil pour cette ordonnée, et $\gamma = -67.37\ s^{-1}$ est la pente; il s'ensuit que les critères absolus et relatifs se rejoignent à la fréquence fondamentale de 132 Hz.

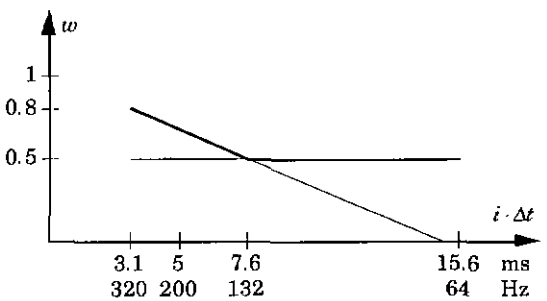

Figure 3.8.g Seuil d'acceptation pour un son laryngé

Nous avons sauté l'étape qui consisterait à recalculer le coefficient de corrélation de plus grande valeur observée r en fonction du signal original non filtré et non réduit. Si le son est réellement laryngé, il contiendra peu de hautes fréquences et la différence ne sera pas significative; sinon, cette étape supplémentaire permettrait peut-être de rejeter quelques cas d'erreur.

3.9 Fréquence fondamentale moyenne

Si T représente l'ensemble des fenêtres considérées comme support d'un son laryngé au sens des conditions exposées aux paragraphes 3.8.3 et 3.8.6, alors la fréquence fondamentale moyenne $\langle F_0 \rangle$ est donnée par

3.9.1

$$\langle F_0 \rangle = \frac{\mathrm{Card}(T)}{\sum_{t \in T} \frac{1}{F_0(t)}}$$

L'équation 3.9.1 paraît tellement triviale qu'il pourrait sembler inutile d'y apporter un quelconque commentaire. Toutefois, il vaut la peine de remarquer que ce sont les périodes dont nous avons décidé de calculer la moyenne, et non les fréquences. Enfin, il faut souligner qu'une étape potentiellement utile a été omise: l'application d'un filtre médian aux périodes fondamentales aurait été bienvenue, mais n'a pas été réalisée.

4 État de l'art

Nous n'avons pas la prétention d'offrir ici un condensé de toutes les recherches et comptes-rendus qui pourraient avoir eu un rapport avec le sujet de cette thèse. Au contraire, nous avons choisi de présenter en toute partialité un nombre restreint de sujets et d'articles. Nous croyons cependant que notre choix, bien que réduit et arbitraire, reflète fidèlement l'état de l'art en reconnaissance de locuteurs indépendante du texte. Si le besoin de situer ce travail dans un cadre plus large que nous ne le faisons devient impérieux, alors nous proposons de consulter des documents tels que [76AtaB, 76RosA, 78JesP, 81CorP, 85DodG, 86FurS, 90BoîR, 90FurS].

4.1 Traits d'identification

L'identification d'une personne par un tiers peut se faire de cinq manières différentes [79WarG]. Suivant les cas on n'identifiera pas la personne elle-même, mais on s'assurera simplement qu'elle a droit à certains privilèges, ou qu'elle est assujettie à certaines corvées. Enfin, il est clair que les cinq manières que nous allons présenter brièvement peuvent se combiner librement entre elles.

Possession: l'individu à reconnaître possède un objet particulier tel qu'un insigne, une clef ou une carte par exemple. L'objet, bien que physiquement transmissible, attribue une identité à son détenteur.

Connaissance: l'individu à reconnaître connaît un code particulier qui lui permet de s'identifier; suivant les cas on parlera de mot de passe, de numéro de client ou de code d'identification personnelle. Sa particularité tient au fait que sa description symbolique est aisée; ce type de code est donc transmissible.

État: l'individu à reconnaître présente des caractéristiques physiques propres à lui seul. C'est le mode de reconnaissance le plus courant dans la vie de tous les jours où il est en particulier très lié aux différences de faciès.

Action: l'individu à reconnaître est apte à se livrer à une action qui le distingue d'autrui. C'est le mode de reconnaissance le plus courant dans toutes les applications légales ou commerciales où il est très lié à la comparaison de signatures.

Reconnaissance: l'individu à reconnaître doit démontrer son aptitude à soi-même reconnaître un élément. Ce mode se distingue du mode de connaissance d'un code par la difficulté inhérente à une description symbolique de l'élément en question. Il est par exemple possible de mémoriser une nuance de couleur, une odeur ou l'apparence détaillée d'un paysage familier sans pour autant être capable d'en transmettre une description. En résumé, l'individu à reconnaître se distingue par une information qu'il recèle, sans qu'il lui soit possible de transmettre l'information elle-même. Par contre, il peut agir en fonction de cette information.

Comme nous l'avions souligné au paragraphe 1.1.1, nous tentons de nous satisfaire de la voix seule pour caractériser de façon automatique un individu, quand bien même nous venons de citer d'autres approches. Afin de pouvoir comparer la pertinence de la voix dans un processus de reconnaissance avec celle d'autres méthodes, nous allons très brièvement donner deux exemples de techniques qui ne sont pas basées sur son traitement automatique mais qui toutes font partie du même mode de reconnaissance: l'état. Nous présenterons d'abord le cas des empreintes génétiques puis celui des empreintes digitales.

4.1.1 Empreintes génétiques

La technique des empreintes génétiques repose sur l'unicité de la constitution de l'agent de réplication d'une cellule, assemblage d'éléments fondamentaux nommés nucléotides dont l'arrangement en double hélice forme une substance acide désoxyribonucléique. Chaque nucléotide porte une parmi deux bases puriques (adénine, guanine) ou deux bases pyrimidiques (cytosine, thymine), dont l'ordre d'agencement peut être considéré en partie comme caractéristique d'un individu, et plus encore d'une espèce; dans le cas d'individus on parle volontiers de gènes allélomorphes, ou allèles, pour décrire les variations au sein d'une population donnée. Globalement, le nombre d'allèles est faible; il représente moins de 0.5% des $3 \cdot 10^9$ bases du génome humain. Cependant, certaines techniques permettent d'isoler quelques fragments de gènes où les allèles sont en concentration relativement élevée. Deux échantillons peuvent ainsi être comparés dans le but de déterminer s'ils proviennent du même individu ou non.

La nature particulièrement invasive de cette méthode de reconnaissance, sa complexité et sa lenteur la rendent inadaptée à des tâches routinières. Son usage est donc réservé à des cas spéciaux, notamment dans le cadre d'enquêtes judiciaires. Si elle est considérée comme efficace pour attester une filiation, par exemple, elle est au contraire mal acceptée en tant que preuve d'identité. Les estimations du taux des erreurs potentiellement commises par cette méthode varient de plusieurs ordres de grandeur: de $1 \cdot 10^{-8}$ à $4 \cdot 10^{-2}$ selon les experts consultés! La controverse porte principalement sur les effets dus à une éventuelle altération de l'échantillon à reconnaître entre le temps où il a été produit et celui où il a été recueilli puis analysé [90NeuP]. Par contre, aucun expert ne conteste le caractère parfaitement individuel de la distribution des allèles, à l'exception évidente de celui de jumeaux monozygotes.

4.1.2 Empreintes digitales

L'identification de personnes au moyen d'empreintes digitales fut proposée dès 1880, en substitution du système anthropométrique introduit par Bertillon qui tenait compte de onze mesures physiques telles que largeur de la tête, taille, longueur des doigts, etc. Les critiques de cet ancien système étaient fondées, puisque par exemple en 1903 on s'avisa que deux malfrats n'étaient pas seulement homonymes (Will West), mais encore se caractérisaient par des mesures pratiquement identiques dans le système Bertillon. La comparaison de leurs empreintes digitales ne révéla toutefois aucune similarité.

Une propriété des empreintes digitales est leur pérennité, à un facteur d'échelle près qui se manifeste surtout lors de la croissance de l'individu qui les porte. Une autre propriété est la grande variété d'empreintes rencontrées dans le monde; on considère généralement la confusion entre deux personnes comme une impossibilité pratique. L'acquisition d'empreintes est un procédé instantané et non invasif, puisqu'elle est réalisable de manière optique. Enfin, l'extraction de certaines caractéristiques descriptives d'une empreinte permet de réduire la capacité de stockage nécessaire et facilite leur comparaison automatique. Malgré ces qualités, l'utilisation courante d'empreintes digitales ne connaît pas de développement commercial important; elle est par exemple moins populaire que celle des motifs rétiniens [89FitK]. Nous lierons cet insuccès à la crainte d'une intrusion de nature policière dans la vie privée de l'utilisateur, en raison de la similarité des techniques d'investigation criminelle et de celles adaptées au domaine public.

Figure 4.1.a Empreinte digitale agrandie du pouce droit de l'auteur

4.2 Reconnaissance humaine

La reconnaissance par un être humain de la voix d'un autre procède de nombreux aspects. Pour s'en convaincre, il suffit de ne considérer que quelques-uns des indices utilisés par un auditeur pour reconnaître un locuteur, essayer d'imaginer une expérience adaptée à la mise en évidence de caractéristiques précises pour chacun d'eux, puis tenir compte de toutes les synergies possibles. Par exemple, on considérera la langue utilisée, l'accent régional, le contenu sémantique, la forme émotionnelle, la rapidité d'élocution, la modulation de la voix, sa puissance, son timbre, etc. Certains de ces indices appartiennent au domaine de la psychologie, d'autres à celui de la psychophysique, d'autres à la sensorimétrie, tandis que d'autres encore sont purement acoustiques. De plus, les conclusions d'un auditeur sur l'identité du locuteur portent sur de multiples aspects [89GiaA]. Il s'ensuit que vouloir présenter ici l'état de l'art dans chacun de ces domaines serait trop ambitieux; nous laissons ce soin aux auteurs des documents proposés en tête de ce chapitre et nous nous contenterons de quatre exemples. Les trois premiers diffèrent entre eux par le type de tâche réalisée mais possèdent en commun le fait que la reconnaissance est basée sur des sons; le quatrième exemple fait aussi appel à une prise de décision humaine, cependant la caractéristique sur laquelle est basée la reconnaissance n'est plus sonore mais visuelle.

4.2.1 Tâche d'identification sans rejet

Le premier exemple [63ComA] traite de une tâche d'identification 1 à n où les conditions expérimentales sont bien définies. 15 auditeurs ont tenté d'identifier 9 locuteurs mâles dont ils étaient familiers, cette familiarité ayant permis de limiter les effets d'une mémorisation imparfaite de la voix de chacun. Le matériel acoustique utilisé était constitué de 15 répétitions, par locuteur, de la seule voyelle [i] tenue pendant une durée adéquate, ce qui a permis d'annihiler tout effet lié à des variations du contenu et de la forme du message.

Notons que, dans une tâche d'identification 1 à $n = 9$, la chance seule explique déjà un taux de réussite de 11%. Il est donc nécessaire d'établir un seuil minimal pour le taux obtenu par l'expérience si l'on veut pouvoir affirmer avec une confiance donnée que la capacité observée d'identification des locuteurs par les auditeurs n'est pas due à un hasard favorable. Si l'on veut être certain à 95% de la réalité de l'identification, ce seuil vaut 28%. Or, les taux de réussite mesurés sont dépassés pour toutes les durées testées, qui varient de 0.025 s à 1.5 s; le taux de réussite le plus élevé valant 65% et étant atteint pour une durée de 0.75 s, il résulte de cette étude que les auditeurs ont été capables d'identifier les locuteurs avec une marge de 37%.

4.2.2 Tâche d'identification avec rejet

Le deuxième exemple [91Kre]] utilise un matériel sonore constitué de conversations téléphoniques et traite à la fois d'identification 1 à $n+1$ et de vérification. Dans le cas de l'identification avec rejet, 100 auditeurs ont eu chacun environ 100 s pour se familiariser avec une unique voix mâle inconnue ($n = 1$). Leur capacité à identifier cette voix a été testée après un délai d'une semaine, l'examinateur leur présentant alors 10 voix de locuteurs du même sexe tout en précisant que la voix à identifier pouvait tout aussi bien s'y trouver plusieurs fois ou ne pas s'y trouver du tout (en réalité, elle s'y trouvait une fois exactement). Comme aucune prétention d'identité n'intervient, il s'agit nécessairement d'une tâche d'identification, même si l'apprentissage réalisé par les auditeurs ne porte que sur un seul locuteur. Leur tâche consistait en outre à préciser la confiance qu'ils ressentaient envers chacune de leurs réponses, une autre équipe de 50 auditeurs ayant au préalable émis un jugement sur la difficulté supposée à reconnaître chacun des locuteurs. Il résulte de cette première partie de l'étude que

1) le taux de confusion est nécessairement nul puisque la base de donnée ne comprend qu'un seul locuteur ($\rho_c = 0$). Le taux de méprise ρ_m est faible et vaut 6.1%; il peut éventuellement être interprété ici comme équivalent à un taux de fausse acceptation ρ_a. Le taux de dédain ρ_d est plus élevé et vaut 37%; il peut éventuellement être interprété ici comme un taux de faux rejet ρ_r. Cette différence entre les deux taux peut être liée au mode de reconnaissance et de mémorisation de locuteurs par les auditeurs, mais il peut aussi par exemple avoir été crée involontairement par l'examinateur dans sa présentation aux auditeurs de leur tâche, à supposer qu'il ait insisté trop lourdement sur la possibilité ou l'impossibilité d'apparition du locuteur à identifier parmi les voix présentées. Ce point montre la difficulté inhérente à toute expérience où le jugement humain intervient;

2) les voix estimées comme les plus difficiles à reconnaître par les 50 auditeurs indépendants se sont avérées être celles qui sont source du nombre minimal d'erreurs! Ce fait est interprété par les auteurs comme la manifestation des termes en lesquels les auditeurs mémorisent les voix: un prototype et un ensemble de déviations par rapport à celui-ci. Au cours du temps, les déviations seraient oubliées de telle sorte que les réponses d'identification convergeraient vers les voix les plus typiques. Or, une voix qui se démarque peu du prototype sera considérée comme difficile à reconnaître, aucune caractéristique saillante ne pouvant lui être associée.

4.2.3 Tâche de vérification

Dans le cas de **v** la tâche de vérification, les mêmes 10 locuteurs que dans l'expérience précédente ont servi de matériel de test; ils étaient inconnus des 24 nouveaux auditeurs, tous différents des 150 déjà rencontrés. Le travail des auditeurs consistait à déterminer si une paire de locutions d'une durée de 2 s chacune et séparées par 1.5 s provenaient du même locuteur ou non; il s'ensuit qu'il ne s'agit pas à proprement parler d'un tâche de vérification mais plutôt d'une tâche de comparaison, car aucune prétention d'identité explicite n'intervient. Par analogie, nous continuerons malgré tout à parler ici de vérification. Le nombre de paires de voix présentées valait $30 = 10 \times 3$ pour les tests qui auraient dû conduire à une décision d'homogénéité, et $90 = 10 \times 9$ pour les tests qui auraient dû conduire à une décision d'hétérogénéité. Il résulte de cette seconde partie de l'étude que

1) le taux de fausse acceptation ρ_a vaut 11.4% tandis que le taux de faux rejet ρ_r vaut 19.7%;

2) les voix estimées comme les plus difficiles à reconnaître par les 50 auditeurs indépendants n'ont pas eu d'effet sur le taux de fausse acceptation. Le taux de faux rejet y a été lui aussi insensible, à condition qu'elles aient été présentées en seconde position dans la paire. En revanche, les auteurs ont pu montrer une corrélation positive entre la difficulté présumée de reconnaissance et le taux de faux rejet lorsque les voix difficiles étaient présentées en premier.

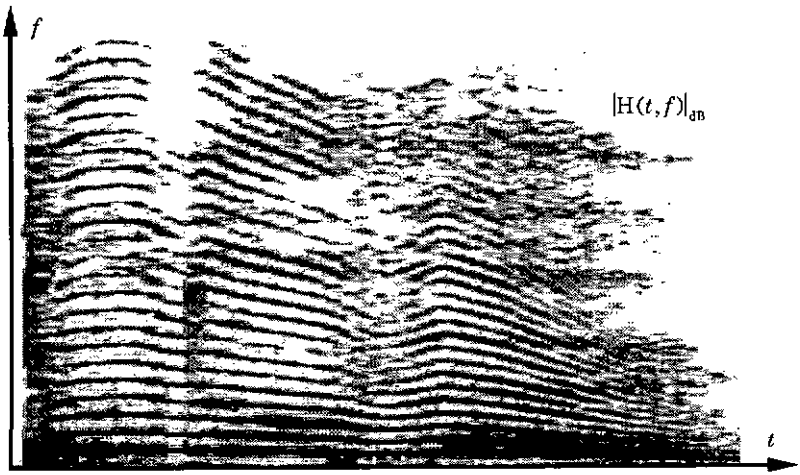

Figure 4.2.a Sonagramme

4.2.4 Sonagrammes

La méthode des sonagrammes est historiquement une des premières à avoir fait intervenir un dispositif technique autre qu'un simple appareil d'enregistrement ou de restitution de sons. La propriété de ce dispositif, un sonagraphe, destiné à faciliter la reconnaissance humaine de locuteurs, est de représenter le message vocal sous une forme visuelle telle que son déroulement temporel soit accessible intégralement et simultanément par l'humain en charge de reconnaître le locuteur, sans pour autant nuire à la représentation de ce message. On y parvient en construisant un espace où l'une des dimensions est graduée en unités de temps et l'autre en unités de fréquence; la représentation dans cet espace de la série des spectres d'amplitudes à court terme contenus dans le signal

acoustique de parole est appelée sonagramme, constituée finalement d'une image où la densité spectrale est codée par une intensité lumineuse.

Certains experts se sont spécialisés dans la reconnaissance de locuteurs fondée sur cette représentation de la parole. Il semble cependant que l'efficacité d'une telle approche soit fortement controversée; certains, par exemple [62Kerl.], prétendent à une qualité excellente de reconnaissance ($\rho_e = 1\%$) tandis que d'autres, par exemple [68SteK], font preuve de plus de circonspection en montrant que la qualité de la reconnaissance auditive reste meilleure que celle de la reconnaissance visuelle, même si l'on choisit pour cette dernière un expert compétent et pour la première un auditeur médiocre. D'autres auteurs encore [88Wil]] proposent de modifier les détails de la transformation utilisée, sans même oser aborder l'aspect de l'efficacité de leur méthode.

4.2.5 Commentaire

Comme nous l'avons vu, il est très difficile d'interpréter les résultats d'expériences telles que celles que nous venons de décrire. Soit les conditions expérimentales sont simplifiées à l'extrême (production vocale élémentaire, identification 1 à n) et les résultats sont interprétables de façon claire mais reflètent trop peu des conditions réelles, soit les contraintes du protocole expérimental sont assouplies au risque d'introduire des facteurs incontrôlables tels que par exemple l'effet de l'expérimentateur en sus de celui des locuteurs et des auditeurs, comme en 4.2.2.1, ou encore faire intervenir des détails que l'on pourrait au premier abord croire insignifiants tels que l'ordre de présentation des stimuli, comme en 4.2.3.2.

		Identification 1 à $n = 9$	Identification 1 à $n+1 = 1+1$	Vérification
		I*	I	V
Confusion	ρ_c	35.0%	00.0%	
Méprise	ρ_m		06.1%	
Dédain	ρ_d		37.0%	
Fausse acceptation	ρ_a			11.4%
Faux rejet	ρ_r			19.7%

Figure 4.2.b Récapitulation des taux d'erreur en reconnaissance humaine

La figure 4.2.b récapitule les taux d'erreur rencontrés ci-dessus. Nous rappelons que la comparaison de ces taux n'est licite que dans la mesure où l'on garde à l'esprit les conditions très différentes dans lesquelles ces expériences ont été menées.

4.3 Reconnaissance automatique

L'automatisation de la reconnaissance de locuteurs, par définition, permet à cette dernière de rendre des jugements objectifs [88NakH]. Cette objectivité n'implique pas nécessairement une qualité supérieure de reconnaissance [89FedA], même si nous pouvons espérer obtenir qu'une approche automatique se montre au moins aussi efficace qu'une approche humaine; par contre, elle permet de focaliser notre intérêt sur les méthodes elles-mêmes et permet d'ignorer toutes les conditions annexes aussi diverses que l'état de fatigue d'un auditeur ou sa bonne foi.

Cet approfondissement du domaine exige malheureusement parfois le renoncement à certains des éléments utilisés par l'humain parce que l'état de l'art ne permet pas de les traduire en termes techniques; nous pensons par exemple aux apports émotionnels et sémantiques qui parfois aussi révèlent l'identité du locuteur. Cette perte qualitative pourra toutefois être compensée par l'augmentation de précision quantitative qu'apporte une approche utilisant un calculateur numérique. En outre, du point de vue économique, le monde occidental est accoutumé à considérer le remplacement d'un homme par une machine comme profitable, et l'automatisation d'une tâche se traduit souvent par un gain financier.

En l'état actuel, l'application la plus développée de la reconnaissance automatique de locuteur est le contrôle d'accès, où le dispositif de reconnaissance sert généralement de cerbère à un lieu particulier, le mot de passe introduit au paragraphe 1.3.3 étant le degré de dépendance du texte privilégié par ce mode de fonctionnement [88AttJ, 89FitK]. L'autre extrême de l'échelle des degrés de dépendance est occupé par l'indépendance totale qui concerne principalement les techniques d'investigation judiciaire, un domaine controversé dont nous ne tenons pas à débattre ici. Enfin, les méthodes de reconnaissance développées dans cette thèse se rapportent essentiellement à la reconnaissance indépendante du texte; nous verrons plus tard quelles en sont les applications potentielles.

Comme déjà mentionné au paragraphe 1.3.4, l'indépendance du texte s'obtient souvent en s'intéressant à la densité de probabilité d'une suite de vecteurs caractéristiques à court terme, extraits d'une locution, dont l'estimation se fait sur un temps suffisamment long pour pouvoir modéliser le comportement global du locuteur; remarquons au passage que les vecteurs caractéristiques utilisés à ce jour sont presque toujours une représentation particulière du filtre de synthèse de l'analyse par prédiction linéaire. Cette statistique à long terme est ensuite codée de façon adéquate; la représentation obtenue constitue la référence du locuteur. Nous allons donner ci-dessous deux exemples de codage de la densité de probabilité: le premier en utilise une représentation très rudimentaire puisqu'il ne s'intéresse qu'à sa moyenne; on trouve ce type d'approche par exemple dans [77Mar], 81ShrM]. Le second en conserve un représentation discrète mais y est plus fidèle, par le biais d'une technique nommée quantification vectorielle; on trouve ce type de codage par exemple dans [80BuzA].

4.3.1 Moyenne à long terme

Le premier exemple que nous avons choisi de présenter ici est extrait de l'article [82ShrM]. Dans cette expérience, 12 locuteurs mâles ont prêté leur voix à raison de 6 sessions d'une durée de 10 min chacune. Par session, une durée de 60 s a été échantillonnée à une cadence de 10^4 Hz puis numérisée et analysée par tranches contiguës de 0.020 s, sans recouvrement; les tranches correspondant à du silence ont été rejetées, ce qui toutefois n'a pas empêché de conserver au moins 20 s par session. Chaque tranche retenue a livré divers vecteurs caractéristiques liés à une analyse par prédiction linéaire d'ordre $p = 12$: il s'agit en particulier des coefficients de prédiction linéaire a_k, des coefficients de réflexion K_m, des coefficients logarithmiques du rapport des aires g_m et des coefficients cepstraux $c(n)$. Tous ces ensembles de coefficients, sans exception, sont une représentation du filtre de synthèse.

Chaque vecteur caractéristique a ensuite été soumis à une transformation de Karhunen-Loève dépendante du locuteur, estimée sur une durée de 40 s. Enfin, une durée de 100 s, obtenue en faisant usage de 5 des 6 sessions, a permis d'estimer la valeur moyenne des paramètres transformés. La session non utilisée a fourni 4 vecteurs moyens de test, estimés sur une durée de 5 s chacun, puis chaque session a servi de source de vecteurs de test à tour de rôle, tandis que les autres prenaient (ou conservaient) le rôle de la référence.

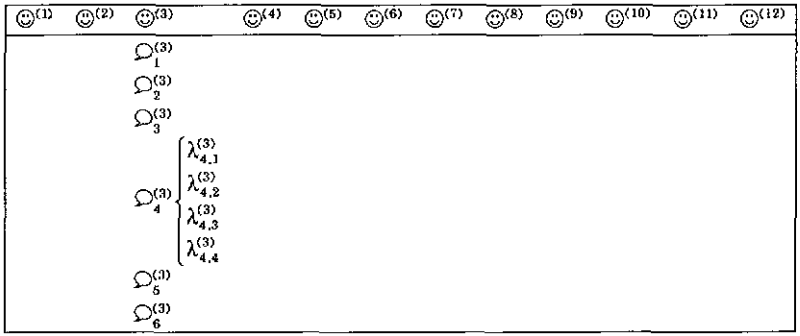

Figure 4.3.a Douze locuteurs, six locutions et quatre vecteurs de test

La figure 4.3.a illustre le protocole par un exemple où le troisième locuteur voit toutes ses locutions être utilisées pour la construction d'une référence moyenne, à l'exception de la quatrième, découpée en quatre vecteurs de tests. Dans un cas de fonctionnement idéal, une tâche de vérification comparant un de ces vecteurs avec la référence du locuteur considéré produirait une décision d'homogénéité. Par permutation des sessions de test, le nombre maximal de décisions correctes d'homogénéité vaut donc 24 pour ce locuteur; c'est aussi le nombre maximal possible de tentatives d'identification 1 à n. Dans un cas de fonctionnement idéal, une tâche de vérification comparant un de ces vecteurs à la référence d'un locuteur différent produirait une décision d'hétérogénéité. Le nombre maximal de décisions correctes d'hétérogénéité vaut donc $6\times(12-1)\times(6\times4)=1584$ pour ce locuteur; parmi celles-ci, seules $1584/6=264$ ont été réalisées par les auteurs.

La figure 4.3.b montre les résultats obtenus selon le protocole d'expérience décrit. On constate que les seuils de vérification ont été imposés a posteriori de sorte à rendre égaux les taux de fausse acceptation et de faux rejet. Malheureusement, les auteurs ont omis de spécifier si chaque référence possède son propre seuil, ou si un seuil unique est associé à chaque locuteur, ou encore si une seule valeur globale de seuil a été utilisée.

	∎ 1 à $n = 12$	▼
	ρ_c	$\rho_a = \rho_r$
$\langle a_k \rangle$	11.5%	2.9%
$\langle c(n) \rangle$	08.3%	2.9%
$\langle K_m \rangle$	07.8%	2.7%
$\langle g_m \rangle$	06.8%	3.0%

Figure 4.3.b Taux d'erreur selon [82ShrM]

■ 4.3.2 Erreur moyenne de quantification vectorielle

Le second exemple que nous avons choisi de présenter ici est extrait de l'article [88SooF], où l'auteur examine l'efficacité d'une représentation discrète de la densité de probabilité d'un vecteur caractéristique obtenue par un algorithme de classification, par exemple celui des nuées dynamiques. Si l'on associe à chaque noyau un indice différent, alors il devient possible de classer puis de quantifier tout point de l'espace dans lequel sont plongés les vecteurs caractéristiques; on parle de la quantification vectorielle du paragraphe 2.3.4.A. Il suffit ensuite de considérer une mesure de la distorsion entre les vecteurs originaux et les vecteurs quantifiés, comme au paragraphe 2.3.4.B; on peut alors comparer un ensemble de vecteurs de test avec la référence du locuteur constituée par l'ensemble des noyaux, appelé dictionnaire. Le paragraphe cité présentait cette technique de comparaison sans chercher à la justifier; l'objet de ce paragraphe est de combler cette lacune.

A) Principe de la reconnaissance

Dans le but d'obtenir une distance entre une locution de référence et une locution de test, la technique d'estimation de la densité de probabilité du vecteur caractéristique de test a été simplifiée de la manière suivante par les auteurs de l'article: Soit rendue disponible l'estimation de référence. Plutôt que de se livrer à une estimation séparée sur la locution de test et de comparer les deux estimations dans l'espace des représentations statistiques, il est préférable de considérer l'adéquation, à la référence, des vecteurs issus de la locution de test. Le travail s'effectue donc indirectement, dans l'espace des vecteurs caractéristiques, et non plus directement dans l'espace des estimations statistiques, comme c'était le cas au paragraphe 4.3.1.

Cette adéquation se mesure au fait que les vecteurs de test doivent tous être proches des noyaux de la référence. Pour s'en assurer, on réalise la quantifica-

tion vectorielle de la locution de test par la référence et l'on accumule toutes les distorsions observées [85SooF, 86RosA, 86SooF], avec éventuellement une pondération dépendante du noyau retenu [89MasJ, 90EatJ, 91MatT]. La description détaillée de cette méthode a été donnée au paragraphe 2.3.4 pour la mesure de distance et au paragraphe 2.2.2 pour la phase de classification.

Le lecteur attentif remarquera que cette méthode est moins bonne qu'une comparaison dans l'espace statistique, puisqu'une concentration ponctuelle de vecteurs de tests centrés par hasard sur un noyau de la référence produirait une distance nulle, même si la référence annonce une distribution bien plus étalée. Néanmoins, il est fréquent que les locutions de test soient de durée insuffisante pour permettre une estimation robuste de la densité de probabilité des vecteurs caractéristiques retenus; la mesure d'adéquation proposée en est généralement acceptée comme plus représentative.

De notre point de vue, le désavantage majeur de cette approche se fait surtout sentir dans une tâche d'identification, car il est des références qui ont pour propriété de produire des distances plus petites ou plus grandes que d'autres, sans en être pour autant moins ou plus discriminantes. Ce fait complique le calcul de la mesure de vraisemblance de l'homogénéité entre une locution de test et une référence, parce que l'approximation de cette vraisemblance par la simple erreur moyenne de quantification n'est que partiellement valide. Dans une tâche de vérification, un seuil sur cette erreur moyenne suffit à prendre une décision; dans une tâche d'identification, la comparaison de plusieurs d'entre elles devient nécessaire et ne devrait pas se faire directement, en principe.

B) Reconnaissance de locuteurs

Les vecteurs caractéristiques retenus par [88SooF] sont le cepstre à court terme du filtre de synthèse et sa pente temporelle. Pour les obtenir, le signal de parole est filtré passe-bande entre 200 Hz et $3.2 \cdot 10^3$ Hz. La cadence d'échantillonnage est de $6.67 \cdot 10^3$ Hz, tandis que la durée des fenêtres d'estimation est de 0.045 s, espacées régulièrement de 0.015 s. Une préaccentuation de coefficient $\mu = 0.95$ est appliquée au signal. L'estimation du cepstre provient d'une analyse par prédiction linéaire d'ordre $p = 8$, réalisée par la technique de l'autocorrélation. L'ordre cepstral pris en compte n'est pas mentionné par les auteurs, mais on peut supposer qu'il est égal à l'ordre d'analyse.

La pente cepstrale est calculée selon l'expression 3.5.1; en termes de reconnaissance de locuteurs, sa durée d'estimation optimale (au sens de la reconnaissance de locuteurs) varie, au dire des auteurs, entre 0.135 s et 0.195 s, ce qui correspond à respectivement 7 et 11 fenêtres pour le taux de recouvrement et la durée des fenêtres choisies. Il découle du choix de ces paramètres que l'estimation de la dérivée temporelle du cepstre n'est pas de bonne qualité parce qu'elle est insuffisamment locale, ou, ce qui revient au même, trop globale; c'est la raison pour laquelle les auteurs de l'article ont préféré la baptiser pente plutôt que dérivée.

Le nombre de locuteurs participant à cette expérience se compose d'autant d'hommes que de femmes, 5 pour chaque sexe. En cinq sessions réparties sur deux mois les locuteurs ont produit 20 occurrences de chacun des dix chiffres. Ces 200 locutions, recueillies par téléphone, ont été condensées par suppression des pauses. Par locuteur, 100 locutions ont été conservées pour servir de matériel de test E_t dans une tâche d'identification de 1 à n. La durée utilisée pour construire une référence au moyen de l'algorithme des nuées dynamiques correspond à celle des 100 autres locutions. Le nombre de classes est $K = 64$, autant pour les cepstres que pour leur pente; la mesure de distance utilisée pour la fonction d'affectation de la phase de classification, donnée à l'expression 2.2.8, est euclidienne. Par contre, les auteurs n'ont pas cru bon de préciser leur choix des noyaux initiaux ou leur condition de convergence, pas plus qu'ils n'ont précisé la nature exacte de la mesure de vraisemblance adoptée dans la comparaison des erreurs moyennes de quantification.

Les distorsions résultant de la quantification vectorielle ont été calculées de deux façons différentes. La première est le carré d'une métrique euclidienne appliquée à des cepstres tronqués de leur première composante c_0; la mesure de distance d de l'expression 2.3.9 est par conséquent équivalente à d_2^2 selon 2.3.2. La seconde façon en est proche, puisqu'il s'agit du carré d'une métrique euclidienne pondérée. Les coefficients de pondération w_m^2, indépendants du locuteur, sont donnés par l'inverse de la moyenne sur de nombreux locuteurs de la variance du coefficient cepstral d'ordre m. Dans l'expérience décrite, pas moins de 100 locuteurs ont été utilisés pour fournir cette estimation.

C) Résultats

Voici enfin les résultats obtenus par la méthode dite d'erreur moyenne de quantification vectorielle. Rappelons que les valeurs données correspondent à

des taux de confusion ρ_c puisque nous sommes dans une tâche d'identification sans rejet ▮ de 1 à $n = 10$.

La figure 4.3.c présente six résultats différents obtenus par l'usage exclusif de l'un ou de l'autre des vecteurs caractéristiques ou par une combinaison linéaire des distances qu'ils engendrent, et en ignorant ou en prenant en compte une pondération de la distance euclidienne. Dans cette figure, le symbole VQ_c représente l'usage de cepstres instantanés, tandis que le symbole $VQ_{\Delta c}$ concerne celui de cepstres différentiels. Les symboles w et ¬w annoncent respectivement la prise en compte ou non de la pondération dans le calcul de distance.

ρ_c	¬w	w
$VQ_{\Delta c}$	24.5%	17.4%
VQ_c	15.7%	11.7%
$VQ_{\Delta c} \wedge VQ_c$	10.9%	07.2%

Figure 4.3.c Taux de confusion en quantification vectorielle selon [88SooF]

■ 4.4 Commentaires

Nous venons de présenter quatre articles d'auteurs s'intéressant à la reconnaissance de locuteurs, les deux premiers traitant de la reconnaissance humaine et les deux derniers de la reconnaissance automatique. Nous avons rencontré cinq mesures différentes des imperfections de la reconnaissance: confusion ρ_c, mépris ρ_m ou dédain ρ_d, fausse acceptation ρ_a ou faux rejet ρ_r. Nous avons encore appris à connaître quatre bases de données disparates quant au contenu, à l'ampleur et aux conditions d'acquisition. C'est donc avec précautions qu'il faut considérer la figure suivante, qui résume les résultats présentés.

		[63ComA]	[91KreJ]	[82ShrM]	[88SooF]
		☜	☜	$\langle c(n) \rangle$	$VQ_c \wedge$ ☎
Confusion	ρ_c	35.0%	00.0%	8.3%	11.7%
Méprise	ρ_m		06.1%		
Dédain	ρ_d		37.0%		
Fausse acceptation	ρ_a		11.4%	2.9%	
Faux rejet	ρ_r		19.7%	2.9%	

Figure 4.4.a Résumé de résultats de la littérature

4.4.1 Bases de données

Dans la suite de cette thèse nous aurions volontiers utilisé une des bases de données des études publiées. Malheureusement, les contacts que nous avons tenté de nouer dans le but d'en obtenir une copie se sont révélés infructueux, soit que 8 parmi les 14 auteurs contactés n'aient pas répondu, soit que leur base de données fût inadaptée à nos besoins car incomplète ou trop petite. Nous présentons ci-dessous un résumé des réponses intéressantes.

- Y. Bennani, du laboratoire de recherche en informatique, université de Paris-Sud, ainsi que J. W. Fussel, de US Department of Defense, proposent tous deux l'utilisation de la base de données TIMIT. Celle-ci contient 10 locutions pour chacun des 104 locuteurs et des 52 locutrices la constituant. Ces locutions offrent l'avantage d'être étiquetées phonétiquement; cependant, la durée de chacune d'elles est trop courte pour nos besoins, quelques essais nous ayant convaincu que seule une vingtaine de secondes de parole est disponible par locuteur, ce qui est manifestement trop peu à la lumière de [79Mar]].
- S. Furui, de NTT Human Interfaces Laboratories, nous propose une base de données en langue japonaise, contenant 23 locuteurs et 13 locutrices, couvrant en 3 sessions une période de six mois. Notre inexpérience de la langue japonaise, en particulier notre incapacité à reconnaître l'état émotif du locuteur, nous ont conduit à renoncer à cette base de données puisque nous prétendons à l'indépendance du texte, par opposition à l'indépendance totale.
- A. E. Rosenberg, de AT&T Bell Laboratories, nous signale l'existence de deux bases de données sans accepter formellement de nous les mettre à disposition. Le contenu de ces bases de données est une série de coefficients d'autocorrélation, ce qui ne nous convient pas.
- M. Savic, de Renssealer School of Engineering, répond que la base de données qu'il utilise est propriété du gouvernement américain et ne peut nous être mise à disposition.
- J. J. Godfrey, de Linguistic Data Consortium, non contacté, vend depuis avril 1993 la base de données SWITCHBOARD. Saluons son gigantisme (250 heures de conversations téléphoniques); sur les 550 locuteurs qu'elle contient, 50 offrent un nombre de locutions suffisantes pour servir nos desseins. La date trop tardive à laquelle cette base de données est devenue disponible est cependant le motif majeur de notre renoncement à l'utiliser.

Cette difficulté à se procurer du matériel d'expérimentation ne paraît pas être que la nôtre, car contrairement au domaine de la reconnaissance de la parole

où certaines bases de données sont disponibles et reconnues comme étalons, celui de la reconnaissance de locuteurs ne paraît pas jouir de pareils avantages. De fait, nous ne connaissons pas d'exemple où des auteurs utiliseraient une base de données commune sans pour autant appartenir au même laboratoire.

Nous avons donc été contraint d'établir nos propres bases de données, que nous acceptons volontiers de mettre à disposition des chercheurs intéressés. La description de leurs conditions d'acquisition, de leur contenu et de leur exploitation fait l'objet du chapitre suivant.

5 Bases de données

Nous avons construit deux bases de données. Cette redondance apparente n'en est plus une quand on considère le fait que la première est destinée à servir de matériel d'expérimentation bien contrôlé, tandis que la seconde tolère une variabilité plus grande. Il en découle ainsi différents degrés de difficulté, la première de nos bases de données posant un problème de reconnaissance de locuteurs plus facile à résoudre que la seconde. Nous allons décrire ci-dessous le contenu de ces bases de données et énoncer notre façon de les exploiter.

5.1 Vocation d'une base de données

La vocation d'une base de données de reconnaissance de locuteurs est de satisfaire simultanément au moins deux usages: le premier est de permettre l'apprentissage, opération qui consiste en la construction d'un classificateur; le second est de permettre le test, opération qui consiste en l'estimation de la qualité du classificateur précédemment établi.

Par construction du classificateur nous entendons l'opération qui consiste à déterminer (explicitement ou implicitement) la frontière entre les classes qu'il distingue, sur la base de E_a un ensemble d'apprentissage formé d'échantillons représentatifs de la densité de probabilité de chaque classe. Cette définition est très générale. Par exemple, elle s'applique au cas du classificateur à deux classes $\{h\}$ et $\{\bar{h}\}$ rencontré au paragraphe 2.5; elle y recouvre à la fois la détermination du vecteur représentatif nécessaire à l'opération de comparaison, et l'établissement du seuil de décision.

Par estimation de la qualité du classificateur nous entendons l'opération qui consiste à lui faire classer E_t un ensemble de test formé d'échantillons représentatifs de la densité de probabilité de chaque classe qu'il distingue, et à dénombrer *les erreurs commises de sorte à pouvoir en déduire les taux d'erreur.* Tout l'art consiste à séparer l'ensemble $E = E_a \cup E_t$ des éléments de la base de

données en ensembles d'apprentissage E_a et de test E_t adaptés respectivement à l'un et à l'autre des deux emplois.

On considère généralement deux méthodes de séparation [72FukK]: la première, nommée méthode C, fait usage de tous les éléments pour la construction d'un unique classificateur et réutilise ces mêmes éléments pour le tester ($E_t \cap E_a \neq \varnothing$). L'estimation de la qualité $1-\rho$ de ce classificateur pêche par optimisme. La seconde méthode, nommée méthode U, réserve à l'usage exclusif de test un certain nombre d'éléments de la base de données et en utilise les autres pour la construction du classificateur ($E_t \cap E_a = \varnothing$). Cette méthode pêche par pessimisme.

La famille des méthodes U est populaire en raison même de son pessimisme [88LiK]. Un membre de cette famille est particulièrement apprécié; il s'agit de la technique de l'exclusion du test (en anglais: leave-one-out). Dans cette dernière, un seul élément est réservé à l'usage du test du classificateur ($\mathrm{Card}(E_t) = 1$), tandis que la construction de celui-ci est assurée par tous les éléments restants ($E_a = E \setminus E_t$). L'unique élément de test est classifié, le résultat de la justesse de classification contribue à l'estimation du taux d'erreur ρ, puis cet élément de test réintègre l'ensemble servant à la construction du prochain classificateur tandis qu'un autre vient prendre sa place. Ainsi, tous les éléments auront servi de test indépendant, mais il aura fallu construire autant de classificateurs; toutefois, l'estimation moyenne de qualité $\langle 1-\rho \rangle$ qui résulte de cette technique est considérée comme robuste.

5.1.1 Pessimisme et optimisme

Construisons un classificateur sur la base d'un ensemble d'échantillons d'apprentissage E_a; soit ρ la probabilité d'erreur de ce classificateur observée sur un ensemble d'échantillons de test E_t. Soit encore U l'ensemble de tous les échantillons possibles, le cardinal de cet ensemble étant à considérer comme infini

5.1.1 $$\rho(E_a, E_t) \leq 1 \qquad \forall E_a, E_t \subseteq U$$

Faute de mieux, nous considérerons pour une tâche de vérification que les probabilités a priori d'apparition de couples (*voix*, *identité*) réellement homogènes et réellement hétérogènes sont identiques. En reprenant les notations du paragraphe 2.4.4, nous avons donc

5.1.2 $$\rho(E_a, E_t) = \frac{\overline{A} + \overline{R}}{N}$$

Posons maintenant une première hypothèse qui demande que tout classificateur construit sur la base d'un ensemble d'apprentissage donné (en l'occurrence E_t) soit optimal pour ce même ensemble, au sens du critère suivant qui rend explicite le fait qu'il n'existe pas d'ensemble $E_a \neq E_t$ pouvant servir à l'apprentissage et permettant de diminuer le taux d'erreur observé en utilisant E_t comme ensemble de test

5.1.3 $$H_1: \quad \exists E_a \subset U \mid \rho(E_a, E_t) < \rho(E_t, E_t) \qquad \forall E_t \subseteq U$$

Nous tirons de cette hypothèse les conclusions suivantes

5.1.4 $$\begin{cases} \rho(U,U) \leq \rho(E_a, U) \\ \rho(E_a, E_b) \leq \rho(U, E_a) \end{cases}$$

En considérant E_a comme une variable aléatoire, nous pouvons calculer l'espérance mathématique E des deux termes de l'expression 5.1.4

5.1.5 $$\begin{cases} E(\rho(U,U)) = \rho(U,U) \leq E(\rho(E_a, U)) \\ E(\rho(E_a, E_a)) \leq E(\rho(U, E_a)) \end{cases}$$

Posons une seconde hypothèse qui demande que le test d'un nombre de plus en plus grand d'ensembles E_a, dont la variété reflète de mieux en mieux celle de U, se traduise par une convergence de la moyenne des taux d'erreur vers la probabilité d'erreur du classificateur, pour autant qu'il soit idéal vis-à-vis de U au sens de la première hypothèse

5.1.6 $$H_2: \quad E(\rho(U, E_a)) = \rho(U,U)$$

Il s'ensuit la relation suivante

5.1.7 $$E(\rho(E_a, E_a)) \leq \rho(U,U) \leq E(\rho(E_a, U))$$

- Optimisme de la méthode C: en interprétant $\rho(U,U)$ comme la probabilité d'erreur d'une méthode de reconnaissance, nous voyons que le membre de gauche de l'expression 5.1.7 y est inférieur ou égal. Il s'ensuit que, en

moyenne, lorsque l'on utilise les mêmes échantillons pour l'apprentissage et pour le test ($E_t = E_a$) le taux d'erreur observé est inférieur à la probabilité d'erreur.

- Pessimisme de la méthode U: lorsque U la distribution vraie des échantillons est utilisée pour le test ($E_t = U$), alors le taux d'erreur observé est en moyenne supérieur à la probabilité d'erreur. Dans ce dernier cas, il est fréquent en pratique de se contenter d'une approximation de U par un ensemble d'éléments de test qui ne dépendent pas de E_a.

La relation 5.1.7 n'est vérifiée que si les hypothèses H_1 et H_2 le sont aussi. Quant à la première de ces hypothèses, on peut montrer que certains classificateurs obéissent à la condition 5.1.3; c'est par exemple le cas pour un classificateur baysien. La seconde se laisse moins bien appréhender par la théorie; en pratique cependant, on considère que la condition 5.1.6 est généralement respectée.

5.2 Base de données I

Cette première base de donnée sert à mener, dans des conditions bien contrôlées, nos investigations de reconnaissance de locuteurs indépendante du texte; nous voulons en outre qu'elle se distingue par certaines conditions restrictives aptes à faciliter, dans un premier temps, la tâche de reconnaissance. Nous nous sommes donc limité à un échantillon restreint de la gent humaine, nous avons contrôlé le texte parlé par les locuteurs et nous avons surveillé les conditions d'acquisition.

5.2.1 Représentativité

Nous disposons de 10 locuteurs, 9 hommes et 1 femme, qui couvrent la tranche d'âge comprise entre vingt et quarante ans. Aucun effort n'a été entrepris pour sélectionner ces locuteurs sur la base de leur caractéristique vocale supposée; seule leur disponibilité est intervenue dans ce choix. Leur familiarité avec la langue française était très différente de cas en cas, de mauvaise à excellente. Vu notre prétention à l'indépendance du texte, nous ne considérons pas ce dernier élément comme important. Par contre, soulignons ici le fait que nous avons réalisé une unique campagne d'acquisition pour la constitution de cette base de données, en date du 20 février 1989; par conséquent, chaque locuteur ne nous a offert qu'une seule session.

5.2.2 Contenu

Les locutions contenues par cette base de données sont de nature identique, quel que soit le locuteur. Une locution consiste en une série de 20 mots courts, tous différents, dont l'ordre est aléatoire au sein de chacune des séries qui ont pu être prononcées par les locuteurs. La liste de ces mots est donnée à la figure 5.2.a, où l'on voit que la langue française est utilisée et qu'il s'agit principalement de chiffres et de nombres monosyllabiques. Chacun de ces 20 mots apparaît une fois exactement au sein d'une série.

Un	Deux	Trois	Quatre	Cinq
Six	Sept	Huit	Neuf	Dix
Onze	Douze	Treize	Quinze	Seize
Vingt	Trente	Cent	Mille	Chiffre

Figure 5.2.a Contenu des locutions

Cette base de données satisfait précisément les conditions du vocabulaire restreint énoncées au paragraphe 1.3.3. Cependant, nous n'avons à aucun moment fait usage de la connaissance explicite des mots prononcés; de fait, nous pouvons donc prétendre à l'indépendance du texte.

5.2.3 Acquisition

Chacun des 10 locuteurs a produit 8 séries de 20 mots; nous appellerons désormais locution une quelconque de ces séries. La durée de chaque locution est de 15 s exactement. La reproductibilité de cette durée, quel que soit le locuteur, a été obtenue par un mécanisme de synchronisation par dictée: les mots à prononcer s'affichaient sur un écran avec un rythme imposé et régulier que devait suivre le locuteur. Il avait cependant toute liberté quant à l'instant de départ d'une série, grâce à une touche du clavier qui lui permettait de la déclencher.

L'enregistrement des locutions s'est fait dans une salle calme, sans précautions acoustiques particulières. La distance entre les locuteurs et le microphone `Superscope EC-7` utilisé était libre de toute contrainte; elle valait en général 0.5 m environ. Le signal acoustique était transmis à un enregistreur `Marantz SD 275`, muni d'une cassette vierge `Maxell UD I 46`, avec utilisation d'un réducteur de bruit `Dolby c`. La numérisation des signaux de parole a été réalisée par une carte `Data Translation DT 2821` couplée à un micro-ordinateur

Teleprint TDC (compatible IBM AT). Le filtre de garde était un Krohn-Hite 3343, ajusté en passe-bas RC de 3400 Hz de fréquence de coupure, avec une atténuation asymptotique de 48 $dB \cdot oct^{-1}$; la cadence d'échantillonnage était de 8000 Hz et la résolution de 12 bit. Ces valeurs ont été choisies telles qu'elles soient compatibles avec un signal transmis par canal téléphonique; toutefois, aucune bande inférieure de fréquence n'a été coupée. Le gain du système de conversion analogique-numérique a été ajusté manuellement de cas en cas afin d'assurer que le signal numérique ne soit pas saturé, tout en couvrant une plage importante de la résolution (environ 11 bit).

5.2.4 Exploitation

Nous avons décidé d'exploiter notre première base de données selon une technique issue de la famille des méthodes C. Sachant que nous menons l'opération de classification sur la base d'une décision $\mathbf{D}$ fondée sur le calcul d'une distance entre le vecteur représentatif de la référence du locuteur et un vecteur de test, présentons ici le détail de notre procédé: nous avons réservé une locution de notre base de données à l'établissement du vecteur représentatif que nous avons ensuite utilisé dans le calcul des distances envers toutes les locutions restantes. Comme énoncé au paragraphe 1.6, nous nous concentrons sur une tâche de vérification; nous avons donc cherché dans l'espace des distances un seuil de décision entre domaine homogène et domaine hétérogène qui soit optimal au sens de l'erreur équitable. Nous avons ensuite dénombré le nombre d'erreurs $\overline{A}$ et $\overline{R}$ qui résultent de l'application de ce seuil aux données en cours d'examen; enfin, nous avons recommencé ces opérations pour la référence suivante jusqu'à épuisement de la base de données. En résumé, nous avons

5.2.1
$$\begin{cases} E_a = E \\ E_t = E \setminus \{\lambda_j^{(k)}\} \end{cases}$$

Si l'ensemble d'apprentissage E_a est séparé en ensemble E_r servant à l'établissement des vecteurs représentatifs et en ensemble E_s servant à l'établissement des seuils, alors

5.2.2
$$\begin{cases} E_r = \{\lambda_j^{(k)}\} \\ E_s = E_a \setminus E_r \end{cases}$$

Si l'on veut distinguer dans E_e l'ensemble E_h servant à établir les distances du domaine homogène de $E_{\bar{h}}$ celui servant à établir celles du domaine hétérogène, alors

5.2.3 $$E_e = E_h \cup E_{\bar{h}} \qquad \begin{cases} E_h = \left\{ \lambda_l^{(k)} \mid l \neq j \right\} \\ E_{\bar{h}} = \left\{ \lambda_l^{(u)} \mid u \neq k \right\} \end{cases}$$

Bien que cette approche s'inspire de la technique de l'exclusion de test (famille des méthodes U), il apparaît clairement de cette description que nul vecteur de test indépendant n'intervient, ce qui apparente indiscutablement cette technique à la famille des méthodes C; cependant, nous continuerons à la nommer technique de l'exclusion car nous sous-entendons exclusion du vecteur représentatif. Alors que la technique de l'exclusion du test exigeait, par locuteur, la construction d'autant de classificateurs que d'éléments contenus dans la base de données, notre technique de l'exclusion du vecteur représentatif se contente, par locuteur, d'autant de classificateurs que d'éléments de la base de données appartenant au dit locuteur. Nous limitons donc le nombre de classificateurs à construire en le faisant passer respectivement de 800 à 80.

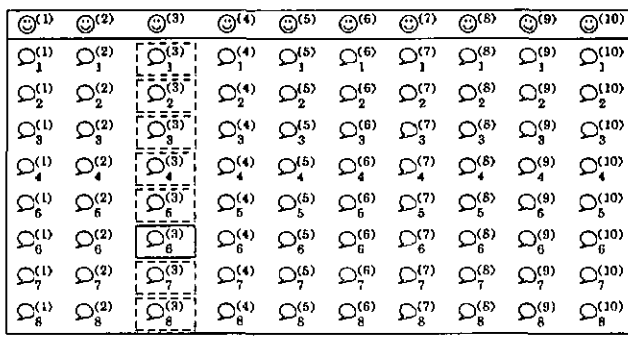

Figure 5.2.b Construction et jugement d'un classificateur

La figure 5.2.b présente un exemple d'exploitation de notre première base de données où l'on réserve $\Omega_6^{(3)}$ la sixième locution du troisième locuteur $\left\{ \textcircled{3}^{(3)} \right\}$ à l'usage du vecteur représentatif (encadrement plein), où l'on utilise les 7 autres locutions de ce locuteur pour obtenir les distances du domaine homogène (encadrements brisés) et où l'on établit les distances du domaine hétérogène avec les 72 locutions restantes (pas d'encadrement). Les ensembles de distances

obtenus permettent de construire un diagramme des taux d'erreur semblable à celui de la figure 2.5.a, où l'on comprend maintenant pourquoi la quantification était plus grossière pour la courbe de faux rejet ρ_r (7 échantillons) qu'elle ne l'était pour la courbe de fausse acceptation ρ_a (72 échantillons). On peut donc qualifier les 80 classificateurs annoncés en faisant jouer le rôle de source du vecteur représentatif à toutes les locutions de la base de données, à tour de rôle.

5.3 Base de données II

Cette seconde base de donnée sert à mener nos investigations de reconnaissance de locuteurs indépendante du texte dans un contexte plus riche que précédemment. En effet, nous nous sommes efforcé d'augmenter le nombre de cobayes tout en respectant la parité des sexes. Nous avons de plus abandonné tout contrôle sur le texte parlé par les locuteurs et nous avons admis diverses conditions d'acquisition. En outre, nous avons cherché à établir une technique d'exploitation qui appartienne à la famille des méthodes U. Enfin, nous avons pris en compte la variabilité de la voix des locuteurs entre les sessions. Ces modifications entre la base I et la base II ont pour but de permettre une évaluation des méthodes de reconnaissance qui corresponde mieux aux conditions réelles d'utilisation.

5.3.1 Représentativité

Le contenu de cette seconde base de données est constitué d'émissions radiodiffusées en modulation de fréquence par les chaînes nationales suisses RSR1 et RSR2. Les sujets de ces émissions sont divers; il s'agit par exemple de bulletins météorologiques, d'informations sportives, de journaux parlés, d'émissions littéraires ou scientifiques, de reportages. Leurs dates de radiodiffusion sont comprises entre le 18 février 1992 et le 21 février 1992. Chacun de ces quatre jours de la campagne d'acquisition a fourni une session d'un locuteur; ainsi, au plus quatre sessions sont disponibles pour chacun d'eux, même s'il est fréquent de trouver dans notre base de données des locuteurs n'en ayant produit qu'une seule.

Certains de ces locuteurs sont annonceurs de profession; on peut supposer que cette particularité influence la nature de leur expression vocale, par exemple par un effort délibéré de diction claire. Ce fait serait plutôt de nature à entraver la reconnaissance de locuteurs, puisqu'il est vraisemblable que le but recherché

par les annonceurs est de se rapprocher le plus possible d'une voix au ton neutre, accessible à tous les auditeurs. Or nous avons vu au paragraphe 4.2.2 que ce type de voix était jugé comme le plus difficile à identifier, au moins dans un contexte de reconnaissance humaine.

5.3.2 Acquisition

Nous ignorons tout des conditions de prise de son. En particulier nous ne contrôlons pas l'acoustique de la salle, nous ne savons pas si un filtre a été employé pour l'un ou l'autre des locuteurs et nous n'avons pas connaissance du type de microphone employé, qui peut donc différer d'une session à l'autre pour le même locuteur; nous sommes donc contraint d'admettre des variations éventuelles du canal de transmission [85GisH]. La radiodiffusion des émissions en modulation de fréquence a utilisé les canaux $95.1 \cdot 10^6$ Hz et $99.3 \cdot 10^6$ Hz. Leur réception et leur enregistrement ont été réalisés par une minichaîne Sony `MHC-1600 CD`, munie de cassettes vierges `Maxell UD II 60` et `Maxell XL II-S 90`, avec utilisation d'un réducteur de bruit `Dolby C`. Le filtre de garde était incorporé à une carte `Digidesign Audiomedia` couplée à un ordinateur `Macintosh IIfx` et utilisée pour la numérisation des signaux de parole; la cadence d'échantillonnage était de 8000 Hz et la résolution de 16 bit.

5.3.3 Contenu

Afin de ne conserver que l'information pertinente, nous avons édité les dialogues, ainsi que les monologues entrecoupés de plages musicales, et nous en avons extrait manuellement avec soin les passages où seule la voix d'un locuteur donné était présente, sans pour autant en supprimer les pauses naturelles du langage. Une session d'un locuteur est réalisée par l'aboutement de ces extraits. Ce travail nous a permis de récolter des échantillons de voix de 18 femmes et de 43 hommes. Les durées disponibles, par session, varient entre 19 s et 425 s; par locuteur, ces durées varient globalement entre 26 s et 1121 s. Chaque colonne de la figure 5.3.a résume le nombre de locuteurs disponibles pour un nombre de sessions donné, où F signifie féminin et M masculin, où le symbole ≥ annonce que la ligne correspondante contient le nombre de locuteurs disponibles pour un nombre de sessions égal ou supérieur à celui reporté en colonne et où le symbole = annonce que le nombre de locuteurs disponibles est exact au lieu d'être cumulé.

Nous avons décidé de découper chaque session en fragments d'une durée constante égale à 8 s. Cette découpe s'est réalisée sans égard au texte: le début

du premier fragment est synchronisé avec le début de la session, puis le début de chaque fragment suivant est synchronisé avec la fin de celui qui le précède. Le dernier fragment étant généralement d'une durée inférieure à la durée prescrite, nous l'avons rejeté.

	1	2	3	4
F ≥	18	9	6	3
F =	9	3	3	3
M ≥	43	9	7	5
M =	34	2	2	5

Figure 5.3.a Répartition par nombre de sessions

F	18	19	20	21
XxXx	6	4	13	10
AlVi	52	42		
AMSa	28			
AnMA	21	22	14	18
CaMi			12	
ChPa	10			
ElGo			11	2
FlKu				4
FrDC			51	48
GeBr			14	
IsMo		12	13	11
MaCo	5		8	8
Made	14	16	13	18
MaHu			35	
MaSa	36	29		35
MDBo				13
MiDB		6		
RuDr	13			

M	18	19	20	21
AlKo		16		
AlTi				17
AnBe	10	11		14
AnLa		17		
BeJa				13
BeZi	9	16	19	14
ChRo	15			
ChSu				20
ClFr	4			
DaFA		16		12
DaFi	27	28	43	40
DaMi				8
DrMé			19	
EmMa	12			
FlRo	22			
FrBa	15			
FrKl	38			
FrLe		10	10	10
GéMé		11		
GePo	34	37	9	39
GiBa		18		
JaSE		50		

M	18	19	20	21
JBHe	3	2	2	3
JCMa				15
JDBa			33	
JFMo				5
JJBe	11	6	11	7
JLRi	8			
JMVo		25		
LaBo	18			7
LuMé			27	
MaMa				24
MiSk	14			
PADe				8
PaNu			19	
PhHé		20		
PhTh				14
PiBa			22	
PiDu				18
PiMe		3		
RoCa	53			
RoTi			18	
XaPe		19		

Figure 5.3.b Nombre de fragments disponibles

La figure 5.3.b montre le nombre de fragments d'une durée complète de 8 s disponibles par session ainsi que les étiquettes associées aux différents locuteurs. Ces sessions s'identifient par le quantième du mois et sont reportées en colonne, tandis que les symboles F et M ont le même sens qu'à la figure 5.3.a.

■ 5.3.4 Exploitation

Nous avons décidé d'utiliser une technique de la famille des méthodes U; il s'ensuit que nous devons construire une partition de notre base de données E sous la forme d'ensembles disjoints d'apprentissage E_a et de test E_t. En outre, nous avons admis une condition encore plus sévère sur l'ensemble d'apprentissage: nous exigeons que les ensembles d'établissement des vecteurs représentatifs E_r et des seuils E_s soient eux aussi disjoints. L'ensemble de test est quant à lui soumis à une contrainte supplémentaire qui veut que les locuteurs du domaine hétérogène soient tous différents de ceux rencontrés en apprentissage. Enfin, nous imposons de respecter la parité des sexes. Il suit de ces considérations que nous découpons notre base de données en quatre parties de volume approximativement égal E_r, E_s, $E_{h,t}$ et $E_{\bar{h},t}$.

En phase de test, la prise en compte de la variabilité à long terme d'un locuteur s'obtient en choisissant des éléments de l'ensemble E_t issus d'une session qui n'ait pas déjà été utilisée en phase d'apprentissage; l'apprentissage doit donc nécessairement permettre d'ajuster les seuils de décision en prévision de cette variabilité à long terme. Pour y parvenir, nous séparons les sessions correspondant aux ensembles E_r et E_s.

Finalement, nous avons réservé à l'ensemble E_r servant à la création des vecteurs représentatifs une première session parmi celles disponibles. Nous avons encore réservé une deuxième session à l'ensemble E_s servant à la détermination des seuils de distance; enfin, une troisième session est utilisée pour l'ensemble E_t servant à tester indépendamment la qualité des classificateurs. Il découle de ces considérations et de l'examen de la figure 5.3.a que 13 locuteurs au plus nous offrent un nombre de sessions suffisant pour satisfaire nos besoins.

Certains de ces locuteurs n'ont malheureusement pas parlé pendant une durée suffisante pour nous convenir, car nous voulons au moins disposer de 10 fragments. La combinaison de cette exigence avec celle de la parité des sexes implique que seuls $5+5=10$ locuteurs de notre seconde base de données peuvent

engendrer un classificateur, alors que la liberté de choix des autres locuteurs, aptes à jouer le rôle de test indépendant, est plus grande. Nous avons donc conservé $6+6=12$ locuteurs supplémentaires, à raison d'une session unique de 10 fragments pour chacun. Il s'ensuit que nous n'utilisons finalement que 22 locuteurs parmi les 61 que contient notre base de données complète, chacun des locuteurs de référence fournissant une durée de 240 s de parole, et chacun des locuteurs de test indépendant fournissant une durée de 80 s de parole, pauses d'élocution ou de respiration incluses.

F	18	19	20	21
XxXx	6	4	10	10
AnMA	10	10		10
IsMo		10	10	10
Made	10	10		10
MaSa	10	10		10
AMSa	10			
ChPa	10			
ElGo			10	
FrDC			10	
MaHu			10	
RuDr	10			

M	18	19	20	21
AnBe	10	10		10
BeZi		10	10	10
DaFi		10	10	10
FrLe		10	10	10
GePo	10	10		10
FrKl	10			
JaSE		10		
JDBa			10	
JJBe	10			
LuMé			10	
RoCa	10			

Figure 5.3.c Base de données réduite

La figure 5.3.c est la version épurée de la figure 5.3.b. On constate que le locuteur XxXx n'atteint la durée correspondant à 10 fragments par session que si l'on accepte la fusion de deux d'entre elles.

Finalement, l'ensemble d'apprentissage $E_a = E_r \cup E_s$ est donné par

5.3.1
$$\begin{cases} E_r = \left\{ \lambda_j^{(k)} \mid k \in [1,10] \wedge j \in [1,10] \right\} \\ E_s = \left\{ \lambda_j^{(i)} \mid i \in [1,10] \wedge j \in [11,20] \right\} \end{cases}$$

L'ensemble E_s servant à établir les seuils est lui-même découpé en domaines homogènes et domaines hétérogènes

5.3.2
$$E_s = \bigcup_{k=1}^{10} E_{h,s}^{(k)} \cup \bigcup_{k=1}^{10} E_{\bar{h},s}^{(k)}$$

Ces deux domaines d'établissement des seuils sont donnés par

5.3.3
$$\begin{cases} E_{h,s}^{(k)} = \left\{ \lambda_j^{(k)} \mid j \in [11,20] \right\} & \forall k \in [1,10] \\ E_{\bar{h},s}^{(k)} = \left\{ \lambda_j^{(i)} \mid i \in [1,10] \setminus \{k\} \wedge j \in [11,20] \right\} & \forall k \in [1,10] \end{cases}$$

L'ensemble de test E_t est donné par

5.3.4
$$E_t = \bigcup_{k=1}^{10} E_{h,t}^{(k)} \cup \bigcup_{k=1}^{10} E_{\bar{h},t}^{(k)}$$

Les domaines homogènes et hétérogènes de test sont donnés par

5.3.5
$$\begin{cases} E_{h,t}^{(k)} = \left\{ \lambda_j^{(k)} \mid j \in [21,30] \right\} & \forall k \in [1,10] \\ E_{\bar{h},t}^{(k)} = \left\{ \lambda_j^{(i)} \mid i \in [11,22] \wedge j \in [21,30] \right\} & \forall k \in [1,10] \end{cases}$$

■ 5.3.5 Détails pratiques

Voici le détail de notre procédé: nous avons d'abord réservé la session 21 à la construction des vecteurs représentatifs et la session 19 à la détermination des seuils; dans ce type de processus, la session 20 du locuteur XxXx se substituera à sa session 19 chaque fois que ce sera nécessaire. Nous avons alors construit 9 références par locuteur avec 2 fragments consécutifs par vecteur représentatif, comme le montre la figure 5.3.d où le couple de nombres entre parenthèses permet de repérer les références.

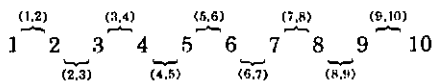

Figure 5.3.d Construction des références

Rappelons que notre but est l'évaluation de méthodes de vérification par acceptation. Pour l'atteindre dans le cas où il est fait usage d'une technique de la famille des méthodes U, il est nécessaire de disposer de seuils de rejet. Par choix arbitraire, nous avons fait équivaloir ces seuils aux seuils d'erreur équitable décrits au paragraphe 2.5.

Il nous faut donc maintenant déterminer le seuil d'erreur équitable associé à une référence par l'observation des distances obtenues pour le domaine homo-

gène entre le vecteur représentatif de cette référence et tous les 10 fragments de la session 19 du locuteur considéré, et pour le domaine hétérogène entre le vecteur représentatif de cette référence et 24 fragments tirés au sort parmi ceux de la session 19 d'autres locuteurs. Ce tirage au sort a été guidé par les principes suivants: seuls les 10 locuteurs engendrant des références peuvent y participer, car nous voulons imposer les conditions réelles, où la connaissance de la variabilité des locuteurs est imparfaite; parmi les 24 fragments à tirer au sort, 12 proviennent de locuteurs masculins et autant de locuteurs féminins, car nous voulons faire respecter la parité des sexes; chaque locuteur fournit 3 fragments exactement, car aucun locuteur ne doit être représenté plus qu'un autre.

Après avoir déterminé le seuil d'erreur équitable sur la base de l'ensemble des locuteurs de référence, nous pouvons établir la valeur des distances de la référence en cours d'examen et des échantillons du domaine homogène ou hétérogène qui n'ont pas encore été utilisés; ici interviennent donc enfin les sessions 18 ou 20, selon le locuteur considéré (18 et 19 pour le locuteur XxXx). Nous avons jugé la qualité du classificateur en prenant en compte le même nombre de distances des domaines de rejet ou d'acceptation que celui rencontré dans l'étape de construction du classificateur: 10 fragments du locuteur considéré et 24 fragments tirés au sort parmi ceux d'autres locuteurs. Les principes d'équité du tirage au sort ont été conservés; ils se traduisent par un choix aléatoire de 2 fragments dans chaque session des 12 locuteurs indépendants.

La figure 5.3.e donne l'exemple du classificateur associé à la référence (1,2) de la session 21 du locuteur AnMa, où nous avons utilisé le symbole P pour marquer les fragments utilisés dans la construction de la référence, le symbole Σ pour marquer ceux en rapport avec la partie homogène de la détemination du seuil d'erreur équitable, et le symbole σ pour la partie hétérogène. On constate dans cet exemple que les principes guidant le tirage au sort ont nécessité l'exclusion du locuteur AnBe; à cet égard, nous avouons n'avoir pas cherché à obtenir une égalité de traitement exacte envers tous les locuteurs, car le tirage au sort du locuteur exclu a été conduit, à chaque nouvelle référence, sans autres contraintes que celle de la parité des sexes. L'estimation de la qualité du classificateur se réalise à l'aide des fragments marqués Θ pour le domaine homogène et marqués θ pour le domaine hétérogène.

Si l'on examine un grand nombre d'exemples construits sur le modèle de la figure 5.3.e, alors on constate l'existence d'une trop grande fréquence d'appari-

tion de paires de fragments consécutifs dans la partie servant à la détermination des seuils. Ce fait s'explique par notre façon d'exploiter le générateur de nombres pseudo-aléatoires mis à notre disposition par le système d'exploitation VMS V5.4-2: dans le choix du quantième fragment, le nombre flottant issu de ce générateur a simplement été multiplié par une constante adéquate, puis tronqué. Ce mode d'action revient à rejeter une partie de la quantité d'information disponible. Dans ces conditions, les propriétés statistiques du générateur ne sont plus garanties et nous en voyons ici les effets. Nous avons toutefois admis l'hypothèse selon laquelle cet aspect influence le résultat global d'une façon trop mineure pour mériter d'être approfondie.

Sexe	Nom	Session	1	2	3	4	5	6	7	8	9	10	ϵ
F	AnMa	21	P	P									E_r
F	XxXx	20		σ	σ					σ			$E_{\bar{h},s}$
F	AnMa	19	Σ	Σ	Σ	Σ	Σ	Σ	Σ	Σ	Σ	Σ	$E_{h,s}$
F	lsMo	19		σ							σ	σ	$E_{\bar{h},s}$
F	Made	19				σ	σ				σ		$E_{\bar{h},s}$
F	MaSa	19		σ	σ				σ				$E_{\bar{h},s}$
M	AnBe	19											
M	BeZi	19				σ	σ					σ	$E_{\bar{h},s}$
M	DaFi	19	σ	σ								σ	$E_{\bar{h},s}$
M	FrLe	19					σ	σ				σ	$E_{\bar{h},s}$
M	GePo	19	σ	σ						σ			$E_{\bar{h},s}$
F	AnMa	18	Θ	Θ	Θ	Θ	Θ	Θ	Θ	Θ	Θ	Θ	$E_{h,t}$
F	AMSa	18	θ				θ						$E_{\bar{h},t}$
F	ChPa	18						θ			θ		$E_{\bar{h},t}$
F	ElGo	20							θ	θ			$E_{\bar{h},t}$
F	FrDC	20						θ				θ	$E_{\bar{h},t}$
F	MaHu	20				θ		θ					$E_{\bar{h},t}$
F	RuDr	18		θ							θ		$E_{\bar{h},t}$
M	FrKl	18	θ								θ		$E_{\bar{h},t}$
M	JaSE	19	θ	θ									$E_{\bar{h},t}$
M	JDBa	20								θ	θ		$E_{\bar{h},t}$
M	JJBe	18		θ								θ	$E_{\bar{h},t}$
M	LuMé	20					θ			θ			$E_{\bar{h},t}$
M	RoCa	18	θ				θ						$E_{\bar{h},t}$

Figure 5.3.e Exemple d'estimation d'un classificateur

Nous venons de construire 90 classificateurs en considérant l'ensemble des sessions 21 pour la génération des références et l'ensemble des sessions 19 pour la détermination des seuils. Il suffit de permuter l'usage de ces deux sessions pour doubler le nombre de classificateurs dont ont désire juger la qualité moyenne, opération à laquelle nous nous sommes livré sans pour autant remettre en cause la parenté de notre technique d'exploitation de cette seconde base de données avec la famille des méthodes U, puisque nous continuons à respecter nos principes ($E_a \cap E_t = \varnothing$ et $E_r \cap E_t = \varnothing$). Par contre, nous n'avons pas entrepris l'inclusion d'autres sessions dans cette astuce de permutation.

5.4 Comparaison des bases de données

La comparaison de bases de données fait intervenir de nombreuses considérations, par exemple sur le nombre de locuteurs utilisés, sur le nombre de sessions disponibles et la durée des intervalles qui les séparent, sur la qualité et la stabilité des conditions d'acquisition ou encore sur le degré de dépendance du texte prononcé. L'exploitation de ces bases de données peut se distinguer non seulement par son principe (par exemple méthode C ou méthode U, inclusion ou non de E_r dans E_h) mais encore de façon significative par des choix tels que le mélange ou la séparation des sessions, ou par des paramètres tels que la durée des fragments de test et de ceux destinés à la construction des références. Enfin, la pondération appliquée à toutes ces caractéristiques dépend fortement du point de vue que l'on cherche à mettre en valeur (rapidité de calcul, conservation des performances face à un canal bruité, simplicité d'apprentissage par exemple).

Nous proposons ici un critère objectif de comparaison facile à déterminer: il s'agit du nombre de distances calculées pour une technique de reconnaissance donnée, par base de données. Les figures 5.4.a à 5.4.f donnent ce nombre pour chaque expérience menée sur les bases de données rencontrées dans cette thèse. On constate que celles que nous avons construites, conjointement avec notre façon de les exploiter, soutiennent sereinement la comparaison avec les articles que nous avons choisi de présenter ici. En particulier, il faudrait même encore doubler le nombre annoncé de distances calculées pour notre seconde base de données si l'on acceptait de tenir compte de la phase d'estimation des seuils.

9 locuteurs = 0 F + 9 M
1 session / locuteur
15 locutions / session
0.75 s / locution
1 référence / locuteur
$\mathrm{Card}(E_a) \to \infty$
$\mathrm{Card}(E_t)$ = 15 locutions / locuteur
$E_t \subseteq E_a$
15 auditeurs
$2025 \times$ ▮

Figure 5.4.a Nombre de distances pour [63ComA], §4.2.1

10 locuteurs = 0 F + 10 M	10 locuteurs = 0 F + 10 M
2 sessions / locuteur	2 sessions / locuteur
	3 locutions / session
95 s / session	2 s / locution
1 référence / locuteur	
$\mathrm{Card}(E_a)$ = 1 session / référence	$\mathrm{Card}(E_h)$ = 3 paires de locutions / locuteur
$\mathrm{Card}(E_t)$ = 1 session / locuteur	$\mathrm{Card}(E_{\bar{h}})$ = 90 paires de locutions
$E_a \cap E_t = \varnothing$	
100 auditeurs	24 auditeurs
$1000 \times$ ▮	$2880 \times$ ▼

Figure 5.4.b Nombre de distances pour [91KreJ], §4.2.2 et §4.2.3

12 locuteurs = 0 F + 12 M
6 sessions / locuteur
4 locutions / session
5 s / locution $\in E_t$
6 références / locuteur
100 s / référence $\in E_a$
$\mathrm{Card}(E_h)$ = 4 tests / référence
$\mathrm{Card}(E_{\bar{h}})$ = 44 tests / référence
$E_t \subseteq E_a$
$3456 \times$ ▼

Figure 5.4.c Nombre de distances pour [82ShrM], §4.3.1

10 locuteurs = 5 F + 5 M
5 sessions / locuteur
40 locutions / session
1 locution = 1 chiffre $\in \{0,1,2,3,4,5,6,7,8,9\}$
1 référence / locuteur
Card(E_a) = 100 locutions / locuteur
Card(E_t) = 100 locutions / locuteur
$E_a \cap E_t = \emptyset$
1000 × ▮

Figure 5.4.d Nombre de distances pour [88SooF], §4.3.2

10 locuteurs = 1 F + 9 M
1 session / locuteur
8 locutions / session
15 s / locution $\in E_t$
8 références / locuteur
15 s / référence $\in E_a$
Card(E_h) = 56 tests / locuteur
Card($E_{\bar{h}}$) = 576 tests / locuteur
$E_t \subseteq E_a$
6320 × ▼

Figure 5.4.e Nombre de distances pour notre base de données I, §5.2

10 locuteurs = 5 F + 5 M $\subset E_a$
12 locuteurs = 6 F + 6 M $\subset E_t$
3 sessions / locuteur
10 locutions / session
8 s / locution $\in E_t$
18 références / locuteur
16 s / référence $\in E_a$
Card(E_h) = 10 tests / référence
Card($E_{\bar{h}}$) = 24 tests / référence
$E_a \cap E_t = \emptyset$
6120 × ▼

Figure 5.4.f Nombre de distances pour notre base de données II, §5.3

6 Répétition d'expériences

Nous avons voulu tester sur nos propres bases de données les méthodes classiques de reconnaissance automatique du locuteur, dans le but de comparer les performances que nous en obtenons avec celles de la littérature et d'en déduire l'importance ou l'insignifiance éventuelle des différences de condition d'acquisition, d'analyse ou d'exploitation; ainsi, nous avons notamment répété les expériences des paragraphes 4.3.1 [82SbrM] et 4.3.2 [88SooF]. En particulier, nous avons appliqué la méthode de la moyenne à long terme au cepstre complexe du filtre de synthèse de l'analyse par prédiction linéaire en réalisant la comparaison selon une distance euclidienne, euclidienne pondérée et de Mahalanobis. Quant à la méthode de l'erreur moyenne de quantification vectorielle, nous l'avons appliquée aux cepstres complexes du filtre de synthèse ainsi qu'aux cepstres complexes différentiels du filtre de synthèse. Nous terminons ce chapitre par une méthode de reconnaissance basée sur la comparaison de fréquences fondamentales moyennes.

6.1 Cepstre complexe moyen du filtre de synthèse

Les propriétés homomorphiques de l'analyse cepstrale [68OppA] étayent l'affirmation selon laquelle la moyenne temporelle $\langle c^{(k)}(n) \rangle$, donnée à l'expression 3.4.1, de tous les cepstres à court terme d'une locution $\lambda^{(k)}$ d'un locuteur (k) serait proportionnelle à la somme de deux contributions: d'une part $\langle c_i^{(k)}(n) \rangle$ le cepstre correspondant au filtre créé par le conduit vocal moyen du locuteur, et d'autre part celui correspondant à toutes les fonctions de transfert liées aux conditions d'acquisition supposées invariables dans le temps, telles que par exemple $c_s(n)$ l'acoustique de la salle, $c_m(n)$ la réponse en fréquence du microphone ou $c_e(n)$ de l'enregistreur, $c_b(n)$ celle du matériau de la bande magnétique ou encore $c_f(n)$ celle du filtre de garde

6.1.1 $$\langle c^{(k)}(n)\rangle = \underbrace{\langle c_l^{(k)}(n)\rangle}_{\text{Locuteur}} + \underbrace{c_s(n)+c_m(n)+c_e(n)+c_b(n)+c_f(n)}_{\text{Canal invariant}} \qquad \forall n \geq 0$$

Si l'on admet que les éléments de la seconde contribution ont été identiques pour tous les locuteurs, nous pouvons espérer que la différence entre les cepstres moyens issus de deux locutions $\lambda^{(i)}$ et $\lambda^{(k)}$ décrit directement la différence morphologique entre les deux locuteurs (i) et (k) qui les ont prononcées [72FurS]. Cependant, rappelons ici que l'hypothèse d'uniformité des canaux de transmission est mieux vérifiée par la première base de données que par la seconde; il s'ensuit que nous nous attendons à voir ce fait se refléter par une dégradation de la qualité de la reconnaissance lorsque nous passerons de l'une à l'autre.

6.1.1 Principe du cepstre complexe moyen

Un vecteur caractéristique moyen $\langle \mathbf{c}^{(k)} \rangle$ d'un locuteur (k) peut être construit en considérant plusieurs ordres cepstraux n dans l'expression 3.4.1. Pour la reconnaissance de locuteurs, il est fréquent de renoncer au coefficient $\langle c(0) \rangle$ et de ne considérer les coefficients d'ordre supérieur que jusqu'à l'ordre p d'analyse

6.1.2 $$\langle \mathbf{c}^{(k)} \rangle = \begin{pmatrix} \langle c^{(k)}(1) \rangle \\ \langle c^{(k)}(2) \rangle \\ \vdots \\ \langle c^{(k)}(p) \rangle \end{pmatrix}$$

Finalement, la décision $\mathbf{D}$ de l'expression 2.5.1 est basée sur le calcul d'une distance $d(\mathbf{x},\mathbf{y})$ entre un vecteur caractéristique moyen $\mathbf{x} = \langle \mathbf{c}^{(k)} \rangle$ considéré comme référence, et un autre vecteur caractéristique moyen $\mathbf{y} = \langle \mathbf{c}^{(i)} \rangle$ considéré comme vecteur de test.

La figure 6.1.a montre deux cepstres complexes moyens du filtre de synthèse issus de deux locuteurs masculins différents, tandis que la figure 6.1.b montre une superposition de cepstres complexes moyens issus du premier locuteur et la figure 6.1.c une superposition de cepstres complexes moyens issus du second locuteur.

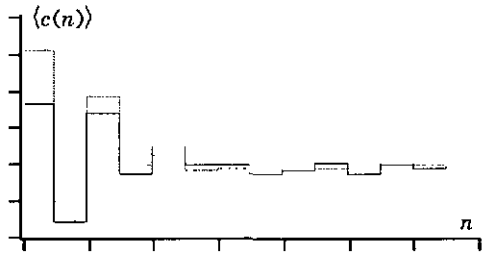

Figure 6.1.a Paire de cepstres complexes moyens

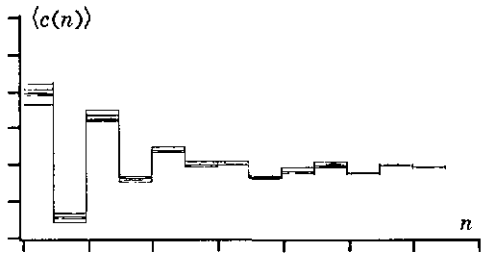

Figure 6.1.b Ensemble de cepstres complexes moyens du premier locuteur

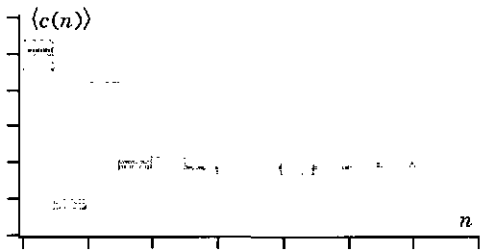

Figure 6.1.c Ensemble de cepstres complexes moyens du second locuteur

La différence la plus perceptible entre les cepstres complexes moyens de l'un ou de l'autre locuteur s'observe pour les coefficients d'ordre faible. Par exemple, le coefficient $\langle c(1) \rangle$ du premier locuteur, auquel nous avons associé des traits pleins, paraît posséder une valeur systématiquement inférieure à celle du même coefficient du second locuteur, auquel nous avons associé des traits brisés. C'est l'ensemble de ces différences que nous sommes résolu à exploiter.

6.1.2 Conditions d'analyse

Les conditions d'acquisition ayant déjà été décrites aux paragraphes 4.3.1 et 5.2.3, nous n'y reviendrons pas. Par contre, les paramètres présidant à l'analyse cepstrale du paragraphe 3.3 n'ont pas encore été spécifiés. L'objet de la figure 6.1.d est de combler cette lacune.

			[82ShrM]	Base I	Base II
Passe-bas	Hz		4500	3400	~ 3500
Passe-haut	Hz		150	0	0
Echantillonnage	Hz		10000	8000	8000
Résolution	bit		14	12	16
Pauses			voix laryngée	incluses	incluses
Fenêtre		w	Hamming	Bartlett	Bartlett
Support	s		0.020	0.030	0.030
Pas	s		0.020	0.010	0.010
Préaccentuation		μ	0.94	0.95	0.95
Ordre		p	12	14	14
Distance		d	Karhunen-Loève	euclidienne	pondérée

Figure 6.1.d Conditions d'analyse

On voit que la rubrique **Ordre** de la figure 6.1.d ne livre qu'une seule valeur, valable à la fois pour l'analyse par prédiction linéaire et pour la transformation de son résultat lors de l'application de la récurrence 3.3.1. La raison en est que l'on considère généralement un nombre de coefficients cepstraux qui n'est pas différent de l'ordre d'analyse p de l'équation 3.2.1, tout en renonçant au coefficient cepstral à l'origine $c(0)$. L'abandon de ce coefficient résulte du fait qu'il ne dépend que de la condition de normalisation unitaire du premier coefficient du filtre de synthèse ($a_0 = 1$). Il s'ensuit que sa valeur n'est pas directement dépendante du locuteur, au contraire du coefficient cepstral à l'origine que l'on obtiendrait si l'on estimait le cepstre du signal de parole $s(n)$, sans passer par l'analyse par prédiction linéaire [91Fus].

La comparaison de nos conditions avec celles de [82ShrM] fait apparaître quelques différences que nous supposons mineures, comme par exemple celles qui se rapportent au taux de recouvrement, au type de fenêtre ou encore à la résolution numérique. Nous considérons au contraire comme majeure la différence entre notre approche et celle que proposent les auteurs de l'article

[82ShrM] dans la sélection des éléments du signal qui participent au calcul de leur valeur moyenne. Alors que nous acceptons la totalité de la locution comme source de vecteurs caractéristiques, y compris les pauses, qu'elles soient d'élocution, de respiration, de surprise ou dues à une subite envie d'éternuer, les auteurs de [82ShrM] ne tolèrent que les vecteurs caractéristiques correspondant à une voix laryngée.

■ 6.1.3 Cepstre complexe moyen et base de données I

Les figures 6.1.e et 6.1.f présentent respectivement $\overline{A}$ les erreurs d'acceptation observées sur le domaine hétérogène et $\overline{R}$ les erreurs de rejet observées sur le domaine homogène, quand le vecteur caractéristique est constitué par le cepstre moyen $\langle c(n) \rangle$ et quand la mesure de comparaison est une distance euclidienne. La méthodologie présidant à cette expérience a été décrite au paragraphe 5.2.

$64 \cdot \rho_a$	☺(1)	☺(2)	☺(3)	☺(4)	☺(5)	☺(6)	☺(7)	☺(8)	☺(9)	☺(10)	☺(i)
☺(1)		0	0	0	0	0	0	2	0	0	2
☺(2)	0		1	2	3	24	9	3	0	0	42
☺(3)	0	0		0	0	0	0	0	0	0	0
☺(4)	0	29	26		5	3	0	0	0	14	77
☺(5)	0	19	5	2		3	0	0	14	0	43
☺(6)	0	39	0	0	0		0	0	3	0	42
☺(7)	0	31	1	0	0	3		12	0	0	47
☺(8)	0	8	8	0	0	0	4		0	0	20
☺(9)	0	6	0	0	6	9	0	0		0	21
☺(10)	0	0	0	0	0	0	0	0	0		0
☺(i)	0	132	41	4	14	42	13	17	17	14	294
$\langle \rho_a \rangle$											5.1%

Figure 6.1.e Fausses acceptations

$56 \cdot \rho_r$	☺(1)	☺(2)	☺(3)	☺(4)	☺(5)	☺(6)	☺(7)	☺(8)	☺(9)	☺(10)	☺(k)
	0	13	3	0	0	5	0	1	1	1	24
$\langle \rho_r \rangle$											4.3%

Figure 6.1.f Faux rejets

Chaque colonne des figures 6.1.e et 6.1.f représente l'ensemble des erreurs commises pour toutes les tâches de vérification associées à la référence du locuteur (k). Chaque ligne de la matrice du domaine hétérogène représente les fausses acceptations $\overline{A}$ commises en tentant d'utiliser comme imposteur le locuteur (i) correspondant; en accord avec le paragraphe 5.1, 64 est le nombre maximal d'erreurs possibles par élément de cette matrice, puisque chaque imposteur potentiel (i) offre 8 locutions à chacune des 8 références du locuteur (k). Dans le cas de la matrice répertoriant les faux rejets $\overline{R}$, ce nombre maximal vaut 56. Nous avons encore donné dans les marges de ces matrices la somme des erreurs selon leurs lignes et leurs colonnes, ainsi que les taux moyens d'erreur.

Illustrons l'interprétation de ces matrices par quelques exemples choisis arbitrairement: on constate que, pour cette méthode, les références associées à la locutrice $\{☺^{(6)}\}$ (l'unique élément féminin de notre première base de données) se sont laissées berner 3 fois par le locuteur $\{☺^{(7)}\}$, alors que les références associées à ce dernier se sont montrées intransigeantes face à la première. Globalement, les références de cette même locutrice $\{☺^{(6)}\}$ ont d'ailleurs conduit à 42 fausses acceptations, c'est-à-dire à 42 décisions incorrectes d'homogénéité entre les auteurs des locutions de test et l'identité à laquelle ils ont prétendu; ce laxisme ne se retrouve pas dans le domaine homogène, la locutrice en question s'étant montrée peu accommodante envers sa personne puisqu'elle a rejeté pas moins de 5 fois la maternité de ses propres locutions. Par contre, ses tentatives d'imposture se sont conclues 42 fois par un succès, alors que celles du locuteur $\{☺^{(7)}\}$ ont réussi 47 fois.

Les seuils de décision du paragraphe 2.3 ont été choisis de sorte à respecter au mieux, indépendamment pour chacun des 80 classificateurs, l'équilibre entre les faux rejets et les fausses acceptations; c'est cette indépendance qui explique l'asymétrie de la matrice des erreurs d'acceptation. Le résultat global découlant du choix de ces seuils est un taux $\frac{1}{2}\cdot(\rho_a+\rho_r)$ valant 4.7%. Nous appellerons cette grandeur un taux moyen d'erreur équitable, par opposition à la fois au taux d'erreur équitable, dont il n'est que l'estimation, et au taux moyen d'erreur, que nous rencontrerons quand nous ferons usage de seuils de décision a priori.

Nous constatons que notre taux moyen d'erreur équitable vaut 1.8% de plus que le taux atteint par [82ShrM]. Rappelons que, chez ce dernier, la durée d'estimation d'une référence vaut 100 s, pauses exclues, alors que la nôtre est de

15 s seulement, pauses incluses; en revanche, les durées d'estimation d'un vecteur de test valent respectivement 5 s et 15 s.

6.1.4 Cepstre complexe moyen et base de données II

Cette seconde base de données est attendue comme plus difficile que la première, en raison à la fois d'un nombre plus élevé de locuteurs et d'une variabilité plus grande du texte prononcé. Le fait que plusieurs sessions interviennent complique encore le problème de reconnaissance en raison de l'apport accru des fluctuations de la voix d'un locuteur au cours du temps [74FurS]. Si la durée d'estimation des références est peu modifiée, passant de 15 s à 16 s, en revanche celle des tests diminue et ne vaut plus que 8 s; enfin, notre méthode d'exploitation de cette base de données implique l'usage de seuils a priori, alors que jusque là nous n'avions fait usage que de seuils a posteriori. Nous allons examiner ci-dessous l'ampleur des conséquences de ces surcroîts de difficultés.

$36 \cdot \rho_a$	XxXx	AnMa	IsMo	Made	MaSa	AnBe	BeZi	DaFi	FrLe	GePo	
AMSa	25	29	3	0	13	0	12	16	0	0	98
ChPa	6	8	0	0	0	0	8	8	0	0	30
ElGo	25	27	7	2	19	1	22	11	0	0	114
FrDC	36	36	24	35	30	21	33	36	1	3	255
MaHu	24	21	1	14	4	2	19	21	0	3	109
RuDr	20	16	3	0	15	0	6	8	0	0	68
FrKl	33	28	12	13	17	10	27	30	0	0	170
JaSE	24	17	7	3	7	0	14	12	0	0	84
JDBa	33	33	20	18	26	13	33	31	6	8	221
JJBe	36	36	28	29	36	12	36	36	3	7	259
LuMé	36	36	35	36	36	25	36	36	12	14	302
RoCa	3	4	0	0	0	0	0	4	0	0	11
	301	291	140	150	203	84	246	249	22	35	1721
$\langle \rho_a \rangle$											39.8%

Figure 6.1.g Fausses acceptations

$180 \cdot \rho_r$	XxXx	AnMa	IsMo	Made	MaSa	AnBe	BeZi	DaFi	FrLe	GePo	
	104	109	117	159	132	89	80	177	18	22	1007
$\langle \rho_r \rangle$											55.9%

Figure 6.1.h Faux rejets

A) Cepstre complexe moyen en métrique euclidienne

Les figures 6.1.g et 6.1.h sont l'équivalent des figures 6.1.e et 6.1.f dans le cas de la seconde base de données. La galanterie nous a poussé à y disposer en première place les ressortissantes du beau sexe, autant selon les colonnes que selon les lignes; les locutrices sont explicitement isolées des locuteurs par une ligne de séparation. La procédure d'exploitation de cette base de données a été déjà expliquée au paragraphe 5.2; elle conduit à exactement 36 actes de vérification pour chaque élément de la matrice du domaine hétérogène et 180 actes de vérification par élément du domaine homogène. Il découle de ces figures que le taux de fausse acceptation ρ_a vaut 39.8% et que celui de faux rejet ρ_r vaut 55.9%, le taux moyen d'erreur valant donc 47.9%.

Ce résultat est exécrable. Il a été obtenu en comparant un vecteur de test et un vecteur de référence au moyen de la distance euclidienne décrite au paragraphe 2.3.1; or, d'autres métriques classiques existent et ont été proposées pour la reconnaissance de locuteurs, par exemple par les auteurs des documents [83ShrM, 88NodH, 89BasM]. Nous allons utiliser deux d'entre elles dans l'espoir d'améliorer les résultats de reconnaissance.

B) Cepstre complexe moyen en métrique euclidienne pondérée

Les bons résultats obtenus lors du test de la méthode du cepstre complexe moyen du filtre de synthèse sur notre première base de données nous encouragent à ne pas abandonner tout espoir; nous avons donc tenté d'améliorer les résultats obtenus sur notre seconde base de données en choisissant comme autre méthode de comparaison la distance euclidienne pondérée décrite au paragraphe 2.3.2, le rôle du vecteur de pondération **w** étant tenu par l'inverse des coefficients de la diagonale de la matrice de covariance établie en utilisant les fragments de l'ensemble d'apprentissage.

Dans le détail, pour chacune des deux sessions correspondant respectivement à E_r et E_a dans l'ensemble d'apprentissage E_a, nous avons pris, par locuteur, les 10 cepstres moyens disponibles et nous en avons construit une matrice de covariance. Le calcul de la moyenne de ces 20 matrices nous a permis d'établir une matrice unique dont les éléments de la diagonale principale, identiques à la variance σ^2 des composantes des vecteurs cepstraux moyens, ont été inversés et ont fourni les poids w_i^2 de l'expression 2.3.4. Les résultats obtenus par l'usage de cette métrique sont donnés aux figures 6.1.i et 6.1.j, pour lesquelles nous espérons qu'il ne soit plus nécessaire de préciser l'ordonnance.

$36 \cdot \rho_a$	XxXx	AnMa	IsMo	Made	MaSa	AnBe	BeZi	DaFi	FrLe	GePo	
AMSa	0	0	0	11	0	0	12	2	0	0	25
ChPa	14	4	11	11	11	0	0	0	0	0	51
ElGo	8	0	16	9	1	7	6	6	4	0	57
FrDC	0	0	0	3	0	3	0	0	6	0	9
MaHu	15	2	3	29	18	0	7	14	11	0	102
RuDr	0	4	0	1	1	0	0	2	1	0	9
FrKl	0	2	0	3	0	35	0	1	0	0	6
JaSE	0	18	0	0	0	4	20	30	13	0	116
JDBa	0	0	1	2	0	5	0	8	7	2	24
JJBe	0	0	0	0	0	12	0	11	4	12	32
LuMé	0	0	0	0	0	0	2	3	0	0	17
RoCa	0	0	0	0	0	0	0	0	0	0	0
	37	30	31	69	31	66	47	77	46	14	448
$\langle \rho_a \rangle$											10.4%

Figure 6.1.i Fausses acceptations

$180 \cdot \rho_r$	XxXx	AnMa	IsMo	Made	MaSa	AnBe	BeZi	DaFi	FrLe	GePo	
	27	26	18	48	79	67	56	35	3	36	395
$\langle \rho_r \rangle$											21.9%

Figure 6.1.j Faux rejets

Les taux de fausse acceptation ρ_a et de faux rejet ρ_r valent respectivement 10.4% et 21.9%, ce qui nous conduit à un taux moyen d'erreur valant 16.2%. Bien que l'utilisation d'une distance euclidienne pondérée ait permis une réduction importante de ce taux d'erreur, nous devons constater une différence sensible quant à son effet sur le taux de fausse acceptation et sur celui de faux rejet, ce dernier prévalant largement.

C) Cepstre complexe moyen en métrique de Mahalanobis

Les taux de fausse acceptation ρ_a et de faux rejet ρ_r résultant de l'usage de la métrique de Mahalanobis décrite au paragraphe 2.3.3 valent respectivement 6.2% et 25.7%, conduisant certes à une moyenne à peine inférieure à celle obtenue dans le cas de la métrique euclidienne pondérée, mais aussi à une différence encore plus marquée entre les deux taux d'erreur.

Cette différence s'explique par le fait que la détermination des seuils de classification entre domaine hétérogène et domaine homogène est imparfaite. Cette imperfection résulte d'une adaptation trop grande des éléments de pondération aux données mêmes dont ils sont issus, typique d'une spécialisation excessive des classificateurs. En effet, considérons le nombre de valeurs scalaires ayant servi à déterminer les pondérations relativement au nombre de pondérations résultantes. Dans ce calcul, nous prenons en compte le fait que nous disposons de 14 valeurs indépendantes par vecteur cepstral moyen, de 2 sessions, de 10 fragments par locuteur et de 10 locuteurs.

Distance	poids	données	rapport
d_2	0	2800	∞
$d_{2,\text{pond}}$	14	2800	200
$d_{2,\text{Maha}}$	105	2800	27

Figure 6.1.k Représentativité des facteurs de pondération

La figure 6.1.k permet de constater qu'un facteur de pondération d'une métrique de Mahalanobis représente globalement bien moins de données qu'un facteur de pondération de la métrique euclidienne.

Si cette représentativité des facteurs de pondération était sans effet sur les classificateurs, alors nous pourrions nous attendre à ce que les caractéristiques associées au domaine homogène varient peu entre les conditions d'apprentissage $E_r \cup E_{h,s}$ et celles de test $E_{h,t}$; par contre ceci n'est plus vrai pour le domaine hétérogène car la variabilité des locuteurs s'y modifie de $E_{\bar{h},s}$ à $E_{\bar{h},t}$, ne fût-ce que parce que le nombre de locuteurs différents est plus élevé sur l'ensemble de test. Dans ce cas, la distribution des distances observées sur le domaine hétérogène s'élargirait plus que celle des distances homogènes et le seuil estimé sur l'ensemble d'apprentissage deviendrait trop grand relativement à celui qui conviendrait au mieux sur l'ensemble de test.

En pratique toutefois, constatons qu'un autre effet s'oppose à celui auquel nous nous attendons, au point de le dominer. La spécialisation excessive vis-à-vis de l'ensemble d'apprentissage conduit à un élargissement marqué de la distribution des distances quand nous passons de $E_r \cup E_{h,s}$ à $E_{h,t}$, car les pondérations de Mahalanobis sont trop bien adaptées à l'ensemble d'apprentissage et trop peu à celui de test. Dans ce cas, la distribution des distances observées sur le domaine

hétérogène s'élargit moins que celle des distances homogènes et le seuil estimé sur l'ensemble d'apprentissage devient trop petit relativement à celui qui conviendrait au mieux sur l'ensemble de test.

On pourrait se poser la question de savoir pourquoi nous n'avons pas réalisé l'estimation de la matrice de covariance sur un ensemble de locuteurs d'une part qui n'interviendrait ni dans la construction des classificateurs ni dans l'estimation de leur qualité, et qui serait d'autre part d'un cardinal suffisant, autant du point de vue des données disponibles que de celui de la variabilité des locuteurs en nombre et dans le temps. La réponse est que nous avons voulu nous placer dans les conditions habituelles d'utilisation de la reconnaissance automatique de locuteurs, où les données observables sont limitées; nous nous sommes ainsi contraint à ne pas utiliser d'autres éléments que ceux à disposition dans la base de données réduite présentée à la figure 5.3.c.

En résumé, l'usage d'une pondération de Mahalanobis a entraîné des seuils de décision trop petits par rapport à ceux correspondant à l'erreur équitable, ce qui se manifeste par exemple dans les diagrammes de vérification. Expliquons cette observation en confrontant l'ensemble de détermination des seuils $E_s \subset E_a$ et l'ensemble de test E_t grâce aux diagrammes de vérification ainsi qu'aux histogrammes des distances de chacun des deux domaines.

L'histogramme $\mathbb{H}(E_s)$ de la figure 6.1.l présente une estimation de la densité de probabilité des distances que l'on rencontre lors de la phase de construction des classificateurs utilisant une métrique de Mahalanobis. Ces distances ont été centrées par la soustraction du seuil de décision μ correspondant, ce qui explique l'apparition de distances négatives pour le domaine homogène qui forme un pic plus étroit et de mode inférieur à celui du domaine hétérogène. Le second histogramme $\mathbb{H}(E_t)$ a été établi lors de la phase de test, avec soustraction des seuils issus de la phase précédente. On constate que les modes des deux domaines se sont très peu déplacés; par contre les deux distributions se sont élargies, principalement du côté des distances de valeur élevée, avec pour corollaire une diminution de la hauteur des pics modaux. Ceci tout à la fois conforte notre attente d'une diversité plus grande du domaine hétérogène et met en évidence la faiblesse de la métrique de Mahalanobis quand la matrice de covariance est estimée de façon peu robuste sur un univers de locuteurs trop petit. En effet, si cette estimation était convenable, le domaine homogène n'aurait pas dû se modifier entre la phase d'apprentissage et la phase de test.

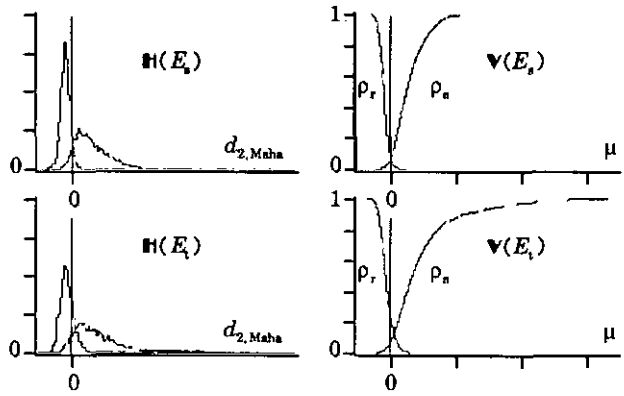

Figure 6.1.l Cepstres moyens et distances de Mahalanobis

La partie droite de la figure 6.1.l présente les diagrammes de vérification $\mathbf{V}(E_s)$ et $\mathbf{V}(E_t)$ correspondant aux histogrammes de la partie gauche de cette même figure. On y constate clairement d'une part l'inadéquation de la détermination des seuils d'erreur équitable, et d'autre part l'extension du support des distances du domaine hétérogène observées dans la phase d'estimation de la qualité des classificateurs.

En raison de l'inadéquation citée, nous renonçons à utiliser la métrique de Mahalanobis dans la suite de cette thèse; par contre, nous continuerons à exploiter les avantages liés à sa version simplifiée que représente la métrique euclidienne pondérée.

6.1.5 Résumé

Le vecteur caractéristique que nous avons utilisé dans les expériences que nous venons de décrire est la moyenne temporelle de cepstres complexes à court terme du filtre de synthèse rencontrés dans notre première et dans notre seconde base de données. Les seuils de vérification ont été déterminés comme étant ceux susceptibles de produire un taux d'erreur équitable sur les ensembles ayant servi à la construction des classificateurs. Il découle de notre méthodologie d'exploitation des bases de données qu'il s'agit de seuils a posteriori pour la base I et a priori pour la base II. La figure 6.1.m résume les taux d'erreurs rencontrés dans l'exercice des trois métriques que l'on rencontre le plus fréquemment dans la littérature.

Vecteur	Distance	Base	ρ_a	ρ_r	$\frac{1}{2}(\rho_a+\rho_r)$
$\langle c(n)\rangle$	d_2	I	5.1%	4.3%	4.7%
$\langle c(n)\rangle$	d_2	II	39.8%	55.9%	47.9%
$\langle c(n)\rangle$	$d_{2,\text{pond}}$	II	10.4%	21.9%	16.2%
$\langle c(n)\rangle$	$d_{2,\text{Maha}}$	II	6.2%	25.7%	15.9%

Figure 6.1.m Résumé des taux d'erreur

Constatons que le surcroît de difficultés attendu entre la première et la seconde base de données s'est manifesté de façon sensible; tentons ici d'en déterminer les responsabilités. Les coupables potentiels que nous suspectons déjà sont la méthodologie d'établissement des seuils (a priori ou a posteriori), la multiplication des sessions ou encore la robustesse de notre estimation des poids (pour les distances qui en font usage).

6.1.6 Analyse des sources potentielles d'erreur

Tentons de séparer, pour la seconde base de données, les effets individuels de chaque cause potentielle d'erreurs; par exemple, il est possible d'obtenir un taux moyen d'erreur équitable valant seulement 6.6% si nous remplaçons les seuils a posteriori par des seuils a priori et si nous renonçons à introduire d'autres sessions que celles ayant servi à la fois à la construction des vecteurs représentatifs et à l'estimation de la matrice de covariance. Formellement, nous aurions

6.1.3
$$\begin{cases} E_r = \left\{\lambda_j^{(k)} \mid k\in[1,10]\wedge j\in[1,10]\right\} & \\ E_{h,a}^{(k)} = \left\{\lambda_j^{(k)} \mid j\in[11,20]\right\} & \forall k\in[1,10] \\ E_{h,a}^{(k)} = \left\{\lambda_j^{(i)} \mid i\in[1,10]\setminus\{k\}\wedge j\in[11,20]\right\} & \forall k\in[1,10] \\ E_{h,1}^{(k)} = E_{h,a}^{(k)} & \forall k\in[1,10] \\ E_{h,t}^{(k)} = E_{h,a}^{(k)} & \forall k\in[1,10] \end{cases}$$

Dans ces conditions, allégées puisqu'on y aurait remplacé une méthode U par une méthode C, une métrique de Mahalanobis serait nécessaire pour atteindre le résultat annoncé; le taux moyen d'erreur équitable associé à une métrique euclidienne ou euclidienne pondérée y vaudrait respectivement 16.7% et 9.0%.

Si l'on décidait de ne renoncer qu'au caractère a priori des seuils, sans se défaire de la session supplémentaire et des 12 locuteurs qu'elle fait apparaître, alors les taux associés aux métriques euclidiennes, euclidiennes pondérées et

de Mahalanobis deviendraient respectivement 46.6%, 12.1% et 11.1%. Formellement, nous aurions

6.1.4
$$\begin{cases} E_r = \left\{\lambda_j^{(k)} \mid k \in [1,10] \wedge j \in [1,10]\right\} & \\ E_{k,s}^{(k)} = \left\{\lambda_j^{(k)} \mid j \in [21,30]\right\} & \forall k \in [1,10] \\ E_{k,s}^{(k)} = \left\{\lambda_j^{(i)} \mid i \in [11,22] \setminus \{k\} \wedge j \in [21,30]\right\} & \forall k \in [1,10] \\ E_{k,t}^{(k)} = E_{k,s}^{(k)} & \forall k \in [1,10] \\ E_{\bar{k},t}^{(k)} = E_{\bar{k},s}^{(k)} & \forall k \in [1,10] \end{cases}$$

Si nous introduisons $E_w \subseteq E_s$ le sous-ensemble de l'ensemble d'apprentissage qui sert à l'estimation des pondérations, alors les conditions ci-dessus peuvent être abrégées respectivement par

6.1.5
$$\begin{cases} E_s = E_t \subseteq E_w \\ E_s = E_t \not\subset E_w \end{cases}$$

Les différents autres cas à considérer sont

6.1.6
$$\begin{cases} E_s \cap E_t = \varnothing \quad \wedge \quad E_s \subseteq E_w \quad \wedge \quad E_t \subseteq E_w \\ E_s \cap E_t = E_s \cap E_w = \varnothing \quad \wedge \quad E_t \subseteq E_w \\ E_s \cap E_t = E_t \cap E_w = \varnothing \quad \wedge \quad E_s \subseteq E_w \\ E_s \cap E_t = E_s \cap E_w = E_t \cap E_w = \varnothing \end{cases}$$

Profitons ici de rappeler que la façon ordinaire d'exploiter notre seconde base de données correspond au troisième cas de l'expression 6.1.6 qui correspond, comme le quatrième cas, à une méthode U. Les deux premiers cas doivent être considérés comme une méthode C puisque $E_t \subseteq E_w$ et $E_w \subseteq E_s$.

La figure 6.1.n résume les résultats que nous venons d'énoncer et montre que, en d'autres termes, il est évident que l'introduction de la session indépendante amène la plus grande part de la dégradation de la qualité des classificateurs, ce qui innocente partiellement le choix des seuils de décision. Par contre, il est incertain que l'augmentation de la variabilité des locuteurs en soit la cause plus que le défaut de robustesse de l'estimation de la matrice de covariance.

Distance	Sessions	a posteriori	a priori
d_2	1	16.7%	
	> 1	46.6%	47.9%
$d_{2,\text{pond}}$	1	9.0%	
	> 1	12.1%	16.2%
$d_{2,\text{Maha}}$	1	6.6%	
	> 1	11.1%	15.9%

Figure 6.1.n Modifications de l'exploitation de la base de données II

6.2 Erreur moyenne de quantification vectorielle

Dans ce paragraphe, nous allons discuter en profondeur la méthode de l'erreur moyenne de quantification vectorielle, méthode dont nous avions présenté le formalisme aux paragraphes 2.2.2 quant au processus d'établissement du vecteur représentatif et 2.3.4 quant au processus de comparaison. Nous allons appliquer cette méthode à deux vecteurs caractéristiques différents qui sont d'une part la suite des cepstres complexes du filtre de synthèse et d'autre part leur pente temporelle.

6.2.1 Principe

Un locuteur émet de la parole intelligible en produisant dans un ordre judicieux une série de sons stables entrecoupés de transitions bien choisies. Le nombre de sons stables émis est petit; une vision très simplifiée permet de les associer directement aux différents sons perçus comme caractéristiques et interprétés par l'être humain comme, par exemple, relatifs à une certaine voyelle.

Considérons la densité de probabilité de vecteurs caractéristiques. Si l'existence d'états stables est avérée, alors cette densité possède plusieurs modes et chacun d'eux peut servir à la définition d'une classe qui lui soit propre. Cette partition de l'espace des vecteurs caractéristiques peut se faire de diverses manières; l'approximation offerte par des régions de Dirichlet, représentées par un noyau ponctuel, est souvent considérée comme acceptable. En particulier, la classification par nuées dynamiques du paragraphe 2.2.2 est un exemple d'algorithme qui permet de révéler ces classes à l'aide des échantillons typiques de la densité de probabilité que représente l'ensemble X. Rappelons que nous nommons

dictionnaire l'ensemble Y des représentants; il est destiné à permettre l'application de la quantification vectorielle.

Cette dernière autorise le codage d'un ensemble de vecteurs cepstraux en réalisant la substitution des vecteurs originaux par l'étiquette du noyaux des classes auxquelles ils appartiennent; ce processus est une quantification car un univers de nature continue est appliqué sur un ensemble fini de codes discrets. La quantité d'infonnation associée passe donc d'une valeur infinie à une valeur finie et sa diminution se traduit par une erreur de reconstruction. En effet, l'opération de décodage consistant simplement à remplacer les étiquettes par les noyaux, les vecteurs originaux ne sont pas restitués exactement; en fait, tous les vecteurs d'une même classe prennent la valeur du noyau de cette classe après un cycle de codage et de décodage.

L'hypothèse sur laquelle se base la méthode de l'erreur moyenne de quantification vectorielle veut que chaque locuteur possède un registre d'états stables qui se différencie de celui des autres. Par conséquent, un dictionnaire adapté à la reconstitution de la parole d'un locuteur particulier devrait être efficace pour reconstruire toute locution émise par lui-même et inefficace pour reconstruire une locution émise par autrui, l'erreur globale de quantification jouant le rôle de cette mesure d'efficacité.

La figure 6.2.a illustre ce processus. Après avoir récolté suffisamment de parole d'un locuteur de référence (k) et construit son dictionnaire $Y^{(k)}$ (vecteur représentatif), on réalise la quantification vectorielle de la parole d'un locuteur de test (i). Pour ceci, on établit d'abord la suite $\{j\}$ des étiquettes identifiant les noyaux des classes, puis on reconstruit la suite $\left\{\mathbf{y}_j^{(k)}\right\}$ représentant la locution après quantification vectorielle. Un processus de comparaison se charge alors de déterminer une mesure de l'imperfection de reconstruction de chaque vecteur contribuant à l'erreur moyenne de quantification vectorielle $d_{VQ}(X^{(i)}, Y^{(k)})$, où (i) et (k) servent à repérer les locuteurs de test et de référence respectivement. Les détails de classification et des calculs de distance ayant été déjà discutés aux paragraphes 2.2.2 et 2.3.4, nous n'y reviendrons pas ici.

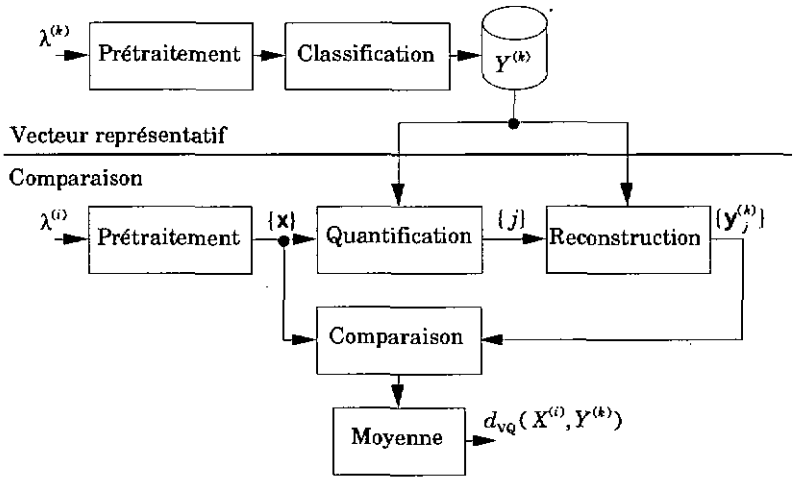

Figure 6.2.a Principe de l'erreur moyenne de quantification vectorielle

■ 6.2.2 Conditions d'analyse

Bien que nos conditions d'établissement des vecteurs caractéristiques cepstraux soient les mêmes que celles rencontrées au paragraphe 6.1.2, nous allons les répéter ici de sorte à pouvoir les comparer avec [88SooF], grâce à la figure 6.2.b.

			[88SooF]	Base I	Base II
Passe-bas	Hz		3200	3400	~3500
Passe-haut	Hz		200	0	0
Echantillonnage	Hz		6670	8000	8000
Résolution	bit		?	12	16
Pauses			exclues	incluses	incluses
Fenêtre		w	Hamming	Bartlett	Bartlett
Support	s		0.045	0.030	0.030
Pas	s		0.015	0.010	0.010
Préaccentuation		μ	0.95	0.95	0.95
Ordre		p	8	14	14
Nombre de classes		K	64	32	32

Figure 6.2.b Conditions d'analyse

La comparaison de nos conditions avec celles des auteurs de l'article [88SooF] fait apparaître peu de différences essentielles. Notons pourtant qu'ils ont porté une attention exclusive aux segments du signal susceptibles de contenir de la parole puisqu'ils ont rejeté les parties considérées comme pause; à cet égard, l'algorithme de segmentation entre pause et parole n'est pas décrit dans l'article cité. Une autre différence importante naît du choix de l'ordre d'analyse, nettement plus petit que celui que nous avons retenu. Enfin, un paramètre est particulièrement pertinent dans le cadre de ce paragraphe: il s'agit du nombre de classes. Nous voyons que notre choix est plus modeste; cependant, les mêmes auteurs ont publié certaines expériences dans l'article [86RosA] qui montrent que le nombre de classes que nous avons retenu reste acceptable.

Il reste encore à préciser les paramètres présidant à la création des dictionnaires et ceux de la mesure de l'erreur de quantification. Quant aux premiers, si nous savons qu'une des nombreuses variantes de l'algorithme des nuées dynamiques est utilisée par les auteurs de [88SooF], pourtant aucun détail de réalisation n'y est explicite. Nous ne connaissons ni le choix des conditions initiales, ni surtout la condition d'arrêt de la classification; la métrique utilisée dans la mesure d'agrégation des classes n'y est pas non plus signalée explicitement. Quant aux secondes, nous les avons déjà décrites au paragraphe 4.3.2.B.

Pour notre part, l'algorithme de classification automatique que nous avons retenu est aussi celui des nuées dynamiques, exposé au paragraphe 2.2.2. Or, si nous savons qu'il converge toujours, cependant nous savons aussi que cette convergence n'est pas absolue: il se peut que le minimum atteint par le critère de l'expression 2.2.12 ne soit que local. Ce fait nous contraint à construire une bonne partition initiale Y_0. Nous avons reporté à l'annexe A.5 les détails de construction de ces partitions initiales pour la base de donnée I et pour la base de données II. Notons ici que l'importance du choix de la partition initiale ne nous a pas paru très grande, pour autant que la convergence soit atteinte.

A) Particularités de la comparaison par quantification vectorielle

Contrairement à [88SooF], qui considère la moyenne des erreurs de quantification en utilisant pour la mesure de distance d de l'expression 2.3.9 le carré d'une distance euclidienne ($d = d_2^2$), quelques expériences préliminaires nous ont persuadé du fait que l'utilisation d'une distance euclidienne est plus raisonnable ($d = d_2$); la mesure de dissimilitude que nous avons adoptée est par conséquent la moyenne des erreurs de quantification.

Nous justifierons ce choix par le fait que le carré d'une mesure euclidienne de distance, éventuellement pondérée, accentue l'importance de quelques vecteurs excentriques et porte préjudice au poids des vecteurs les plus communs. Or, de deux choses l'une: ou bien un locuteur peut être bien caractérisé par des tics de langages, peu fréquents mais spécifiques, les vecteurs excentriques méritant dans ce cas l'importance que leur accorde le carré de la métrique euclidienne, ou bien ces vecteurs excentriques doivent être considérés comme peu représentatifs et anecdotiques. Il semble en pratique que la seconde hypothèse soit la plus juste.

6.2.3 Quantification vectorielle de cepstres complexes et base de données I

Les figures 6.2.c et 6.2.d donnent les matrices de fausses acceptations et de faux rejets dans le cas de la quantification vectorielle de cepstres du filtre de synthèse. On en déduit un taux de fausse acceptation ρ_a valant 2.6% et un taux de faux rejet ρ_r valant 2.3%. Le taux moyen d'erreur équitable vaut donc 2.5%.

$64 \cdot \rho_a$	☺(1)	☺(2)	☺(3)	☺(4)	☺(5)	☺(6)	☺(7)	☺(8)	☺(9)	☺(10)	☺(k)
☺(1)		0	0	0	0	0	0	0	0	0	0
☺(2)	0		11	3	2	0	0	0	0	1	17
☺(3)	0	0		2	0	0	0	0	0	0	2
☺(4)	0	0	1		0	0	0	0	0	0	1
☺(5)	0	3	44	20		0	0	0	0	0	67
☺(6)	0	0	0	0	0		0	0	0	0	0
☺(7)	0	3	4	0	0	0		3	0	15	25
☺(8)	1	1	12	0	0	0	0		0	2	16
☺(9)	0	0	19	2	0	0	0	0		0	21
☺(10)	0	0	3	0	0	0	0	0	0		3
☺(i)	1	7	94	27	2	0	0	3	0	18	152
$\langle \rho_a \rangle$											2.6%

Figure 6.2.c Fausses acceptations

$56 \cdot \rho_r$	☺(1)	☺(2)	☺(3)	☺(4)	☺(5)	☺(6)	☺(7)	☺(8)	☺(9)	☺(10)	☺(k)
	0	0	10	1	0	0	0	0	0	2	13
$\langle \rho_r \rangle$											2.3%

Figure 6.2.d Faux rejets

On constate dans ce cas que plus de la moitié des erreurs est liée aux références du locuteur d'étiquette $\{☺^{(3)}\}$; la méthode de l'erreur moyenne de quantification vectorielle ne paraît donc pas y être propice. Par contre, elle est mieux adaptée au locuteur d'étiquette $\{☺^{(2)}\}$ que ne l'était la méthode du cepstre moyen. Ce fait est une illustration de ce que la notion de facilité de reconnaissance d'un locuteur doit être rendue relative à la méthode employée.

■ 6.2.4 Quantification vectorielle de cepstres complexes différentiels et base de données I

Les figures 6.2.e et 6.2.f donnent les matrices des erreurs de décision obtenues par l'usage de la pente cepstrale définie à l'expression 3.5.1. Nous avons rendu minimal le support de l'estimation de cette pente en imposant $T = 1$. Ce choix correspond à une durée valant 0.050 s, où les deux seuls cepstres qui interviennent possèdent une partie commune de 0.010 s.

$64 \cdot \rho_a$	$☺^{(1)}$	$☺^{(2)}$	$☺^{(3)}$	$☺^{(4)}$	$☺^{(5)}$	$☺^{(6)}$	$☺^{(7)}$	$☺^{(8)}$	$☺^{(9)}$	$☺^{(10)}$	$☺^{(i)}$
$☺^{(1)}$		39	15	50	7	64	25	57	10	13	280
$☺^{(2)}$	0		0	8	0	64	1	27	0	0	100
$☺^{(3)}$	43	48		48	37	64	48	56	34	46	424
$☺^{(4)}$	0	2	0		0	64	0	7	0	0	73
$☺^{(5)}$	35	59	42	59		64	41	62	39	37	438
$☺^{(6)}$	0	0	0	0	0		0	0	0	0	0
$☺^{(7)}$	19	33	17	35	3	64		61	2	14	248
$☺^{(8)}$	0	0	0	0	0	60	0		0	0	60
$☺^{(9)}$	40	53	38	56	40	64	42	61		41	435
$☺^{(10)}$	33	55	36	63	18	64	42	64	20		395
$☺^{(j)}$	170	289	148	319	105	572	199	395	105	151	2453
$\langle \rho_a \rangle$											42.6%

Figure 6.2.e Fausses acceptations

$56 \cdot \rho_r$	$☺^{(1)}$	$☺^{(2)}$	$☺^{(3)}$	$☺^{(4)}$	$☺^{(5)}$	$☺^{(6)}$	$☺^{(7)}$	$☺^{(8)}$	$☺^{(9)}$	$☺^{(10)}$	$☺^{(i)}$
	17	28	14	31	11	56	19	39	10	14	239
$\langle \rho_r \rangle$											42.7%

Figure 6.2.f Faux rejets

On déduit de ces figures un taux de fausse acceptation ρ_a valant 42.6% et un taux de faux rejet ρ_r valant 42.7%. Le taux moyen d'erreur équitable vaut donc 42.6% dans le cas de la quantification vectorielle de pentes cepstrales; nous en concluons que ce vecteur caractéristique ne produit pas de résultats satisfaisants dans les conditions décrites.

6.2.5 Quantification vectorielle de cepstres complexes et base de données II

Nous allons présenter dans ce paragraphe les résultats obtenus sur notre seconde base de données quand nous utilisons la méthode de l'erreur moyenne de quantification vectorielle des cepstres complexes du filtre de synthèse. Nous avons considéré deux mesures de distance, la première étant une métrique euclidienne et la seconde une métrique euclidienne pondérée.

$36 \cdot \rho_a$	XxXx	AnMa	IsMo	Made	MaSa	AnBe	BeZi	DaFi	FrLe	GePo	
AMSa	0	2	0	11	6	0	0	0	0	0	19
ChPa	2	1	1	0	3	0	0	0	0	0	7
ElGo	0	0	5	1	1	0	0	0	0	0	7
FrDC	0	0	0	0	0	0	0	0	0	0	0
MaHu	3	4	3	9	11	0	0	2	0	0	32
RuDr	1	1	0	8	20	0	0	0	0	0	30
FrKl	0	0	0	3	2	0	0	2	0	0	7
JaSE	0	1	0	0	0	0	11	25	0	0	37
JDBa	0	0	0	0	0	0	2	4	0	4	10
JJBe	0	0	0	0	0	0	0	9	0	0	9
LuMé	0	0	0	0	0	3	15	19	0	3	40
RoCa	0	0	0	0	0	0	0	0	0	0	0
	6	9	9	32	43	3	28	61	0	7	198
$\langle \rho_a \rangle$											4.6%

Figure 6.2.g Fausses acceptations

$180 \cdot \rho_r$	XxXx	AnMa	IsMo	Made	MaSa	AnBe	BeZi	DaFi	FrLe	GePo	
	9	13	3	62	14	0	23	13	45	4	186
$\langle \rho_r \rangle$											10.3%

Figure 6.2.h Fausses acceptations

A) Quantification vectorielle en métrique euclidienne
Les figures 6.2.g et 6.2.h montrent les désormais familières matrices de fausses acceptations ou de faux rejets. On en déduit un taux de fausse acceptation ρ_a et un taux de faux rejet ρ_r valant respectivement 4.6% et 10.3%; l'erreur moyenne vaut donc 7.5%.

Ces figures montrent que la méthode de l'erreur moyenne de quantification vectorielle permet une bonne séparation des sexes puisque nous constatons que seulement 7% des erreurs de fausse acceptation concernent des impostures hétérosexuelles tandis que les 93% restants concernent des impostures homosexuelles; quant à la parité des erreurs, le taux d'échec observé sur le domaine féminin est à peu près égal à celui du domaine masculin. A cet égard, nos résultats sont comparables à ceux de [88ChiD] pour une voyelle tenue, dans le cas d'une distance cepstrale.

B) Quantification vectorielle en métrique euclidienne pondérée
Les figures 6.2.i et 6.2.j montrent les matrices des erreurs. On en déduit un taux de fausse acceptation ρ_a et un taux de faux rejet ρ_r valant respectivement 3.3% et 8.2%; l'erreur moyenne vaut donc 5.7% en utilisant la métrique euclidienne pondérée du paragraphe 2.3.2 dans la fonction d'affectation q de l'expression 2.3.9.

Mettons-nous maintenant dans les conditions d'exploitation de notre seconde base de données qui satisfont l'expression 6.1.3 et comparons aux résultats ci-dessus la valeur 3.3% du taux moyen d'erreur équitable obtenu par l'usage de seuils a posteriori sur une session unique. Nous constatons que le taux d'erreur observé sur le domaine hétérogène y est identique si l'on utilise des seuils a priori, ce qui signifie que l'influence sur ρ_a de l'introduction de locuteurs autres que ceux qui ont servi à la construction des vérificateurs est faible. Par contre, le taux d'erreur du domaine homogène s'est accru car la variabilité des locuteurs a augmenté quand nous avons introduit leur troisième session; il s'en est suivi un élargissement du pic associé aux distances de ce domaine. Cette expérience nous permet donc de conclure que dans les conditions d'expérience qui sont les nôtres, et pour la méthode de reconnaissance de locuteurs nommée erreur moyenne de quantification vectorielle, le paramètre influençant le plus les taux d'erreurs est la façon d'utiliser les sessions.

$36 \cdot \rho_a$	XxXx	AnMa	IsMo	Made	MaSa	AnBe	BeZi	DaFi	FrLe	GePo	
AMSa	0	0	0	5	0	0	0	0	0	0	5
ChPa	2	0	1	1	2	0	0	0	0	0	6
ElGo	0	0	4	0	0	0	0	0	0	0	4
FrDC	0	0	1	2	0	0	0	0	0	0	3
MaHu	4	0	5	10	2	0	1	0	0	0	22
RuDr	1	0	0	1	8	0	0	1	0	0	11
FrKl	0	0	0	0	0	0	0	1	0	0	1
JaSE	0	0	0	0	0	0	8	26	0	0	34
JDBa	0	0	0	0	0	1	0	5	0	1	7
JJBe	0	0	0	0	0	0	0	8	0	0	8
LuMé	0	0	0	0	0	6	15	16	0	3	40
RoCa	0	0	0	0	0	0	0	0	0	0	0
	7	0	11	19	12	7	24	57	0	4	141
$\langle \rho_a \rangle$											3.3%

Figure 6.2.i Fausses acceptations

$180 \cdot \rho_r$	XxXx	AnMa	IsMo	Made	MaSa	AnBe	BeZi	DaFi	FrLe	GePo	
	3	18	9	28	12	2	28	28	13	6	147
$\langle \rho_r \rangle$											8.2%

Figure 6.2.j Faux rejets

Nous avions émis au paragraphe 6.1.6 l'hypothèse selon laquelle un défaut de robustesse dans l'estimation du vecteur de pondération était aussi responsable d'une partie des erreurs. Or, nous nous attendons maintenant à pouvoir établir une meilleure estimation de la variance des cepstres instantanés que nous ne pouvions le faire pour les cepstres moyens. En effet, nous disposons de bien plus de données, le facteur multiplicatif étant égal à la quantité de cepstres disponibles dans un fragment, qui vaut 792; il s'ensuit que le rapport de représentativité introduit à la figure 6.1.k vaudrait 158400, comme le montre la figure 6.2.k. Il n'est donc plus possible de prétendre à un défaut de robustesse dans l'acte même d'estimation d'une variance; la variabilité des locuteurs entre les sessions est maintenant seule en cause et explique l'écart observé entre les taux d'erreur.

Distance	poids	données	rapport
d_2	0	2217600	∞
$d_{2,\text{pond}}$	14	2217600	158400
$d_{2,\text{Maha}}$	105	2217600	21120

Figure 6.2.k Représentativité des facteurs de pondération

Bien que la robustesse de la matrice de pondération **W** nécessaire à une distance de Mahalanobis soit maintenant suffisante, nous n'avons pas mené plus avant les expériences qui s'y rapportent. La raison en est un investissement en temps de calcul jugé non rentable face au maigre bénéfice observé sur le seul essai que nous avons conduit, en seuil a posteriori et session unique; le résultat en est un taux moyen d'erreur équitable valant 5.5%. Par comparaison, rappelons que l'usage d'une distance euclidienne pondérée fournit, dans les mêmes conditions, un taux d'erreur ne valant que 3.3%.

6.2.6 Quantification vectorielle de cepstres complexes différentiels et base de données II

Nous allons présenter dans ce paragraphe les résultats obtenus sur notre seconde base de données quand nous utilisons la méthode de l'erreur moyenne de quantification vectorielle des cepstres complexes différentiels du filtre de synthèse. La seule mesure de distance considérée est une métrique euclidienne.

Le support temporel de l'estimation de la pente cepstrale que nous avions choisi au paragraphe 6.2.4 était tel que seule la différence entre deux cepstres intervenait. Tentons maintenant d'améliorer la robustesse de cette estimation par le choix d'un support de plus longue durée; en accord avec les considérations de l'article [88SooF] exposées au paragraphe 4.3.2.B, nous avons imposé $T = 6$, ce qui équivaut à 13 fenêtres et correspond à une durée valant 0.150 s.

Les figures 6.2.l et 6.2.m résument les erreurs de décision résultant de l'utilisation d'une métrique euclidienne pour la méthode de l'erreur moyenne de quantification vectorielle appliquée à des pentes cepstrales. On en déduit un taux de fausse acceptation ρ_a et un taux de faux rejet ρ_r valant respectivement 42.5% et 25.2%; l'erreur moyenne vaut donc 33.8%.

$36 \cdot \rho_a$	XxXx	AnMa	IsMo	Made	MaSa	AnBe	BeZi	DaFi	FrLe	GePo	
AMSa	16	5	11	13	16	0	0	0	0	0	61
ChPa	17	1	24	5	20	0	0	0	0	0	67
ElGo	8	0	18	0	4	0	0	0	0	0	30
FrDC	1	0	3	0	1	0	0	0	0	0	5
MaHu	25	19	15	22	23	9	21	5	4	13	156
RuDr	34	16	36	22	33	7	21	4	10	11	194
FrKl	33	17	35	27	32	12	22	9	19	26	232
JaSE	9	3	6	3	8	2	23	1	6	9	70
JDBa	35	29	35	36	36	36	36	25	34	36	338
JJBe	1	1	3	2	3	4	12	0	2	6	34
LuMé	35	31	34	34	36	34	36	32	32	36	340
RoCa	36	28	31	34	33	31	34	17	31	33	308
	250	150	251	198	245	135	205	93	138	170	1835
$\langle \rho_a \rangle$											42.5%

Figure 6.2.l Fausses acceptations

$180 \cdot \rho_r$	XxXx	AnMa	IsMo	Made	MaSa	AnBe	BeZi	DaFi	FrLe	GePo	
	50	29	73	49	61	15	54	18	80	25	454
$\langle \rho_r \rangle$											25.2%

Figure 6.2.m Faux rejets

Constatons que, malgré notre tentative d'amélioration du support temporel de l'estimation des cepstres complexes différentiels, les résultats ne sont guère meilleurs que ceux que nous avions obtenus sur la première base de données; ils ne sont en tout cas pas assez bons pour nous encourager à chercher une mesure de distance apte à mettre plus en valeur le potentiel de ce vecteur caractéristique pour la reconnaissance de locuteurs indépendante du texte, à supposer que cette mesure existe. A cet égard, l'introduction d'une distorsion bien choisie sur l'axe des fréquences paraît être une voie fructueuse [89XuL1, 89XuL2, 89YongG]. Notons encore, à titre de curiosité, que les références masculines se sont montrées misogynes en rejetant de nombreuses tentatives féminines, alors que l'opposé n'est pas vrai.

6.2.7 Résumé

La figure 6.2.n résume les taux d'erreurs rencontrés. Nous rappelons que le vecteur caractéristique utilisé ici est la série de cepstres complexes à court terme (éventuellement différentiels) des locutions rencontrées dans notre première et dans notre seconde base de données. Les seuils de décision d'une tâche de vérification ont été déterminés comme étant ceux susceptibles de produire un taux d'erreur équitable sur les ensembles d'apprentissage. Il s'ensuit qu'il s'agit de seuils a posteriori pour la base I et a priori pour la base II.

Méthode	Distance	Base	ρ_a	ρ_r	$\frac{1}{2}(\rho_a + \rho_r)$
VQ_c	d_2	I	2.6%	2.3%	2.5%
$VQ_{\Delta c}$	d_2	I	42.6%	42.7%	42.6%
VQ_c	d_2	II	4.6%	10.3%	7.5%
VQ_c	$d_{2,\text{pond}}$	II	3.3%	8.2%	5.7%
$VQ_{\Delta c}$	d_2	II	42.5%	25.2%	33.8%

Figure 6.2.n Résumé des taux d'erreur

Pour la méthode de l'erreur moyenne de quantification vectorielle, nous avons pu mettre en évidence que l'introduction de la session indépendante amène la plus grande part de la dégradation de la qualité des classificateurs, en raison de l'augmentation de la variabilité des locuteurs. Nous avons aussi constaté que cette méthode s'adapte mal à l'usage de la pente cepstrale comme vecteur caractéristique; cet échec ne paraît pas être fondamentalement lié à l'ampleur du support temporel de son estimation.

6.3 Fréquence fondamentale

Nous avons décrit aux paragraphes 3.8 et 3.9 une méthode d'extraction de la fréquence fondamentale moyenne que l'on peut s'attendre à être dépendante du locuteur [90MonA, 90Skv]]. Cependant, certaines raisons pratiques nous conduisent à n'estimer son efficacité pour la reconnaissance de locuteurs indépendante du texte que sur la première base de données; cet inconvénient est mineur car nous savons désormais quelle est l'ordre de grandeur de l'augmentation des taux d'erreurs à laquelle nous devons nous attendre quand nous passons à la base II.

6.3.1 Principe

Les conditions de travail valables pour la méthode de la fréquence fondamentale sont similaires à celles que l'on rencontre ailleurs dans cette thèse: tâche de vérification et exploitation de la base de données selon les procédures décrites au chapitre 5. La seule particularité est liée au choix de la mesure de distance entre deux fréquences fondamentales moyennes: nous avons évalué cette distance simplement par la valeur absolue de leur différence

6.3.1 $$d(F_0^{(i)}, F_0^{(k)}) = \left| F_0^{(k)} - F_0^{(i)} \right|$$

6.3.2 Fréquence fondamentale et base de données I

Les figures 6.3.a et 6.3.b séparent en domaine hétérogène et homogène les matrices des erreurs de décision observées en utilisant comme vecteur caractéristique la fréquence fondamentale moyenne obtenue selon la procédure décrite au paragraphe 3.8.

$64 \cdot \rho_a$	$\odot^{(1)}$	$\odot^{(2)}$	$\odot^{(3)}$	$\odot^{(4)}$	$\odot^{(5)}$	$\odot^{(6)}$	$\odot^{(7)}$	$\odot^{(8)}$	$\odot^{(9)}$	$\odot^{(10)}$	$\odot^{(k)}$
$\odot^{(1)}$		0	0	0	0	0	0	38	2	12	52
$\odot^{(2)}$	0		52	0	1	0	0	0	33	17	103
$\odot^{(3)}$	0	57		0	2	0	0	0	21	5	85
$\odot^{(4)}$	0	0	0		0	0	26	0	0	0	26
$\odot^{(5)}$	0	12	15	18		0	0	0	0	0	45
$\odot^{(6)}$	0	0	0	0	0		0	0	0	0	0
$\odot^{(7)}$	0	0	0	48	0	0		0	0	0	48
$\odot^{(8)}$	64	0	0	0	0	0	0		0	8	72
$\odot^{(9)}$	0	29	14	0	0	0	0	0		49	92
$\odot^{(10)}$	11	7	1	0	0	0	0	4	41		64
$\odot^{(i)}$	75	105	82	66	3	0	26	42	97	91	587
$\langle \rho_a \rangle$											10.2%

Figure 6.3.a Fausses acceptations

$56 \cdot \rho_r$	$\odot^{(1)}$	$\odot^{(2)}$	$\odot^{(3)}$	$\odot^{(4)}$	$\odot^{(5)}$	$\odot^{(6)}$	$\odot^{(7)}$	$\odot^{(8)}$	$\odot^{(9)}$	$\odot^{(10)}$	$\odot^{(k)}$
	8	12	5	8	0	0	2	0	8	9	52
$\langle \rho_r \rangle$											9.3%

Figure 6.3.b Faux rejets

Nous déduisons de ces figures un taux de fausse acceptation ρ_a et un taux de faux rejet ρ_r valant respectivement 10.2% et 9.3%; le taux moyen d'erreur équitable vaut donc 9.7% pour cette méthode.

6.4 Comparaison des méthodes

Les figures 6.1.m et 6.2.n nous permettent de nous faire une idée de la valeur relative des méthodes pour lesquelles nous avons répété les expériences de reconnaissance dans des conditions proches de celles rencontrées dans la littérature. Il en ressort clairement que la méthode du cepstre moyen est moins efficace que ne l'est celle de l'erreur moyenne de quantification vectorielle de cepstres instantanés, qui sort victorieuse de la comparaison. Par contre, des résultats décourageants s'obtiennent quand cette même quantification vectorielle s'applique à la pente cepstrale. La méthode de la fréquence fondamentale moyenne fournit un résultat médiocre sans toutefois être aussi mauvais que le précédent.

L'usage d'une pondération dans le calcul de distances est bienvenu, à condition que la détermination de la pondération soit robuste. Nous observons que ce dernier critère n'est pas respecté par la pondération de Mahalanobis pour des raisons inhérentes au nombre d'échantillons disponibles. Nous observons encore que les échantillons sont en nombre suffisant pour la détermination des valeurs de pondération euclidienne, bien que le nombre de sessions dont ils proviennent ne soit pas encore assez grand pour assurer une robustesse suffisante face à l'introduction de locutions de test issues de sessions inconnues à l'apprentissage.

Enfin, conformément à notre attente, la première base de données s'est montrée plus facile à dompter que la seconde, le facteur prépondérant qui gouverne cette facilité de reconnaissance nous paraissant être plus lié au choix des sessions qu'il ne paraît être lié à l'utilisation de seuils a posteriori, par opposition aux seuils a priori. De ce point de vue, la dégradation s'est fait sentir plus sur la méthode du cepstre moyen que sur celle de l'erreur moyenne de quantification vectorielle.

6.4.1 Comparaison aux résultats de la littérature

La comparaison directe de nos résultats avec ceux de la figure 4.4.a est rendue difficile par des différences méthodologiques et paramétriques. Pour ce qui est

du cepstre moyen, les seuils de décision d'une tâche de vérification ayant été établis a posteriori par les auteurs de l'article [82ShrM], les conditions les plus proches de celles de la littérature sont celles rencontrées sur notre première base de données. Relativement à notre résultat, [82ShrM] réduit le taux d'erreur de 40% environ au prix d'une augmentation de 650% de la durée nécessaire à l'établissement d'un vecteur représentatif, la durée d'apprentissage pouvant être considérée comme encore plus longue puisque la transformation de Karhunen-Loève dont il est fait usage nécessite des données supplémentaires; globalement, l'augmentation de durée d'apprentissage atteint 1000% environ.

Pour ce qui est de la méthode de la quantification vectorielle, les auteurs de l'article [88SooF] dissertant d'une tâche d'identification 1 à n, aucune comparaison directe ne convient. Plutôt que tenter une extrapolation de taux de confusion à partir de taux d'erreur équitable nous préférons citer, pour la même méthode, les résultats d'une tâche de vérification d'un autre article des mêmes auteurs [86RosA]. Dans ce dernier, pour le même nombre de noyaux (K = 32 / 32), une taille très similaire de l'ensemble servant à la classification (1450 / 1584 vecteurs caractéristiques) et moitié de la nôtre quant au test (360 / 792 vecteurs caractéristiques), et pour une méthodologie où une session sert à la construction des références et où une autre est réservée à l'établissement de seuils a posteriori, un taux moyen d'erreur équitable valant 3.1% est observé. Une pondération ayant été utilisée dans le calcul de distances, les auteurs de [86RosA] font preuve d'une diminution de 5% relativement à notre taux d'erreur obtenu dans des conditions semblables (3.3%).

En conclusion, les résultats que nous avons obtenus en répétant sur nos bases de données les expériences classiques de reconnaissance de locuteur corroborent les résultats publiés dans la littérature, à condition de prendre parfois en compte quelques facteurs correctifs liés à l'utilisation de paramètres différents, comme par exemple la durée d'apprentissage.

7 Contributions originales

Nous ne nous contentons pas, dans cette thèse, de seulement reproduire, comparer et vérifier selon une méthodologie identique et sur nos propres bases de données la validité d'expériences classiques. Nous avons aussi l'ambition de présenter deux vecteurs caractéristiques nouveaux que nous espérons aptes à résoudre le problème de la reconnaissance de locuteurs indépendante du texte. Nous nommerons conformité le premier vecteur et cepstre réel du résidu le second. Enfin, une originalité supplémentaire est de considérer que l'identité d'un locuteur se manifeste de façons tellement variées que presque tout ce que l'on peut extraire du signal de parole porte une fraction de ces manifestations; notre but n'est donc pas nécessairement de trouver la représentation élémentaire unique et la métrique associée qui offrent les taux d'échec les plus bas, mais plutôt de trouver des caractéristiques suffisamment différentes les unes des autres pour que leur combinaison soit efficace. Pour décrire ces différences nous parlerons d'indépendance ou d'orthogonalité, sans que ces termes doivent être pris dans leur sens technique précis, respectivement probabiliste ou vectoriel. Finalement, nous montrerons comment utiliser conjointement les méthodes retenues.

7.1 Conformité

Notre première innovation est née d'une des considérations du paragraphe 4.3.2.A qui montrait comment la méthode de l'erreur moyenne de quantification vectorielle peut être mise en échec par une locution de test dont les vecteurs appartiendraient à une distribution de mêmes modes que ceux associés à la référence, mais où ces derniers seraient d'amplitude différente. Combinons cette observation avec l'examen de la figure 6.2.a, où l'on voit que la quantification vectorielle a été exploitée en ne tenant compte que de la valeur du noyau le plus proche et non de son identité, les méthodes de reconnaissance rencontrées ne prenant pas explicitement en compte l'information contenue dans la suite (j) des codes identifiant les noyaux des classes; or, notre hypothèse

d'omniprésence de manifestations de l'identité du locuteur nous pousse à examiner cette source potentielle d'infonmations.

7.1.1 Méthode

Nous avons montré au paragraphe 2.3.5 comment estimer la fréquence d'apparition $h^{(i)}$ de chaque noyau d'une locution $\lambda^{(i)}$ d'un locuteur (i),dans le but de pouvoir la comparer à celle que nous obtenons d'une référence et celle d'un vecteur de test. La décision $\mathbf{D}$ de l'expression 2.5.1 étant basée sur le calcul d'une distance $d(\mathbf{h}^{(k)},\mathbf{h}^{(i)})$ entre un vecteur $\mathbf{h}^{(k)}$ considéré comme un vecteur représentatif de référence, et un autre vecteur $\mathbf{h}^{(i)}$ considéré comme vecteur caractéristique de test, il est maintenant nécessaire de préciser comment mesurer cette distance. Par souci de complétude, citons ici une mesure que nous n'utiliserons pas mais qui est néanmoins importante puisqu'elle donne son nom à la méthode que nous sommes en train d'examiner

7.1.1
$$d(\mathbf{h}^{(k)},\mathbf{h}^{(i)}) = \sum_{j=1}^{K} \frac{\left(h_j^{(i)} - h_j^{(k)}\right)^2}{h_j^{(k)}}$$

La distribution de la mesure de distance selon l'expression 7.1.1 suit approximativement une loi en χ^2 à $K-1$ degrés de liberté, quelles que soient les distributions des vecteurs caractéristiques $\mathbf{h}$; c'est la mesure classique de conformité d'un ensemble d'échantillons à l'égard d'une loi de distribution connue. Malgré cet aspect de classicisme séduisant, nous préférons utiliser une distance euclidienne ou euclidienne pondérée, principalement en raison de performances accrues en reconnaissance de locuteur.

7.1.2 Conditions d'analyse

Le nombre $K = 32$ de classes retenu pour les expériences de reconnaissance de locuteurs par quantification vectorielle réalisées au chapitre précédent est trop petit pour espérer représenter fidèlement la parole d'un locuteur; de ce point de vue, la pertinence de l'information qu'on retirerait de l'examen de la liste $\{j\}$ des codes, si l'on utilisait les mêmes dictionnaires que ceux construits aux annexes A.5.1 ou A.5.2, serait sans doute insuffisante. C'est pourquoi nous avons préféré considérer un ensemble de noyaux de cardinal plus élevé, appelé dictionnaire universel Y.

En outre, pour que la comparaison des vecteurs représentatifs $\mathbf{h}^{(k)}$ et caractéristiques $\mathbf{h}^{(i)}$ construits grâce aux listes de codes $\{j\}$ soit utile, il est nécessaire que

les codes obtenus correspondent à des noyaux identiques. Cette condition est facilement satisfaite si l'on décide d'utiliser le même dictionnaire Y pour tous les locuteurs. Nous avons donc rejeté la solution qui consiste à offrir à chaque locuteur de référence un vocabulaire personnalisé, parce que ce dernier ne permettrait de représenter, par principe, que de façon inadaptée les zones de l'espace U où le locuteur considéré possède une faible densité de probabilité de ses vecteurs caractéristiques intermédiaires $\mathbf{x}$. Un autre désavantage pratique lié à l'utilisation de vocabulaires personnalisés est qu'ils nécessitent le calcul des fréquences d'apparition $\mathbf{h}^{(i,k)}$ des noyaux de façon relative au locuteur de référence (k), tandis que l'usage d'un vocabulaire universel Y permet de calculer ces fréquences d'apparition de façon indépendante et autorise par conséquent la comparaison directe d'un vecteur représentatif $\mathbf{h}^{(k)}$ et d'un vecteur caractéristique $\mathbf{h}^{(i)}$, ceci quels que soient (k) et (i).

Nous allons encore préciser ici deux points permettant l'application de la méthode de la conformité. Le premier point concerne la nature des vecteurs caractéristiques intermédiaires: nous avons décidé de faire tenir ce rôle aux vecteurs complexes cepstraux à court terme du filtre de synthèse, les mêmes que ceux que nous avions utilisés jusqu'à maintenant. Le deuxième point concerne la façon de construire le dictionnaire universel Y.

A) Dictionnaire universel de la base de données I

Pour notre première base de données, le nombre de noyaux retenus est $\mathbf{K} = 256$. La partition initiale Y_0 de la construction du dictionnaire universel de la base I a été fournie par la concaténation de 8 des dictionnaires $\mathbf{Y}^{(i)}$ précédemment construits sur cette même base et choisis au hasard. Les échantillons nécessaires au fonctionnement de l'algorithme des nuées dynamiques, qui est bien entendu celui que nous avons retenu pour la construction de ce dictionnaire, représentent l'intégralité de ceux disponibles dans la base de données et sont au nombre de $122480 = 80 \times 1531$. La condition d'arrêt est encore une fois la convergence de l'algorithme. On constate que la région de Dirichlet associée à chaque noyau couvre en moyenne $480 \cong 122480/256$ échantillons.

B) Dictionnaire universel de la base de données II

Pour notre seconde base de données, le nombre de noyaux retenus est $\mathbf{K} = 128$. La partition initiale Y_0 de la construction du dictionnaire universel de la base II a été fournie par les quelques premiers vecteurs rencontrés parmi les échantillons nécessaires au fonctionnement de l'algorithme des nuées dyna-

miques; ces derniers sont constitués par un choix arbitraire de 3 fragments par locuteur pour chacune des deux sessions considérées. Il s'ensuit que le cardinal de l'ensemble des échantillons vaut $47520 = 2 \times 10 \times 3 \times 792$; la condition d'arrêt est encore une fois la convergence de l'algorithme. La représentativité moyenne d'un noyau vaut donc $370 \cong 47520/128$ pour cette base de données.

		Base I	BaseII
Locutions		80	60
Vecteurs	P	1531	792
Dictionnaire	K	256	128
Représentativité		480	370

Figure 7.1.a Conditions de construction des dictionnaires universels

7.1.3 Conformité de cepstres complexes et base de données I

La première expérience réalisée selon la méthode de la conformité utilise comme vecteurs caractéristiques intermédiaires les cepstres complexes à court terme du filtre de synthèse de l'analyse par prédiction linéaire. La comparaison du vecteur représentatif $\mathbf{h}^{(k)}$ et du vecteur caractéristique $\mathbf{h}^{(i)}$ qui en résultent est ici euclidienne. Les figures 7.1.b et 7.1.c séparent en domaine hétérogène et homogène les matrices des erreurs de décision basée sur le vecteur caractéristique constitué par l'histogramme de l'utilisation des noyaux du dictionnaire universel.

Nous observons un taux de fausse acceptation ρ_a valant 3.1% et un taux de faux rejet ρ_r valant 2.9%; le taux moyen d'erreur équitable vaut donc 3.0%. Ce taux d'erreur se situe entre celui que nous avons obtenu en faisant usage de la méthode classique de l'erreur moyenne quantification vectorielle et celui des cepstres moyens; par conséquent, cette nouvelle méthode fournit des résultats acceptables dans nos conditions d'expériences. Du point de vue de l'orthogonalité, la comparaison des figures 6.1.e, 6.2.c et 7.1.b montre que la méthode de la conformité ressemble plus à celle du cepstre moyen qu'elle ne ressemble à celle de l'erreur moyenne de quantification vectorielle; par exemple, on constate que les deux locuteurs dont les références engendrent le plus d'erreurs sont les mêmes pour les deux premières méthodes.

$64 \cdot \rho_a$	☺(1)	☺(2)	☺(3)	☺(4)	☺(6)	☺(6)	☺(7)	☺(8)	☺(9)	☺(10)	☺(k)
☺(1)		0	0	0	0	0	0	0	0	0	0
☺(2)	0		0	0	12	10	0	0	6	0	28
☺(3)	0	0		0	7	1	0	0	1	1	10
☺(4)	0	0	3		0	17	0	0	3	13	36
☺(5)	0	13	2	0		10	0	0	8	2	35
☺(6)	0	0	0	0	0		0	0	0	0	0
☺(7)	0	8	0	0	0	16		0	1	5	30
☺(8)	0	16	1	0	1	16	0		3	1	38
☺(9)	0	0	0	0	0	0	0	0		0	0
☺(10)	0	0	0	0	0	2	0	0	0		2
☺(i)	0	37	6	0	20	72	0	0	22	22	179
$\langle \rho_a \rangle$											3.1%

Figure 7.1.b Fausses acceptations

$56 \cdot \rho_r$	☺(1)	☺(2)	☺(3)	☺(4)	☺(5)	☺(6)	☺(7)	☺(8)	☺(9)	☺(10)	☺(k)
	0	3	1	0	1	7	0	0	2	2	16
$\langle \rho_r \rangle$											2.9%

Figure 7.1.c Faux rejets

■ 7.1.4 Conformité de cepstres complexes et base de données II

Nous avons voulu encore une fois déterminer l'influence des trois métriques classiques sur l'efficacité de la méthode considérée. Voici donc dans l'ordre les résultats de la conformité de cepstres complexes, obtenus sur notre seconde base de données en métrique euclidienne, euclidienne pondérée et de Mahalanobis.

A) Conformité de cepstres complexes en métrique euclidienne

Les figures 7.1.d et 7.1.e sont l'équivalent des figures 7.1.b et 7.1.c dans le cas de la seconde base de données. Il découle de ces figures que le taux de fausse acceptation ρ_a vaut 9.0% et que celui de faux rejet ρ_r vaut 10.7%, le taux moyen d'erreur valant donc 9.9%.

$36 \cdot \rho_a$	XxXx	AnMa	IsMo	Made	MaSa	AnBe	BeZi	DaFi	FrLe	GePo	
AMSa	7	30	3	14	9	0	0	6	0	0	69
ChPa	13	21	3	4	2	0	0	2	0	0	45
ElGo	5	2	17	11	0	0	0	1	0	0	36
FrDC	2	0	13	19	6	0	0	5	0	0	45
MaHu	3	1	11	3	5	0	0	0	0	0	23
RuDr	0	14	1	8	10	0	0	3	0	0	36
FrKl	0	0	0	11	1	0	0	0	0	0	12
JaSE	0	11	0	0	0	0	16	28	1	0	56
JDBa	0	0	0	0	0	0	1	4	2	6	13
JJBe	0	0	0	0	0	0	7	17	9	5	38
LuMé	0	0	0	0	0	2	7	6	0	0	15
RoCa	0	1	0	0	0	0	0	0	0	0	1
	30	80	48	70	33	2	31	72	12	11	389
$\langle \rho_a \rangle$											9.0%

Figure 7.1.d Fausses acceptations

$180 \cdot \rho_r$	XxXx	AnMa	IsMo	Made	MaSa	AnBe	BeZi	DaFi	FrLe	GePo	
	29	6	4	94	14	13	6	12	0	15	193
$\langle \rho_r \rangle$											10.7%

Figure 7.1.e Faux rejets

Si l'on compare ce résultat à ceux obtenus précédemment, on constate que la méthode de la conformité reste plus avantageuse que celle du cepstre moyen, même si l'on fait usage d'une métrique pondérée pour ce dernier. La parité des sexes n'est cependant plus respectée, car on observe que moins du tiers des erreurs est commis par des références de locuteurs masculins.

B) Conformité de cepstres complexes en métrique euclidienne pondérée

Les figures 7.1.f et 7.1.g présentent les résultats obtenus par l'usage d'une métrique enclidienne pondérée. Il découle de ces figures que le taux de fausse acceptation ρ_a vaut 7.3% et que celui de faux rejet ρ_r vaut 11.6%, le taux moyen d'erreur valant donc 9.4%.

$36 \cdot \rho_a$	XxXx	AnMa	IsMo	Made	MaSa	AnBe	BeZi	DaFi	FrLe	GePo	
AMSa	10	22	3	14	6	0	0	4	0	0	59
ChPa	13	7	3	1	2	0	0	0	0	0	26
ElGo	6	0	20	4	0	0	0	1	0	0	31
FrDC	2	0	12	12	2	0	0	2	0	0	30
MaHu	1	0	8	0	2	0	0	1	0	0	12
RuDr	0	7	2	7	7	0	0	1	0	0	24
FrKl	0	0	0	8	1	0	0	1	0	0	10
JaSE	0	5	0	0	0	0	13	33	3	1	55
JDBa	0	0	0	0	0	0	0	2	5	5	12
JJBe	0	0	0	0	0	1	7	26	8	5	47
LuMé	0	0	0	0	0	0	4	3	0	0	7
RoCa	1	0	0	0	0	0	0	0	0	0	1
	33	41	48	46	20	1	24	74	16	11	314
$\langle \rho_a \rangle$											7.3%

Figure 7.1.f Fausses acceptations

$180 \cdot \rho_r$	XxXx	AnMa	IsMo	Made	MaSa	AnBe	BeZi	DaFi	FrLe	GePo	
	31	10	1	102	20	10	12	7	0	15	208
$\langle \rho_r \rangle$											11.6%

Figure 7.1.g Faux rejets

On constate à nouveau que l'usage d'une pondération favorise un peu les faux rejets et limite les fausses acceptations alors que nous aimerions que les seuils établis lors de la construction des classificateurs soient tels que l'on observe une valeur identique pour les taux ρ_a et ρ_r. De ce point de vue, la représentativité, au sens des figures 6.1.k et 6.2.k, d'un poids de la métrique euclidienne pondérée est révélateur; il vaut ici 200, ce que nous considérons pourtant comme une valeur apte à assurer une robustesse suffisante face aux variations des locuteurs. La responsabilité de l'écart entre les taux ρ_a et ρ_r doit par conséquent être attribuée, ici comme au paragraphe 6.1.6, à l'effet d'une spécialisation trop marquée face aux sessions.

Entre une métrique euclidienne et une métrique euclidienne pondérée, le gain du taux moyen d'erreur est minime; on peut cependant se réjouir du fait que la répartition des erreurs entre locuteurs et locutrices est plus harmonieuse lors-

qu'il est fait usage d'une pondération que lorsque la métrique est simplement euclidienne.

C) Conformité de cepstres complexes en métrique de Mahalanobis

Les figures 7.1.h et 7.1.i présentent les résultats obtenus par l'usage d'une métrique de Mahalanobis. Il découle de ces figures que le taux de fausse acceptation ρ_a vaut 0.3% et que celui de faux rejet ρ_r vaut 60.9%, le taux moyen d'erreur valant donc 30.6%, à supposer que cette dernière valeur ait un sens quand celles qui servent à la calculer sont tellement disparates.

$36 \cdot \rho_a$	XxXx	AnMa	IsMo	Made	MaSa	AnBe	BeZi	DaFi	FrLe	GePo	
AMSa	0	6	0	0	0	0	0	0	0	0	6
ChPa	0	3	0	0	0	0	0	0	0	0	3
ElGo	0	0	1	0	0	0	0	0	0	0	1
FrDC	0	0	1	0	0	0	0	0	0	0	1
MaHu	0	0	0	0	0	0	0	0	0	0	0
RuDr	0	0	0	0	0	0	0	0	0	0	0
FrKl	0	0	0	0	0	0	0	0	0	0	0
JaSE	0	0	0	0	0	0	0	0	0	0	0
JDBa	0	0	0	0	0	0	0	0	0	0	0
JJBe	0	0	0	0	0	0	1	0	0	2	3
LuMé	0	0	0	0	0	0	0	0	0	0	0
RoCa	0	0	0	0	0	0	0	0	0	0	0
	0	9	2	0	0	0	1	0	0	2	14
$\langle \rho_a \rangle$											0.3%

Figure 7.1.h Fausses acceptations

$180 \cdot \rho_r$	XxXx	AnMa	IsMo	Made	MaSa	AnBe	BeZi	DaFi	FrLe	GePo	
	44	86	65	178	148	107	77	123	143	126	1097
$\langle \rho_r \rangle$											60.9%

Figure 7.1.i Faux rejets

L'excédent des faux rejets sur les fausses acceptations prend ici une ampleur telle qu'il devient difficile de comparer l'efficacité de la métrique de Mahalanobis à celle des deux précédentes. La représentativité au sens des figures 6.1.k et 6.2.k d'un poids de la métrique nous permet cette fois d'en mieux

comprendre la raison, puisqu'elle ne vaut plus maintenant que 3, ce qui permet d'affirmer avec certitude que la spécialisation des vérificateurs est excessive et conduit à une mauvaise estimation des seuils de décision, le taux de fausse acceptation devenant manifestement trop petit par rapport au taux de faux rejets.

De sorte à pouvoir tout de même comparer les diverses métriques entre elles sans introduire les effets liés au choix a priori des seuils de décision d'une tâche de vérification, nous avons construit la figure 7.1.j qui donne les taux moyens d'erreur équitable obtenus à l'aide de seuils a posteriori. La première colonne de cette figure correspond aux conditions données à l'expression 6.1.3, et la seconde colonne aux conditions données en 6.1.4.

Distance	Construction	Estimation
d_2	5.8%	7.8%
$d_{2,\text{pond}}$	5.2%	7.4%
$d_{2,\text{Mahs}}$	0.0%	8.0%

Figure 7.1.j Taux moyens d'erreur équitable

La pondération a été déterminée sur les sessions servant à la construction des vérificateurs; nous sommes donc dans les conditions respectivement de la première ligne de l'expression 6.1.5 pour la première colonne et de la seconde ligne pour la seconde colonne. On constate que l'effet de cette pondération est bien plus important précisément sur les sessions dont elle est issue qu'il ne l'est sur les sessions indépendantes d'estimation de la qualité des classificateurs, puisque dans le premier cas on parvient même à éliminer toute erreur alors que pareil succès n'est pas observé dans le second cas.

7.1.5 Résumé des résultats de la conformité de cepstres complexes

Nous avons introduit un vecteur caractéristique nouveau qui exprime la fréquence d'apparition des noyaux d'un dictionnaire universel lors de l'application de la quantification vectorielle; nous avons nommé conformité la méthode de reconnaissance basée sur ce vecteur caractéristique. Nous avons réalisé sur nos deux bases de données les expériences de reconnaissance de locuteurs indépendante du texte en appliquant la conformité aux cepstres complexes à court terme du filtre de synthèse; la figure 7.1.k résume les résultats obtenus.

Distance	Base	ρ_a	ρ_r	$\frac{1}{2}(\rho_a + \rho_r)$
d_2	I	3.1%	2.9%	3.0%
d_2	II	9.0%	10.7%	9.9%
$d_{2,\text{pond}}$	II	7.3%	11.6%	9.4%
$d_{2,\text{Maha}}$	II	0.3%	60.9%	30.6%

Figure 7.1.k Résultats de la méthode de la conformité de cepstres complexes

7.2 Résidu

Notre seconde innovation tente aussi de récupérer, comme la conformité, une partie des informations que n'exploitent pas les méthodes classiques; il s'agit dans ce cas du résidu de la prédiction linéaire. Rappelons que cette dernière analyse un signal de parole pour en extraire à la fois un signal d'excitation et le filtre de synthèse destiné à sa mise en forme. Le résidu est identique à l'excitation à un facteur de gain près, comme nous l'avons vu aux chapitres 3.2 et 3.6.

Thèse: En reconnaissance de locuteurs, de très nombreux auteurs s'appliquent à découvrir la façon la plus efficace d'exploiter le filtre de synthèse, car l'expérience montre qu'il contient une part importante de l'information véhiculée par le signal de parole [76SamM, 77Mar], 81FurS1, 81FurS2, 81ShrM, 82MohN, 82SchR, 82ShrM, 83HunM, 83LiK, 83ShrM, 84KraM, 85GisH, 85SooF, 85Wol], 86BirM1, 86BirM2, 86FurS, 86GisH, 86HigA, 86Nai], 86RosA, 86SooF, 88Att], 88BigB, 88ChiD, 88LiK, 88NodH, 88SooF, 88ZheY, 89NodH, 89XuL1, 89XuL2, 89YonG, 89Zal], 90Eat], 90Jua B, 90Ogl], 90RosA, 90SavM, 91BenY, 91GagD, 91HigA1, 91MatT, 91RosA, 92PakM, 93CheM]. Sa représentation cepstrale complexe, par exemple, est alléchante parce qu'elle permet l'usage d'une simple métrique euclidienne pondérée tout en gardant une performance élevée.

Antithèse: Cependant, rien ne permet d'affirmer que cette seule partie des résultats d'une analyse par prédiction linéaire contient l'intégralité des informations pertinentes. La bonne performance de méthodes qui exploitent le filtre de synthèse n'est pas nécessairement synonyme d'exclusivité.

Synthèse: Le but de ce paragraphe est d'explorer certaines méthodes exploitant le vecteur caractéristique constitué par le cepstre réel du résidu, dans l'idée de l'utiliser plus tard comme complément naturel à une méthode faisant usage du cepstre complexe du filtre de synthèse.

7.2.1 Motivation du choix du cepstre réel du résidu pour la reconnaissance de locuteurs

Le résidu de la prédiction linéaire est une composante qui mérite plus de curiosité de la part des chercheurs qu'elle n'en a eue jusqu'à maintenant, puisque sa seule écoute permet déjà aux auditeurs humains de reconnaître des locuteurs [89FeuT]. L'information qu'il véhicule est en outre fortement liée au mode d'excitation du son, qu'il soit laryngé ou dévoisé; par exemple, le résidu est bien adapté à la recherche de la fréquence fondamentale d'un son laryngé, que nous savons dépendante du locuteur. Il découle de ces considérations que l'espoir de découvrir des éléments caractéristiques du locuteur au sein du résidu n'est pas vain.

Si le résidu répond à notre attente sur ce point, alors le fait que son principe de calcul le rende nécessairement indépendant du filtre de synthèse, jusqu'à l'ordre d'analyse, assure une combinaison fructueuse entre les informations qu'il soutire du signal de parole et celles issues des méthodes basées sur le filtre de synthèse, puisque la redondance entre ces deux composantes est faible.

A) Cepstre réel du résidu de la prédiction linéaire

Un vecteur caractéristique n'est vraiment intéressant que s'il satisfait au moins quelques-uns des critères du paragraphe 3.1, ce à quoi parvient justement le cepstre réel du résidu de la prédiction linéaire. Citons le naturel, car il est disponible en tout temps. Citons encore la facilité de mesure, explicitée au paragraphe 3.6; enfin, la robustesse est une qualité que l'on s'attend à rencontrer chez ce vecteur caractéristique [83HunM], parce que les imperfections des canaux de transmission se reflètent principalement dans le filtre de synthèse qui, par construction, ne montre aucune dépendance linéaire envers le résidu, jusqu'à l'ordre d'analyse. Cette robustesse est très désirable [86GisH, 91RosR]. Il nous reste à voir si pérennité, infaillibilité et individualité sont aussi à compter au nombre de ses qualités. A vrai dire, c'est surtout le dernier point cité qui va retenir notre attention dans la suite de ce paragraphe, car c'est l'individualité qui est le point le plus important pour la reconnaissance de locuteurs.

7.2.2 Trois méthodes de reconnaissance basées sur le cepstre réel du résidu

Nous avons exploré l'efficacité du cepstre réel du résidu de la prédiction linéaire, que nous nommerons désormais simplement résidu, au moyen de trois méthodes différentes de reconnaissance de locuteurs indépendante du texte. Il

s'agit de la méthode du résidu moyen, de celle de l'erreur moyenne de quantification vectorielle, et d'une méthode particulière au résidu nommée proéminence.

Résidu moyen: la technique la plus simple d'exploitation du résidu est de calculer sur la durée d'une locution la moyenne des résidus à court terme, puis de comparer à l'aide d'une métrique euclidienne les moyennes obtenues. On peut bien sûr encore chercher à en pondérer les composantes.

Quantification vectorielle: la méthode de la quantification vectorielle ayant donné de bons résultats si l'on utilise le cepstre complexe du filtre de synthèse comme vecteur caractéristique, on peut tenter d'y substituer le résidu.

Proéminence: si l'on observe une série temporelle de résidus à court terme, on en retire la sensation d'un paysage fixe duquel émergent occasionnellement des pics harmoniques, indicateurs de son laryngé, qui se rétractent dès que survient une émission dévoisée; l'emplacement du pic et de ses harmoniques indique la fréquence fondamentale du locuteur. La figure 3.6.a montre ces deux conditions, où un exemple de son laryngé est fourni par le son [a] et où un exemple de son dévoisé est donné par le son [ʃ]. Il devient alors intéressant de ne considérer que les événements liés à la présence de pics caractéristiques des émissions laryngées du locuteur; nous proposons pour cela de soustraire au résidu sa moyenne temporelle et de n'en considérer que les écarts positifs. Cette approche reste néanmoins plus riche que celle de la simple détection d'une fréquence fondamentale, car toutes les périodes sont simultanément prises en compte d'une part, et d'autre part le signal reste défini en tout temps.

A) Justification de la multiplicité des méthodes retenues

Le cepstre réel du résidu étant un vecteur caractéristique inédit, nous ne pouvons plus profiter des acquis de la littérature comme nous l'avions fait avec le cepstre complexe du filtre de synthèse, pour lequel nous savions que l'usage d'une distance euclidienne ou euclidienne pondérée était favorable à l'obtention de bons résultats de reconnaissance. Au contraire, avec le cepstre réel du résidu, nous devons nous contenter d'émettre des hypothèses sur l'efficacité des mesures de distance avant de les confronter à l'expérience. Insistons sur ce sujet en rappelant l'importance de l'harmonie du couple (*transformation, métrique*) déjà soulignée au paragraphe 3.6.3.

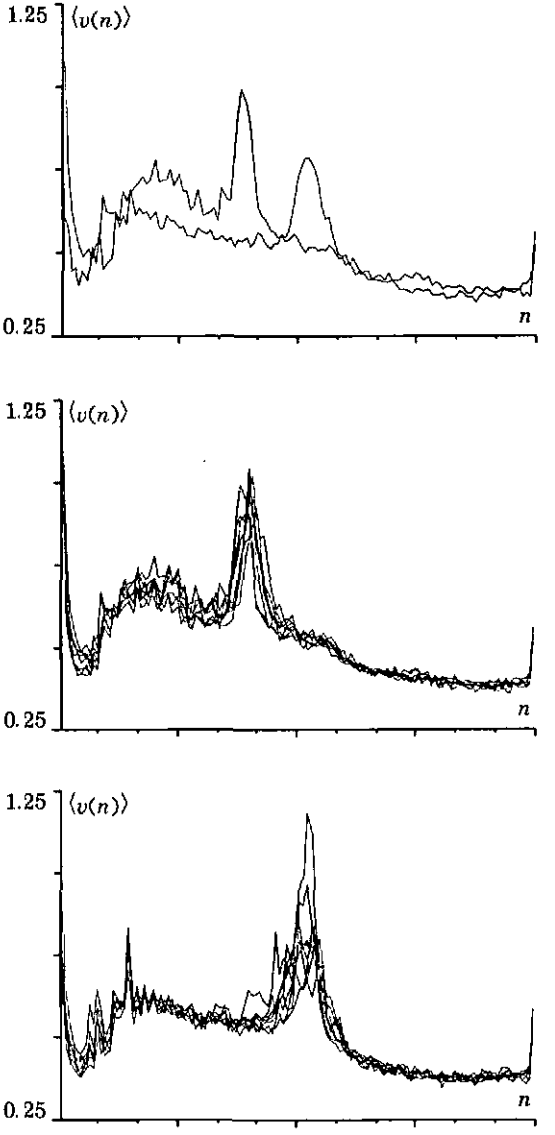

Figure 7.2.a Résidus moyens de deux locuteurs

7.2.3 Principe du résidu moyen

La première partie de la figure 7.2.a montre deux résidus moyens issus respectivement des locuteurs $\{☺^{(1)}\}$ et $\{☺^{(2)}\}$, où l'ordonnée montre la valeur $\langle v(n)\rangle$ atteinte pour la composante n donnée en abscisse. On constate visuellement une différence nette entre ces deux vecteurs caractéristiques; le critère d'individualité du paragraphe 3.1 ne sera toutefois satisfait pleinement que si les locuteurs engendrent des vecteurs qui varient moins que la différence que l'on observe ici.

La deuxième et la troisième partie de la figure 7.2.a permettent de se faire une idée de la pérennité des résidus moyens. L'ensemble de ceux qui ont été produits sur notre première base de données par le locuteur $\{☺^{(1)}\}$, respectivement $\{☺^{(2)}\}$, est reporté sur le même diagramme. On y découvre une stabilité certaine, les résidus moyens s'écartant peu d'un modèle propre à chaque locuteur. Cette assertion, fondée sur des considérations approximatives, nous encourage à entreprendre les expériences quantitatives qui doivent en démontrer la justesse.

7.2.4 Résidu moyen et base de données I

Nous avons testé sur notre première base de données la méthode de reconnaissance de locuteurs indépendante du texte où le processus de comparaison calcule et fournit la distance euclidienne entre deux résidus moyens. La méthodologie d'exploitation de la base de données est bien entendu la même que celle que nous avons appliquée tout au long de cette thèse.

Les figures 7.2.b et 7.2.c donnent les matrices des erreurs de décision qui découlent de l'usage de cette méthode. Nous en déduisons un taux de fausse acceptation ρ_a valant 9.3% et un taux de faux rejet ρ_r valant 9.1%; le taux moyen d'erreur équitable vaut donc 9.2%. Ce résultat est meilleur, quoique de peu, que celui obtenu au paragraphe 6.3.2 pour la méthode de la fréquence fondamentale moyenne, sur la même base de données, ce qui nous rassure pleinement quant à la validité de ce nouveau vecteur caractéristique constitué par le résidu.

$64 \cdot \rho_a$	$\odot^{(1)}$	$\odot^{(2)}$	$\odot^{(3)}$	$\odot^{(4)}$	$\odot^{(5)}$	$\odot^{(6)}$	$\odot^{(7)}$	$\odot^{(7)}$	$\odot^{(8)}$	$\odot^{(10)}$	$\odot^{(i)}$
$\odot^{(1)}$		0	0	0	0	0	0	3	0	8	11
$\odot^{(2)}$	0		35	0	0	0	0	0	10	27	72
$\odot^{(3)}$	0	44		0	0	0	0	0	10	31	85
$\odot^{(4)}$	0	1	4		0	0	21	0	0	6	32
$\odot^{(5)}$	0	33	29	2		0	0	0	4	16	84
$\odot^{(6)}$	0	0	0	0	0		0	0	0	0	0
$\odot^{(7)}$	0	1	1	13	0	0		0	0	9	22
$\odot^{(8)}$	52	0	0	0	0	0	0		1	19	72
$\odot^{(9)}$	0	33	29	0	0	0	0	0		58	120
$\odot^{(10)}$	0	15	10	0	0	0	0	0	13		39
$\odot^{(i)}$	52	127	108	15	0	0	21	3	38	174	537
$\langle \rho_a \rangle$											9.3%

Figure 7.2.b Fausses acceptations

$56 \cdot \rho_r$	$\odot^{(1)}$	$\odot^{(2)}$	$\odot^{(3)}$	$\odot^{(4)}$	$\odot^{(5)}$	$\odot^{(6)}$	$\odot^{(7)}$	$\odot^{(8)}$	$\odot^{(9)}$	$\odot^{(10)}$	$\odot^{(i)}$
	6	12	10	1	0	0	2	0	3	17	51
$\langle \rho_r \rangle$											9.1%

Figure 7.2.c Faux rejets

A) Comparaison du résidu moyen et de la fréquence fondamentale

Si l'on compare la figure 7.2.b à la figure 6.3.a en séparant en deux groupes de même taille les bonnes et les mauvaises références, alors on constate que les meilleurs locuteurs sont les mêmes dans les deux cas; il s'agit plus précisément de $\{\odot^{(4)}, \odot^{(5)}, \odot^{(6)}, \odot^{(7)}, \odot^{(8)}\}$. C'est un indice qui tend à montrer que les deux méthodes en lice non seulement font preuve d'un succès similaire, mais encore livrent des résultats de même nature. En particulier, nous constatons qu'elles sont adaptées l'une et l'autre de façon parfaite à l'unique locutrice $\{\odot^{(6)}\}$ de la base de donnée, puisque les références associées à cette dernière ne se trompent jamais, pas plus que ses vecteurs de test ne conduisent à des décisions erronées.

Les méthodes du résidu moyen et de la fréquence fondamentale ne se complètent pas; il faut donc choisir l'une ou l'autre. Leur efficacité ne permettant pas de les départager, recourrons à des considérations pratiques, ces dernières nous permettant de dégager deux arguments en faveur de la méthode du résidu. Ces

arguments portent sur la facilité de calcul du résidu d'une part et sur son naturel d'autre part.

Quant à celle-là, comme nous l'avons montré au paragraphe 3.8, la détermination d'une fréquence fondamentale fait intervenir de nombreux paramètres qu'il est nécessaire d'ajuster avec art si l'on veut en obtenir de bons résultats. Ceci n'est pas toujours aisé; par exemple, si l'on admet la vue simplificatrice qui assujettit la profondeur de la voix à la taille de l'être humain qui la produit, alors il sera difficile de trouver un jeu de paramètres exploitables à la fois par des fillettes pygmées et par des basketteurs nord-américains.

Quant à celui-ci, la fréquence fondamentale ne satisfait pas complètement le critère du naturel du paragraphe 3.1 puisqu'elle n'est pas disponible en tout temps, en raison du fait qu'elle est définie pour les sons laryngés exclusivement; par contre, le résidu ne souffre d'aucune de ces limitations. Par conséquent, nous considérons ce dernier vecteur caractéristique comme préférable à une fréquence fondamentale sur ce point.

En résumé, la méthode du résidu moyen est apte à se substituer à celle de la fréquence fondamentale moyenne, le gain de la première sur la seconde se trouvant à la fois dans la simplicité de mise en œuvre et dans une disponibilité plus élevée.

■ 7.2.5 Résidu moyen et base de données II

Nous avons testé sur notre seconde base de données deux des trois mesures de distance classiques présentées au paragraphe 2.3. Il s'agit de la distance euclidienne entre deux résidus moyens ainsi que leur distance euclidienne pondérée.

A) Résidu moyen en métrique euclidienne

Les figures 7.2.d et 7.2.e montrent les sempiternelles matrices de fausses acceptations et de faux rejets. On en déduit un taux de fausse acceptation ρ_a et un taux de faux rejet ρ_r valant respectivement 16.3% et 14.8%; l'erreur moyenne vaut donc 15.6%. La métrique y est euclidienne.

$36 \cdot \rho_a$	XxXx	AnMa	IsMo	Made	MaSa	AnBe	BeZi	DaFi	FrLe	GePo	
AMSa	14	11	26	0	2	0	0	0	0	0	53
ChPa	20	34	4	0	33	0	0	0	0	0	91
ElGo	30	25	8	0	24	0	0	0	0	0	87
FrDC	5	15	9	0	23	0	0	0	0	0	52
MaHu	0	1	2	6	18	0	4	0	0	0	31
RuDr	0	0	0	34	2	2	18	23	0	6	85
FrKl	0	0	0	8	0	0	0	35	2	10	55
JaSE	0	0	0	24	0	0	7	36	0	10	77
JDBa	0	0	0	6	0	0	0	15	1	17	39
JJBe	0	0	0	18	0	1	5	17	1	24	66
LuMé	0	0	0	4	0	0	0	31	0	10	45
RoCa	0	0	0	4	0	0	0	21	0	0	25
	69	86	49	104	102	3	34	178	4	77	706
$\langle \rho_a \rangle$											16.3%

Figure 7.2.d Fausses acceptations

$180 \cdot \rho_r$	XxXx	AnMa	IsMo	Made	MaSa	AnBe	BeZi	DaFi	FrLe	GePo	
	43	34	10	11	38	34	28	27	18	24	267
$\langle \rho_r \rangle$											14.8%

Figure 7.2.e Faux rejets

Globalement, on observe une bonne séparation des sexes puisque la fraction des fausses acceptations concernant les confusions hétérosexuelles ne vaut que 17%, tandis que la répartition des autres erreurs du domaine hétérogène est grossièrement équitable. Il est vraisemblable que l'information portée par la fréquence fondamentale, aussi présente dans le résidu, soit responsable de ce comportement, puisqu'il est bien connu que les locutrices possèdent généralement une voix de tessiture plus élevée que celle des locuteurs et qui suffit souvent à les en distinguer. Cependant, la locutrice d'étiquette {Made} ne semble pas se conformer à ce schéma.

Constatons encore que l'augmentation du taux d'erreur due au passage de la base de données *I* à la base *II* est moins importante pour la méthode du résidu moyen qu'elle ne l'a été pour les méthodes du cepstre complexe moyen et de l'erreur moyenne de quantification vectorielle des cepstres complexes à court

terme. Ces taux avaient approximativement triplé quand nous étions passé d'une base à l'autre; le facteur correspondant au cas du résidu moyen est plus petit, ce qui est un fait de nature à encourager l'utilisation du résidu comme vecteur caractéristique dans la reconnaissance de locuteurs. L'explication de cette robustesse accrue relativement à celle observée pour les méthodes usant d'une des formes du filtre de synthèse réside dans l'insensibilité du résidu face à l'influence de certaines modifications des canaux de transmission entre les sessions. Ce peut être aussi la marque d'une pérennité plus grande de ce vecteur caractéristique, assertion qu'il faudrait pourtant confirmer par des expériences représentatives d'intervalles entre les sessions plus longs que ceux dont nous disposons.

B) Résidu moyen en métrique euclidienne pondérée

Les figures 7.2.f et 7.2.g concernent l'application d'une métrique euclidienne pondérée aux résidus moyens. On en déduit un taux de fausse acceptation ρ_a et un taux de faux rejet ρ_r valant respectivement 14.0% et 18.6%; l'erreur moyenne vaut donc 16.3%.

N'oublions pas que la dimension du vecteur caractéristique est environ huit fois plus grande pour le cepstre réel résidu que pour le cepstre complexe du filtre de synthèse; cet état de fait rend encore plus difficile l'estimation de la matrice de covariance nécessaire à la détermination de la pondération. En particulier, si l'on tente d'exploiter l'intégralité de cette matrice en utilisant une métrique de Mahalanobis, alors le rapport de représentativité introduit à la figure 6.1.k ne vaut guère plus que 3; il en résulte un énorme déséquilibre entre les taux de fausse acceptation ρ_a et de faux rejet ρ_r qui vaudraient respectivement 0.4% et 84.6%. Finalement, nous concluons de ces expériences que l'usage d'une pondération n'a pas amené de gain sensible pour la méthode du résidu moyen.

$36 \cdot \rho_a$	XxXx	AnMa	IsMo	Made	MaSa	AnBe	BeZi	DaFi	FrLe	GePo	
AMSa	16	9	31	0	0	0	0	0	0	0	56
ChPa	10	33	23	0	22	0	0	0	0	0	88
ElGo	18	16	18	0	14	0	0	0	0	0	66
FrDC	1	10	16	0	25	0	0	0	0	0	52
MaHu	0	0	0	1	7	0	0	0	0	0	8
RuDr	0	0	0	30	0	0	18	5	0	0	53
FrKl	0	0	0	2	0	0	0	31	2	25	60
JaSE	0	0	0	20	0	0	12	35	13	21	101
JDBa	0	0	0	2	0	0	1	2	2	7	14
JJBe	0	0	0	6	0	0	1	3	10	17	37
LuMé	0	0	0	0	0	0	0	10	3	26	39
RoCa	0	0	0	0	0	0	0	30	0	1	31
	45	68	88	61	68	0	32	116	30	97	605
$\langle \rho_a \rangle$											14.0%

Figure 7.2.f Fausses acceptations

$180 \cdot \rho_r$	XxXx	AnMa	IsMo	Made	MaSa	AnBe	BeZi	DaFi	FrLe	GePo	
	23	24	38	7	42	91	20	8	52	30	335
$\langle \rho_r \rangle$											18.6%

Figure 7.2.g Faux rejets

7.2.6 Erreur moyenne de quantification vectorielle du résidu et base de données I

La méthode de l'erreur moyenne de quantification vectorielle tire son efficacité de la présence, dans une fonction de distribution, de modes caractéristiques d'un individu. Or, nous ne savons pas si la fonction de distribution des cepstres réels du résidu de l'analyse par prédiction linéaire possède ces modes, ni en quelle quantité. Pour tenter d'y voir plus clair, nous avons réalisé une première expérience de reconnaissance de locuteurs indépendante du texte qui utilise cette méthode, dans les conditions ordinaires d'exploitation de notre première base de données.

Le nombre de noyaux retenu est $K = 32$, la construction des dictionnaires est assurée par l'algorithme des nuées dynamiques où les noyaux des dictionnaires initiaux sont donnés par les K premiers vecteurs de chaque référence. Le plus proche voisin est choisi selon un critère de distance euclidienne; les contribu-

tions à l'erreur moyenne correspondent aussi à des distances euclidiennes. Nous donnons aux figures 7.2.h et 7.2.i les traditionnelles matrices d'erreur qui résultent de cette expérience. Nous en déduisons un taux de fausse acceptation ρ_a valant 22.7% et un taux de faux rejet ρ_r valant 23.8%; le taux moyen d'erreur équitable vaut donc 23.2%.

$64 \cdot \rho_a$	☺(1)	☺(2)	☺(3)	☺(4)	☺(6)	☺(6)	☺(7)	☺(8)	☺(9)	☺(10)	☺(k)
☺(1)		0	0	0	0	0	0	0	0	1	1
☺(2)	8		61	0	12	0	0	0	59	36	176
☺(3)	13	11		0	1	0	0	0	10	8	43
☺(4)	31	4	17		7	19	1	5	8	10	102
☺(6)	52	45	64	10		6	0	0	56	48	281
☺(0)	0	0	0	0	0		0	0	0	0	0
☺(7)	64	46	64	64	50	64		60	62	64	538
☺(8)	64	0	0	0	0	1	0		1	19	85
☺(9)	9	0	3	0	0	0	0	0		27	39
☺(10)	16	3	10	0	0	0	0	0	16		45
☺(i)	257	109	219	74	70	90	1	65	212	213	1310
$\langle \rho_a \rangle$											22.7%

Figure 7.2.h Fausses acceptations

$56 \cdot \rho_r$	☺(1)	☺(2)	☺(3)	☺(4)	☺(5)	☺(6)	☺(7)	☺(8)	☺(9)	☺(10)	☺(k)
	24	11	24	8	7	8	0	8	22	21	133
$\langle \rho_r \rangle$											23.8%

Figure 7.2.i Faux rejets

Une explication possible de ce relatif insuccès fait appel à une notion de représentativité dimensionnelle. En effet, si nous imposons $K = 32$, alors le nombre moyen de résidus représentés par un noyau vaut $48 \cong 1531/32$. Or, ces 48 points sont plongés dans un espace à $121 = 240/2 + 1$ composantes, ce qui signifie que la dimension des vecteurs l'emporte largement sur leur nombre alors que le contraire serait souhaitable. Cette observation nous a conduit à mener une exploration systématique de l'effet du nombre de noyaux sur le succès de la reconnaissance.

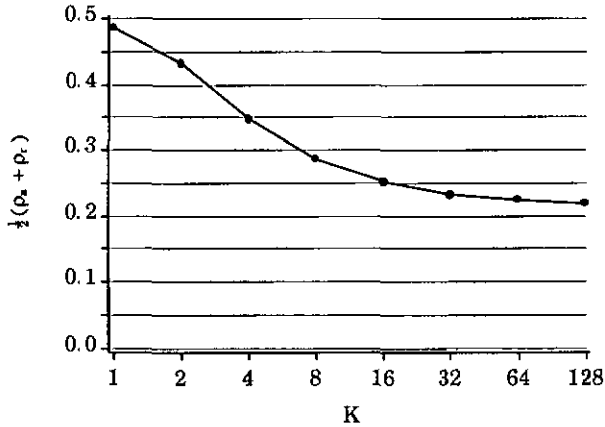

Figure 7.2.j Erreur équitable moyenne en fonction du nombre de noyaux

La figure 7.2.j montre le résultat de cette exploration. Nous avons réalisé 8 expériences où le nombre de noyaux passe de 1 à 128 selon une progression géométrique de facteur 2. Constatons que si nous augmentons la valeur du cardinal de l'ensemble des noyaux, alors l'erreur équitable moyenne décroît mais ne paraît pas avoir la bonne grâce de diminuer au point de franchir une limite de 20%. Il s'ensuit que le mariage du cepstre réel du résidu avec la méthode de l'erreur moyenne de quantification vectorielle n'est pas harmonieux; nous attribuerons la responsabilité de cet insuccès à plusieurs causes probables.

La première explication possible fait appel à un déséquilibre dans la représentativité des noyaux qui permet de supputer une insuffisance de robustesse dans l'estimation de la densité de probabilité des résidus. Nous observons en effet une grande disproportion entre les noyaux puisque le quart seulement d'entre eux code plus de la moitié d'un signal de parole, tandis que moins du quart de ce signal consomme toute une moitié des noyaux. Il s'ensuit que certains d'entre eux représentent une quantité trop faible d'échantillons de la densité de probabilité pour être considérés comme robustes; ceci implique le fait que d'autres sont surexploités, d'où le déséquilibre annoncé.

La deuxième cause probable incrimine la métrique euclidienne utilisée. En effet, nous avons décrit la suite des cepstres réels à court terme du résidu comme un paysage duquel émergent occasionnellement quelques pics. Or, ces pics sont

étroits; l'influence de leur présence, en métrique euclidienne, n'est pas toujours assez sensible par rapport à celle des perturbations intempestives, plus petites, mais distribuées sur plus de composantes, que l'on observe pour tout le reste du paysage.

Enfin, la troisième explication est plus fondamentale, car elle met en doute l'hypothèse même dont seule la satisfaction autorise une mise en œuvre efficace de la méthode de l'erreur moyenne de quantification vectorielle: rien ne pennet de certifier l'existence de plusieurs modes dans la densité de probabilité des résidus. S'il est vrai que les langues à tons peuvent en faire apparaître, la langue française n'en fait pas usage et le domaine des fréquences fondamentales n'y est pas quantifié, au contraire de la musique occidentale traditionnelle qui fait plus volontiers usage de notes de hauteur stable que de notes de hauteur glissante. Les résultats que nous venons d'obtenir en reconnaissance de locuteurs, s'ils ne permettent pas de confirmer explicitement cette hypothèse, en tout cas ne l'infirment pas.

Nous concluons ces expériences par un constat d'échec, car la méthode de l'erreur moyenne de quantification vectorielle du résidu s'est montrée décevante en comparaison des résultats déjà obtenus par la méthode du résidu moyen. Il s'ensuit que nous ne persévérerons pas dans l'examen de cette méthode; nous ne la testerons donc pas sur notre seconde base de données.

7.2.7 Justification du choix de la méthode de la proéminence pour les cepstres réels du résidu

Le résidu nous place dans la situation de l'explorateur en terre inconnue. Nous n'avons en effet pas connaissance a priori de la méthode de comparaison permettant de mesurer l'écart entre deux résidus de sorte à mettre le plus en évidence la part qui est liée à l'identité du locuteur; ce n'est pourtant qu'en découvrant une méthode de mesure efficace que nous pourrons prouver l'existence de cette part. La méthode de comparaison que nous utilisons dans ce paragraphe est une de nos nombreuses tentatives, plus ou moins fructueuses, d'améliorer le résultat déjà obtenu par la simple distance euclidienne entre deux résidus moyens. Il s'agit de la proéminence, décrite en détails au paragraphe 2.3.6.

Comme expliqué au paragraphe 7.2.2, nous avons nommé cette méthode proéminence car nous ne conservons du résidu que les parties saillantes. Cette ap-

proche se base sur le fonctionnement du système auditif humain qui perçoit bien plus volontiers la présence d'une raie spectrale que son absence, en raison du phénomène de masquage [86RosM]; elle profite en outre du fait que la distribution des valeurs prises par les résidus à court terme, composante par composante, ne sont pas symétriques par rapport à leur moyenne.

Nous justifierons cette dernière affirmation en arguant d'une part que, par construction, le cepstre réel du résidu est positif. Il s'ensuit que ses valeurs les plus petites sont nécessairement bornées alors que ses valeurs les plus grandes ne sont limitées que par les conditions de normalisation entre l'excitation et le résidu, ce qui conduit à un déséquilibre inévitable, inhérent à la représentation même du signal. D'autre part, le modèle de production de parole présenté à la figure 3.2.b est assez proche de la réalité; on peut déduire cette affirmation de l'examen de la figure 3.6.a où l'on voit nettement soit une seule période émerger, soit aucune. Il s'ensuit que la présence d'une fréquence fondamentale se traduit par une apparition fugace d'une valeur élevée pour l'une des composantes du résidu, qui y disparaît dès que la fréquence fondamentale varie ou dès que la parole est dévoisée. Ainsi, la distribution des valeurs de cette composante est un magma proche de la moyenne, complété par les quelques îlots de valeurs plus élevées correspondant à des sons laryngés.

■ 7.2.8 Proéminence des cepstres réels du résidu et base de données II

Les figures 7.2.k et 7.2.l montrent la validité de la méthode de la proéminence quand elle s'applique aux cepstres réels du résidu de l'analyse par prédiction linéaire. On en déduit un taux de fausse acceptation ρ_a et un taux de faux rejet ρ_r valant respectivement 11.4% et 14.3%; l'erreur moyenne vaut donc 12.9%. C'est donc à l'aune de la proéminence que nous devons le meilleur résultat que nous ayons pu obtenir en utilisant le cepstre réel du résidu comme vecteur caractéristique, sur notre seconde base de données.

$36 \cdot \rho_a$	XxXx	AnMa	IsMo	Made	MaSa	AnBe	BeZi	DaFi	FrLe	GePo	
AMSa	12	6	25	0	0	0	0	0	0	0	43
ChPa	6	31	18	0	19	0	0	0	0	0	74
ElGo	13	12	17	0	13	0	0	0	0	0	55
FrDC	1	6	11	0	27	0	0	0	0	0	45
MaHu	0	0	0	6	10	0	2	0	0	0	18
RuDr	0	0	0	31	0	0	15	9	0	1	56
FrKl	0	0	0	4	0	0	0	22	0	1	27
JaSE	0	0	0	26	0	0	13	36	0	10	85
JDBa	0	0	0	3	0	0	1	5	1	14	24
JJBe	0	0	0	8	0	0	3	2	0	19	32
LuMé	0	0	0	0	0	0	0	11	0	8	19
RoCa	0	0	0	0	0	0	0	15	0	0	15
	32	55	71	78	69	0	34	100	1	53	493
$\langle \rho_a \rangle$											11.4%

Figure 7.2.k Fausses acceptations

$180 \cdot \rho_r$	XxXx	AnMa	IsMo	Made	MaSa	AnBe	BeZi	DaFi	FrLe	GePo	
	5	18	21	10	39	88	25	18	23	11	258
$\langle \rho_r \rangle$											14.3%

Figure 7.2.l Faux rejets

7.2.9 Récapitulation des méthodes faisant usage de cepstres réels du résidu

Nous avons introduit un vecteur caractéristique nouveau dans le but de recouvrer les parcelles d'information touchant à l'identité d'un locuteur qui ne sont pas présentes dans la suite des filtres de synthèse à court terme que l'on tire d'une analyse par prédiction linéaire. Ce vecteur caractéristique est le cepstre réel du résidu. Nous avons réalisé sur nos deux bases de données certaines expériences de reconnaissance de locuteurs indépendante du texte; ces expériences montrent que le résidu peut résoudre partiellement ce problème. La figure 7.2.m résume les résultats obtenus.

Méthode	Distance	Base	ρ_e	ρ_r	$\frac{1}{2}(\rho_e + \rho_r)$
$\langle v(n) \rangle$	d_2	I	9.3%	9.1%	9.2%
$\langle v(n) \rangle$	d_2	II	16.3%	14.8%	15.6%
$\langle v(n) \rangle$	$d_{2,\text{pond}}$	II	14.0%	18.6%	16.3%
$\langle v(n) \rangle$	$d_{2,\text{Maha}}$	II	0.4%	84.6%	42.5%
VQ_v	d_2	I	22.7%	23.8%	23.2%
Pro	d_{Pro}	II	11.4%	14.3%	12.9%

Figure 7.2.m Résultats des expériences impliquant le résidu

Nous en concluons que, parmi toutes les méthodes exploitant le résidu que nous avons explorées, celle de la proéminence est la plus efficace que nous ayons trouvée, suivie par celle du résidu moyen. Bien que le fait de passer de la méthode de la proéminence à celle des résidus moyens signifie un gain de simplicité, pourtant l'examen de la figure 7.2.m montre que ce gain est entamé par l'augmentation du taux d'erreur, qui croît relativement de 20% environ.

7.3 Choix de méthodes à combiner

La recherche de l'identité d'un locuteur dans un signal de parole est très différente de celle d'une aiguille dans une botte de foin. Dans ce dernier cas, l'aiguille est unique, compacte, et la difficulté vient du fait qu'elle est masquée par de nombreux objets qui lui sont semblables. Dans le premier cas au contraire, l'identité d'un locuteur est une information diffuse, répartie sur de nombreux aspects du signal de parole. La difficulté n'est pas seulement de trouver dans chaque aspect l'information pertinente, quoique partielle, mais encore d'assembler chaque apport de sorte à construire un tout cohérent.

Jusqu'à maintenant, nous nous sommes efforcé de trouver des méthodes qui soient séparément efficaces. Nous avons aussi guidé notre recherche par le souci de choisir des approches qui soient a priori aussi orthogonales que possibles tout en se fondant seulement sur l'analyse par prédiction linéaire, reconnue comme un modèle convenable de l'émission de sons par l'être humain. Ainsi, nous avons montré comment exploiter la part du signal non modélisée par le filtre de synthèse et nommée résidu; nous avons encore montré comment compléter la méthode de l'erreur moyenne de quantification vectorielle par celle de la conformité. Il est donc temps maintenant de combiner ces méthodes;

l'objet de ce paragraphe est de montrer comment choisir les méthodes à combiner, et comment réaliser cette combinaison.

7.3.1 Critère objectif de performance

Le premier point à examiner concerne le choix des méthodes que nous allons combiner. Il serait en effet faux de croire que la multiplication débridée des façons d'extraire les parcelles d'information pertinente noyées dans un flot bruité favorise nécessairement la qualité de la reconnaissance [91WeiS]. Il est sage au contraire de ne considérer qu'un nombre restreint de méthodes. Dans ce paragraphe, nous allons tenter de faire un choix raisonné parmi celles que nous avons rencontrées jusqu'à maintenant, le critère de choix retenu pour l'instant se basant uniquement sur les performance individuelles des méthodes.

Nous disposons globalement de 6 méthodes de reconnaissance de locuteurs indépendante du texte et de quelques variantes se distinguant par l'usage ou l'ignorance d'une pondération. Toutes ces méthodes peuvent être comparées quant au résultat qu'elles fournissent dans les conditions d'expérience de notre seconde base de données. Il s'agit des méthodes du cepstre complexe moyen du filtre de synthèse en distance euclidienne pondérée ($\langle c(n) \rangle$), de l'erreur moyenne de quantification vectorielle du cepstre complexe du filtre de synthèse en distance euclidienne pondérée (VQ_c), de l'erreur moyenne de quantification vectorielle du cepstre complexe différentiel du filtre de synthèse en distance euclidienne ($VQ_{\Delta c}$), de la conformité du cepstre complexe du filtre de synthèse en distance euclidienne pondérée (Conf), du cepstre réel moyen du résidu en distance euclidienne ($\langle v(n) \rangle$), et de la proéminence du cepstre réel du résidu (Pro).

Nous ne disposons malheureusement pas des expériences nécessaires à l'inclusion dans cette comparaison de la méthode de l'erreur moyenne de quantification vectorielle du cepstre réel résidu (VQ_v) ou de celle de la fréquence fondamentale moyenne (F_0). Nous ne déplorons que modérément l'absence de la première, le paragraphe 7.2.6 ayant montré le peu d'intérêt de cette méthode; par contre, l'absence de la seconde est plus regrettable. La raison en est d'une part certaines difficultés d'ordre technique (place disponible sur notre système informatique) et d'autre part des contraintes de calendrier qui nous ont fait renoncer au calcul de la fréquence fondamentale des locutions de notre seconde base de données, ce calcul étant particulièrement coûteux en temps en raison de l'opération de filtrage mentionnée au paragraphe 3.8.4.

La figure 7.3.a classe selon un critère de taux d'erreur moyen croissant les 6 méthodes susceptibles d'être comparées. Les valeurs exactes de ces taux peuvent être retrouvées aux figures 6.1.m ($\langle c(n)\rangle$, 16.2%), 6.2.n ($(VQ_c, VQ_{\Delta c})$, (5.7%,33.8%)), 7.1.k (Conf, 9.4%) et 7.2.m ($(\text{Pro}, \langle v(n)\rangle)$, (12.9%,15.6%)).

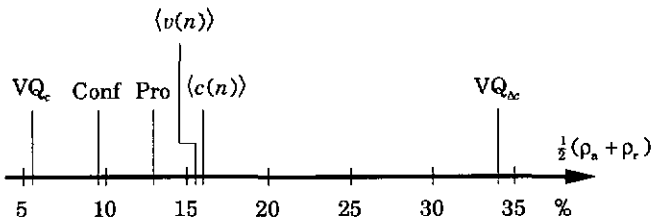

Figure 7.3.a Performance des méthodes

Nous tirons deux conclusions de cette figure. La première porte sur notre incapacité à reproduire les résultats de la référence [86SooF] pour la méthode $VQ_{\Delta c}$, le taux d'erreur de vérification que nous obtenons nous paraissant bien trop élevé eu égard au taux d'erreur d'identification ($\rho_c = 17.4\%$) annoncé par les auteurs de la référence citée. La seconde conclusion est plus réjouissante; elle met en lumière le fait que nos nouvelles méthodes de reconnaissance Conf et Pro se montrent toutes deux plus efficaces que la méthode classique $\langle c(n)\rangle$, même si cette efficacité n'atteint pas celle de la méthode VQ_c. En résumé, le critère de la performance des méthodes nous conduit clairement à rejeter $VQ_{\Delta c}$ et à conserver VQ_c, cette dernière méthode devant être combinée nécessairement avec une de nos deux nouvelles méthodes Conf et Pro.

■ 7.3.2 Critère subjectif de dépendance des méthodes

Le sujet de ce paragraphe concerne la dépendance des méthodes entre elles. Ce critère intervient car la combinaison de plusieurs méthodes ne saurait être fructueuse que lorsque chaque méthode éclaire le problème de la reconnaissance sous un angle particulier; si toutes extraient des données la même information, alors le gain de la combinaison est nul. Cherchons donc maintenant à préciser le degré de dépendance des méthodes candidates; l'approche que nous allons suivre d'abord est subjective et consiste à illustrer la dépendance de certaines paires de méthodes par un diagramme de dispersion.

La motivation du choix des paires de méthodes considère trois aspects. Le premier consiste à séparer les méthodes en deux groupes représentatifs de la

provenance des vecteurs caractéristiques utilisés; nous retrouvons donc l'ensemble $\{VQ_c, Conf, \langle c(n)\rangle, VQ_{\Delta c}\}$ pour le filtre de synthèse de l'analyse par prédiction linéaire et l'ensemble $\{Pro, \langle v(n)\rangle\}$ pour son résidu. Le deuxième aspect prend en compte les conclusions du paragraphe 7.3.1 qui excluent la méthode $VQ_{\Delta c}$ et exigent l'usage de la méthode VQ_c. Le troisième aspect, enfin, sera introduit plus tard. Ces considérations conduisent à la sélection de cinq paires de méthodes donnée à la figure 7.3.b.

ρ_{xy}	VQ_c	Conf	Pro	$\langle v(n)\rangle$	$\langle c(n)\rangle$	$VQ_{\Delta c}$
VQ_c	–					
Conf	C)	–				
Pro	E)	D)	–			
$\langle v(n)\rangle$	–	–	B)	–		
$\langle c(n)\rangle$	–	A)	–	–	–	
$VQ_{\Delta c}$	–	–	–	–	–	–

Figure 7.3.b Sélection des paires de méthodes

A) Conformité de cepstres complexes et cepstre complexe moyen

Les expériences relatives à la méthode de la conformité montrent qu'il est possible d'en obtenir un taux de succès supérieur à celui qu'offre l'usage du cepstre moyen comme vecteur caractéristique; cependant l'examen détaillé des matrices de fausses acceptations et de faux rejets indique en outre que les erreurs commises par les deux méthodes citées sont de même nature. Pour notre seconde base de données, visualisons cette corrélation par le diagramme de dispersion de la figure 7.3.c qui montre en abscisse les distances en métrique euclidienne pondérée obtenues par l'usage de la méthode de la conformité (Conf), et en ordonnée les distances en métrique euclidienne pondérée obtenues par l'usage de la méthode du cepstre complexe moyen ($\langle c(n)\rangle$). Dans ces figures, toutes les distances ont été centrées par soustraction des seuils de décision adéquats, ce qui explique l'apparition de distances négatives. Les contributions du domaine homogène, respectivement hétérogène, y sont présentées de façon séparée d'abord, puis confondue; dans ce dernier cas, le lecteur attentif pourra toutefois encore les distinguer au fait que les premières se présentent par un point de taille plus grosse que les secondes.

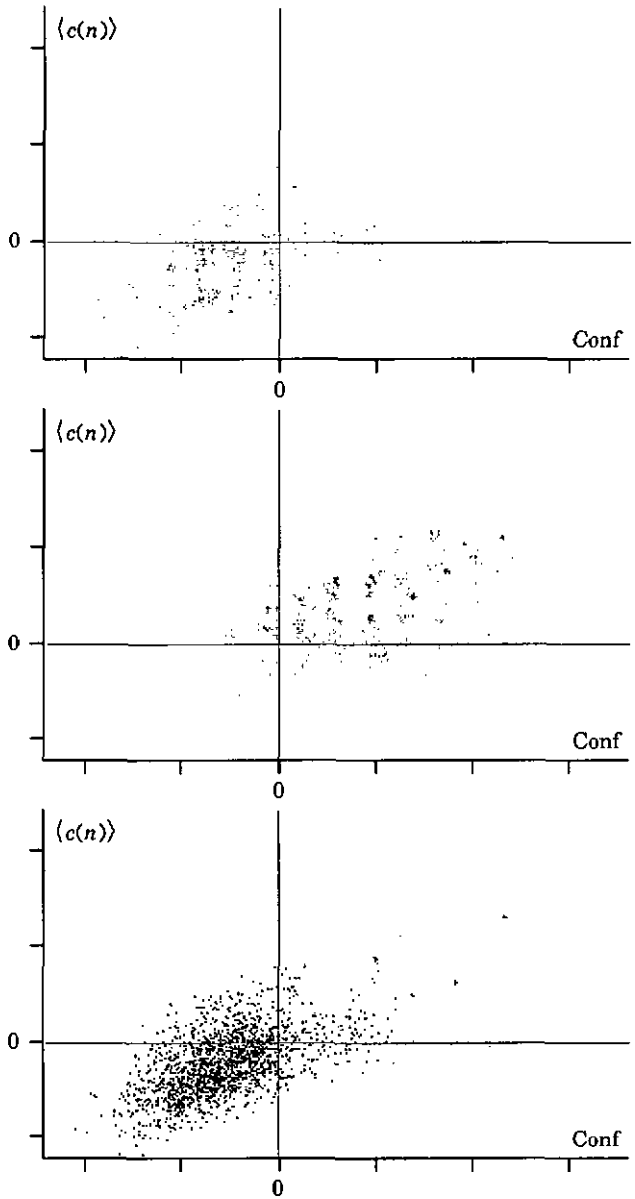

Figure 7.3.c Diagramme de dispersion conformité-cepstre moyen

Nous observons sur le diagramme de la figure 7.3.c que les domaines homogènes et hétérogènes se présentent sous la forme de nuages de points dont l'imbrication est de nature à ne laisser que peu d'espoir à celui qui se donnerait pour mission la recherche d'une frontière de décision sensiblement meilleure, en terme de taux de reconnaissance, que les frontières parallèles aux axes correspondant à l'usage séparé de l'une ou de l'autre méthode. Ce diagramme met donc en évidence le fait que la méthode de la conformité des cepstres complexes du filtre de synthèse se combine mal avec la méthode de leur moyenne.

Nous en concluons que notre méthode inédite de la conformité est susceptible de se substituer à celle du cepstre moyen tout en offrant de meilleures performances et sans changer la nature des informations délivrées. Cette substitution complète la discussion de la sélection des méthodes à combiner et en explicite son troisième aspect, puisque maintenant la raison du rejet de la méthode du cepstre complexe moyen pour des comparaisons ultérieures est éclaircie.

B) Proéminence de cepstres réels et cepstre réel moyen

Les diagrammes de dispersion de la figure 7.3.d permettent une évaluation visuelle du degré de dépendance entre la méthode de la proéminence des cepstres réels du résidu et celle du cepstre réel moyen du résidu. On y trouve en abscisse les distances issues d'une comparaison par proéminence Pro, et en ordonnée les distances obtenues par comparaison selon une métrique euclidienne de cepstres réels moyens du résidu $\langle v(n) \rangle$. Ces diagrammes sont présentés dans les mêmes conditions que celles décrites pour la figure 7.3.c.

Nous constatons que la corrélation entre les mesures issues des deux méthodes est très importante, le diagramme se présentant indiscutablement comme une bande étroite. Le fait que les grandes distances paraissent moins corrélées que les petites nous importe peu, car de toute façon les méthodes prennent dans ce cas une décision correcte. Il s'ensuit que la combinaison de la méthode de la proéminence avec celle du résidu moyen est peu fructueuse; il est donc judicieux de rejeter l'une d'elles. Le critère le plus pertinent dans la comparaison de ces méthodes devient alors celui de la performance. Il conduit à ne conserver que la proéminence dans la suite des comparaisons de ce paragraphe.

L'ensemble des méthodes qui restent en lice contient donc encore les trois éléments $\{VQ_c, Conf, Pro\}$. Nous allons continuer ce paragraphe par un examen systématique de leurs dépendances mutuelles.

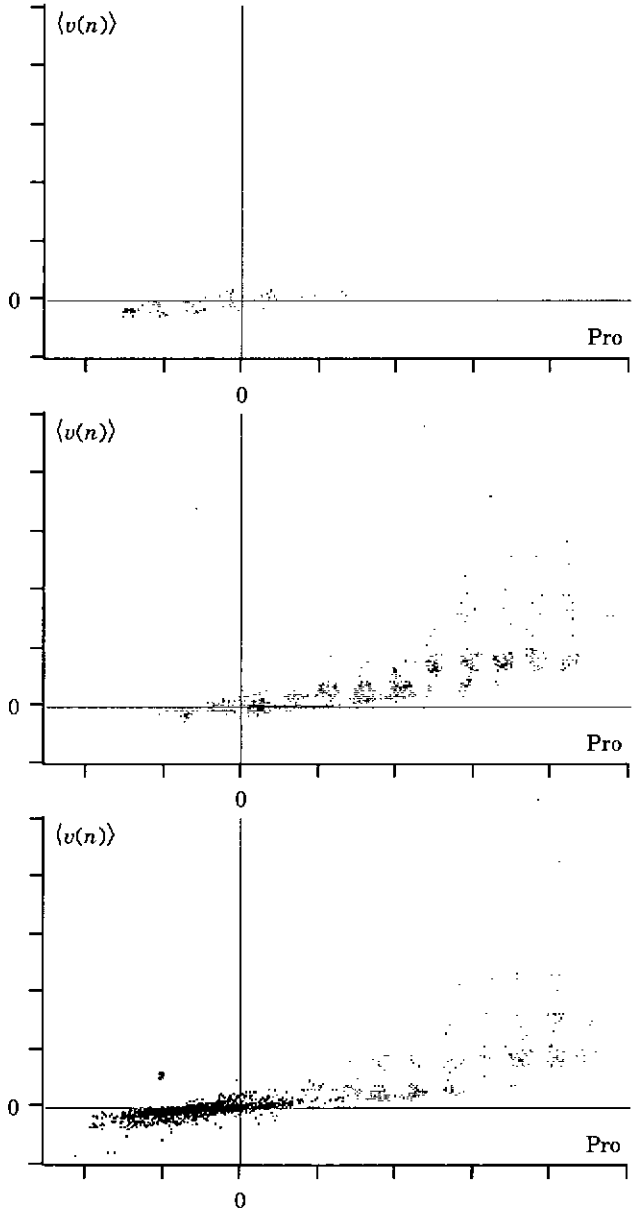

Figure 7.3.d Diagrammes de dispersion proéminence-résidu moyen

C) Quantification vectorielle et conformité

La figure 7.3.e montre les diagrammes de dispersion où l'abscisse porte les distances en métrique euclidienne pondérée obtenues par la méthode de l'erreur moyenne de quantification vectorielle des cepstres complexes du filtre de synthèse; l'ordonnée correspond à la méthode de la conformité des cepstres réels du résidu en métrique euclidienne pondérée. Si l'on compare ces diagrammes à ceux de la figure 7.3.c, on constate que les orthogonalités de chacune des paires de méthodes ne sont pas très différentes; de ce point de vue, la différence d'éclairage des cepstres du filtre de synthèse que l'on espérait voir se manifester entre les méthodes du cepstre moyen, de la conformité, et de l'erreur moyenne de quantification vectorielle, existe mais est peu marquée. De ces trois méthodes, nous conserverons toutefois les deux plus efficaces.

D) Conformité et proéminence

La figure 7.3.f présente les diagrammes de dispersion entre la conformité des cepstres complexes du filtre de synthèse en abscisse, et la proéminence des cepstres réels du résidu en ordonnée. Ces diagrammes sont réjouissants parce que l'on voit aisément que ces deux méthodes font appel à des caractéristiques de nature différente; un indice permettant de confirmer cette affirmation est le fait que l'on trouve des distances simultanément petites pour une méthode et grandes pour l'autre, ou inversement. Ce qui est bien plus important, c'est que l'on puisse concevoir une frontière non parallèle aux axes qui sépare mieux les domaines homogènes et hétérogènes que ne le font les frontières basées exclusivement sur l'une ou sur l'autre des méthodes; à titre de curiosité, on peut d'ailleurs remarquer le fait que, visuellement, l'origine des distances (0,0) paraît devoir se trouver relativement distante de la meilleure frontière rectiligne possible. Nous concluons de ces diagrammes que la combinaison des méthodes concernées aide au processus de reconnaissance.

E) Erreur moyenne de quantification vectorielle et proéminence

La figure 7.3.g est similaire à la précédente; on y trouve la méthode de l'erreur moyenne de quantification vectorielle des cepstres complexes du filtre de synthèse en abscisse et celle de la proéminence des cepstres réels du résidu en ordonnée. Les mêmes conclusions que pour la figure 7.3.f s'appliquent ici à la conformité et à la quantification vectorielle, deux méthodes que nous décidons définitivement de conserver.

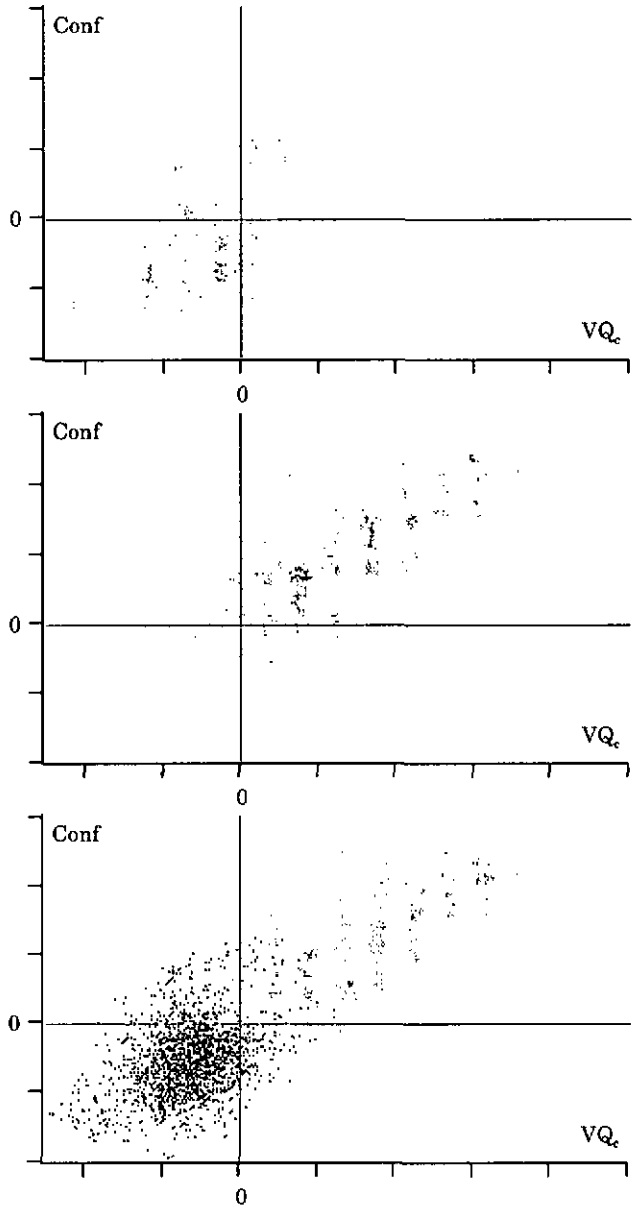

Figure 7.3.e Diagrammes de dispersion quantification vectorielle-conformité

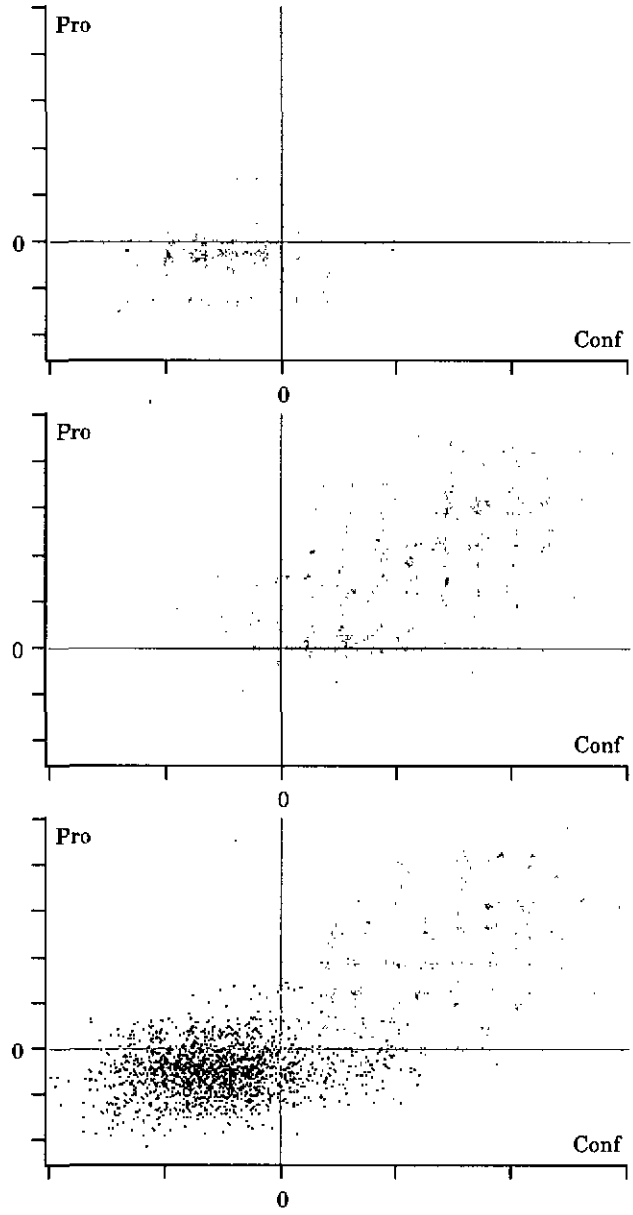

Figure 7.3.f Diagrammes de dispersion conformité-proéminence

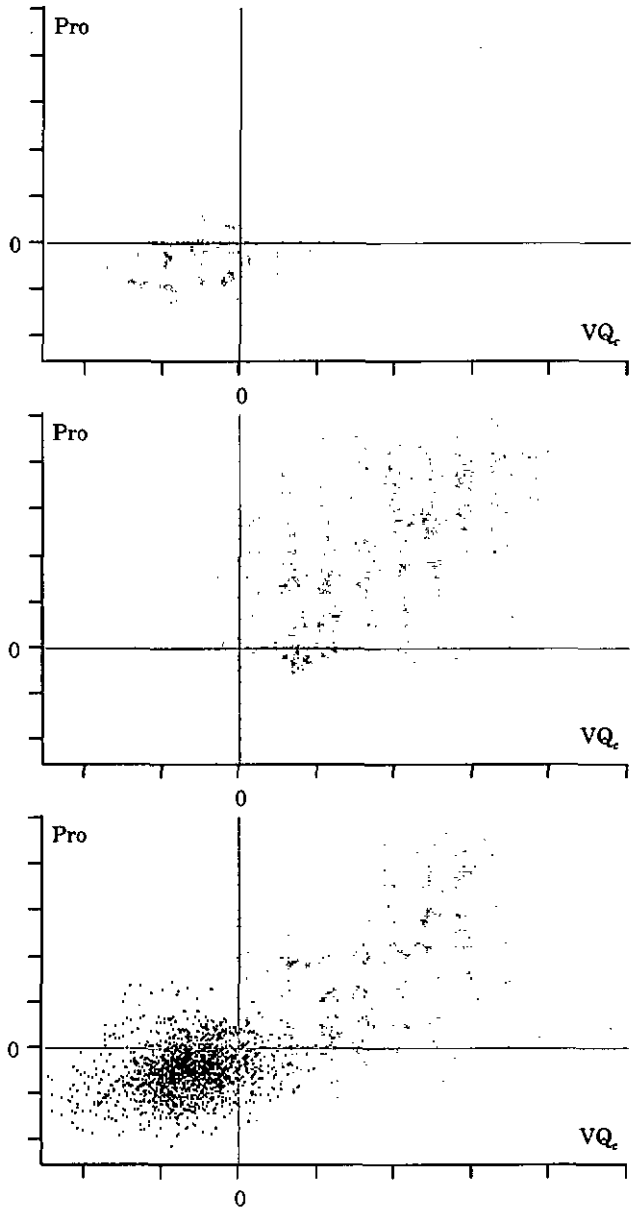

Figure 7.3.g Diagrammes de dispersion quant. vectorielle-proéminence

7.3.3 Critère objectif de dépendance des méthodes

Dans ce paragraphe, nous mesurons objectivement la dépendance linéaire entre les méthodes de reconnaissance grâce au coefficient de corrélation décrit à l'annexe A.3. Les données sur lesquelles porte le calcul de ce coefficient correspondent aux mesures de distance centrées décrites au paragraphe 7.3.2.A; en outre, quand ce calcul fait intervenir simultanément les distances du domaine homogène et celles du domaine hétérogène, nous avons pondéré les contributions de chaque domaine de sorte que E_h et $E_{\bar{h}}$ aient la même importance, malgré le fait que le nombre d'échantillons disponibles dans ces ensembles soient différents.

Les méthodes candidates sont $\{\mathrm{VQ}_c, \mathrm{Conf}, \mathrm{Pro}, \langle v(n)\rangle, \langle c(n)\rangle, \mathrm{VQ}_{\Delta c}\}$. La figure 7.3.h montre les coefficients de corrélation croisée entre toutes ces méthodes quand seul le domaine homogène E_h est considéré, tandis que la figure 7.3.i donne ces coefficients pour le domaine hétérogène $E_{\bar{h}}$. Enfin, la figure 7.3.j livre les coefficients quand les domaines sont confondus.

$\rho_{xy}(E_h)$	VQ_c	Conf	Pro	$\langle v(n)\rangle$	$\langle c(n)\rangle$	$\mathrm{VQ}_{\Delta c}$
VQ_c	1.00					
Conf	0.36	1.00				
Pro	0.24	0.12	1.00			
$\langle v(n)\rangle$	0.28	0.15	0.69	1.00		
$\langle c(n)\rangle$	0.51	0.54	0.29	0.20	1.00	
$\mathrm{VQ}_{\Delta c}$	0.44	0.08	0.10	0.10	−0.00	1.00

Figure 7.3.h Coefficients de corrélations homogènes

$\rho_{xy}(E_{\bar{h}})$	VQ_c	Conf	Pro	$\langle v(n)\rangle$	$\langle c(n)\rangle$	$\mathrm{VQ}_{\Delta c}$
VQ_c	1.00					
Conf	0.76	1.00				
Pro	0.51	0.54	1.00			
$\langle v(n)\rangle$	0.37	0.30	0.77	1.00		
$\langle c(n)\rangle$	0.67	0.65	0.43	0.29	1.00	
$\mathrm{VQ}_{\Delta c}$	0.27	0.08	−0.10	−0.11	−0.09	1.00

Figure 7.3.i Coefficients de corrélations hétérogènes

$\rho_{xy}(E_h \cup E_{\bar{h}})$	VQ_c	Conf	Pro	$\langle v(n) \rangle$	$\langle c(n) \rangle$	$VQ_{\Delta c}$
VQ_c	1.00					
Conf	0.88	1.00				
Pro	0.78	0.77	1.00			
$\langle v(n) \rangle$	0.59	0.57	0.83	1.00		
$\langle c(n) \rangle$	0.84	0.83	0.71	0.54	1.00	
$VQ_{\Delta c}$	0.44	0.31	0.21	0.12	0.20	1.00

Figure 7.3.j Coefficients de corrélations

L'élément encadré de la figure 7.3.j annonce la plus haute valeur non triviale de corrélation. La paire de méthodes auxquelles il correspond est donc formée de celles qui sont les plus corrélées; il s'agit de l'erreur moyenne de quantification vectorielle des cepstres complexes du filtre de synthèse et de leur conformité. Il en découle le fait que les informations délivrées par ces deux méthodes en lice sont de même nature. Nous allons donc les fusionner par la simple addition des distances qu'elles produisent; dans cette opération, la pondération de chacune est telle que la variance des opérandes soit unitaire. La méthode obtenue est nommée (VQ_c,Conf); l'extension naturelle de cette technique et le calcul des coefficients de corrélation correspondants conduit aux figures 7.3.k, 7.3.l, 7.3.m et 7.3.n, où seul le cas des domaines confondus est pris en compte.

ρ_{xy}	(VQ_c,Conf)	Pro	$\langle v(n) \rangle$	$\langle c(n) \rangle$	$VQ_{\Delta c}$
(VQ_c,Conf)	1.00				
Pro	0.80	1.00			
$\langle v(n) \rangle$	0.60	0.83	1.00		
$\langle c(n) \rangle$	0.86	0.71	0.54	1.00	
$VQ_{\Delta c}$	0.39	0.21	0.12	0.20	1.00

Figure 7.3.k Coefficients de corrélations

ρ_{xy}	((VQ_c,Conf),$\langle c(n) \rangle$)	Pro	$\langle v(n) \rangle$	$VQ_{\Delta c}$
((VQ_c,Conf),$\langle c(n) \rangle$)	1.00			
Pro	0.78	1.00		
$\langle v(n) \rangle$	0.59	0.83	1.00	
$VQ_{\Delta c}$	0.31	0.21	0.12	1.00

Figure 7.3.l Coefficients de corrélations

ρ_{xy}	$((VQ_c, Conf), \langle c(n) \rangle)$	$(Pro, \langle v(n) \rangle)$	$VQ_{\Delta c}$
$((VQ_c, Conf), \langle c(n) \rangle)$	1.00		
$(Pro, \langle v(n) \rangle)$	0.72	1.00	
$VQ_{\Delta c}$	0.31	0.17	1.00

Figure 7.3.m Coefficients de corrélations

ρ_{xy}	$(((VQ_c, Conf), \langle c(n) \rangle), (Pro, \langle v(n) \rangle))$	$VQ_{\Delta c}$
$(((VQ_c, Conf), \langle c(n) \rangle), (Pro, \langle v(n) \rangle))$	1.00	
$VQ_{\Delta c}$	0.26	1.00

Figure 7.3.n Coefficients de corrélations

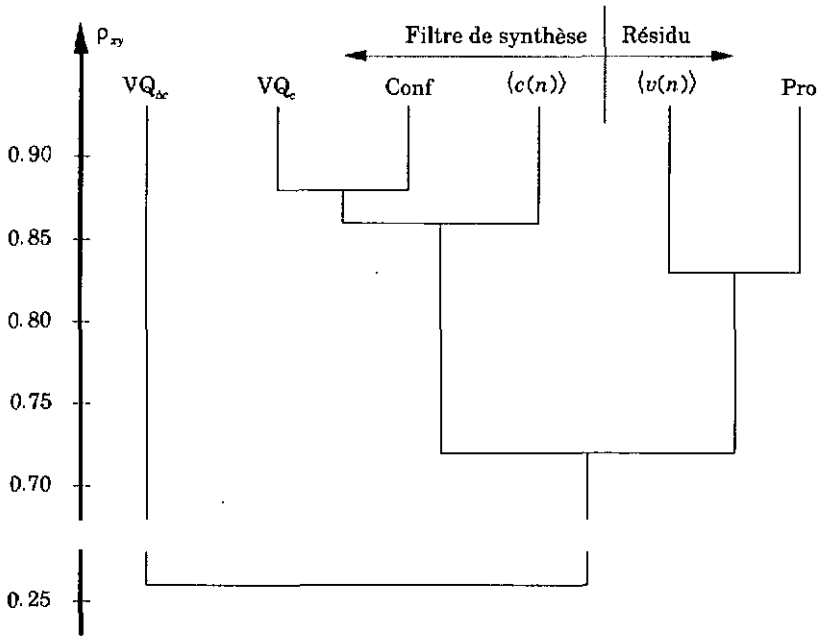

Figure 7.3.o Hiérarchie binaire indicée des méthodes de reconnaissance

En choisissant un indice d'agrégation égal à la valeur du coefficient d'autocorrélation, il devient possible de construire la hiérarchie binaire indicée de la figure

7.3.o qui permet d'illustrer graphiquement le lien de parenté des méthodes. Nous y avons placé à gauche les techniques qui exploitent le filtre de synthèse et à droite celles qui essaient de tirer profit du résidu.

La valeur du coefficient de corrélation entre deux méthodes n'est cependant pas nécessairement liée au succès de leur combinaison; ce n'est qu'un indice potentiel de leur insuccès. En effet, si deux méthodes sont parfaitement corrélées, alors il est certain que leur combinaison n'offre aucun avantage. Par contre, si deux méthodes sont partiellement décorrélées, alors il n'est que possible (mais non certain) que leur combinaison soit fructueuse.

En particulier, nous savons à l'issue du paragraphe 7.3.1 que la méthode de l'erreur moyenne de quantification vectorielle des cepstres complexes différentiels du filtre de synthèse est peu efficace relativement aux autres méthodes. La figure 7.3.o confirme le fait qu'en réalité elle est tellement mauvaise que les décisions qu'elle prend, basées sur la comparaison d'un vecteur représentatif et d'un ensemble de vecteurs caractéristiques, ne corroborent en aucune façon les décisions, souvent correctes, des autres méthodes; elle ne se prête donc pas à la combinaison. Cette remarque permet de ne pas remettre en question notre décision de rejeter la méthode $\mathrm{VQ}_{\Delta c}$, bien qu'elle montre la plus forte indépendance.

Concluons ce paragraphe en constatant que la hiérarchie binaire indicée de la figure 7.3.o agrège séparément d'une part toutes les méthodes tirant profit du filtre de synthèse (à l'exception évidente de $\mathrm{VQ}_{\Delta c}$) et d'autre part toutes les méthodes tirant profit du résidu; ces deux groupes séparés ne sont eux-mêmes réunis que lorsque le sommet de la hiérarchie est atteint. Cette figure est donc une preuve du bien-fondé de notre approche, puisqu'elle démontre que les informations délivrées par chaque groupe de méthodes sont de nature différente.

7.3.4 Choix de trois méthodes

L'éventail des 6 méthodes examinées est $\{\mathrm{VQ}_c, \mathrm{Conf}, \mathrm{Pro}, \langle v(n)\rangle, \langle c(n)\rangle, \mathrm{VQ}_{\Delta c}\}$. Le critère de performance objective du paragraphe 7.3.1 a permis d'en rejeter la dernière méthode ($\mathrm{VQ}_{\Delta c}$); en relation avec le critère subjectif de dépendance du paragraphe 7.3.2, il a de plus conduit au rejet des méthodes $\langle c(n)\rangle$ et $\langle v(n)\rangle$. L'ensemble des méthodes rescapées est par conséquent constitué par $\{\mathrm{VQ}_c, \mathrm{Conf}, \mathrm{Pro}\}$. Le coefficient de corrélation du paragraphe 7.3.3 encourage l'usage conjoint d'un membre de la famille des méthodes exploitant le résidu

avec celui d'un membre de la famille des méthodes exploitant le filtre de synthèse; comme une seule des méthodes encore en lice traite du résidu, nous la conservons définitivement (Pro). Enfin, la figure 7.3.e nous montre une indépendance subjective suffisante entre les méthodes VQ_c et Conf pour que nous ne renoncions ni à l'une, ni à l'autre.

En conclusion, nous avons retenu 3 méthodes différentes et indépendamment efficaces, robustes chacune envers un type particulier de dégradation du signal [84KraM]. Ces méthodes sont l'erreur moyenne de quantification vectorielle des cepstres complexes du filtre de synthèse (VQ_c), leur conformité (Conf) et la proéminence des cepstres réels du résidu (Pro). Nous venons de nous convaincre que l'espoir n'est pas vain de les voir se corriger mutuellement lorsque la tâche de vérification du locuteur liée à l'une de ces méthodes rend une décision incorrecte d'homogénéité de la paire de données constituée par l'entrée vocale et l'identité. Il nous reste à mener les expériences quantitatives qui permettront de montrer que ce choix répond à notre attente.

7.4 Combinaison des trois méthodes retenues et base II

Il existe plusieurs façons de combiner les méthodes de reconnaissance entre elles. Par exemple, on peut faire appel à la combinaison logique des décisions des méthodes individuelles [82MohN], ou à la combinaison des distances qu'elles mesurent. C'est à cette dernière classe de techniques qu'appartient le discriminant linéaire de Fisher, présenté à l'annexe A.4, que nous avons choisi comme première méthode de combinaison, en raison de son approche formelle et déterministe. Nous avons ensuite tenté l'utilisation d'une seconde technique de combinaison nommée pouvoir discriminant.

7.4.1 Principe de combinaison des méthodes

Le principe de combinaison est le même pour les deux techniques examinées et consiste à calculer une distance sous la forme d'une somme pondérée de contributions des méthodes individuelles de reconnaissance. Si d_{VQ}, d_{Conf} et d_{Pro} représentent les distances obtenues par les méthodes que nous avons choisies au paragraphe 7.3, alors la combinaison linéaire de ces distances est

7.4.1 $$d = \mathbf{w}^T \cdot \mathbf{d} \qquad \mathbf{d} = \begin{pmatrix} d_{VQ} \\ d_{Conf} \\ d_{Pro} \end{pmatrix}$$

Cette mesure de distance d est ensuite introduite dans le processus de décision ◘ de l'expression 2.5.1 où le seuil μ est déterminé selon la procédure ordinaire décrite au paragraphe 5.3.4 qui associe à chaque référence son propre seuil. Si nous interprétons la pondération $\mathbf{w}$ comme un vecteur directeur normal au plan de séparation entre domaine homogène et domaine hétérogène, plongé dans l'espace des distances, alors nous admettons pour la suite de ce paragraphe qu'il existe autant de plans de séparations que de seuils différents; ils possèdent cependant tous la même orientation, donnée par $\mathbf{w}$.

■ 7.4.2 Combinaison par discriminant linéaire de Fisher

Le discriminant linéaire de Fisher sépare par un plan l'espace des distances obtenues par chacune des méthodes. Ainsi, pour un point donné de cet espace, une des composantes porte une distance au sens de l'erreur moyenne de quantification, une autre composante celle obtenue par la méthode de la conformité et la dernière porte la distance due à la méthode de la proéminence. Pour calculer la pondération $\mathbf{w}$, nous avons estimé les matrices de dispersion $\mathbf{S}_h$ et $\mathbf{S}_{\bar{h}}$ de l'expression A.4.5 en tenant compte de toutes les distances à disposition issues des comparaisons de vecteurs représentatifs et de vecteurs caractéristiques des deux sessions qui forment l'ensemble d'apprentissage E_a. Il résulte de l'application de l'équation A.4.6 un vecteur $\mathbf{w}$, normal au plan de séparation entre domaine homogène et domaine hétérogène. Les composantes de ce vecteur pouvant être considérées comme représentatives d'une pondération sur les distances issues des méthodes retenues, il est instructif de connaître leur valeur.

Méthode	w
VQ_c	0.43
Conf	0.72
Pro	0.54

Figure 7.4.a Vecteur directeur de Fisher

L'interprétation de ces pondérations ne prend de sens que si l'on admet que les méthodes fournissent des distances de dimension comparable. Pour cela, les composantes de la figure 7.4.a ne sont valables que dans le cas où l'on consi-

dère que toutes les distances entre une identité ω_k et une locution $\lambda_j^{(i)}$ sont centrées et normales. L'opération de centrage, déjà rencontrée lors de l'établissement des diagrammes de dispersion, consiste à soustraire le seuil $\mu^{(k)}$ associé à la référence considérée; l'opération de normalisation consiste à diviser le résultat par la dispersion $\sqrt{\sigma_h^2+\sigma_{\bar{h}}^2}$ associée individuellement à chaque composante d_{VQ}, d_{Conf} et d_{Pro}.

L'opération de centrage, locale parce que dépendante de la référence, possède ici une importance identique à celle qu'elle avait aux diagrammes de dispersion du paragraphe 7.3.2. Pour une méthode donnée, elle permet en effet de comparer les distances par rapport à une origine commune, quel que soit le vecteur représentatif de référence. L'opération de normalisation étant linéaire et globale, son effet sur le taux de reconnaissance est nul; son avantage est cependant de permettre la comparaison de l'importance accordée, selon le critère de Fisher, à chacune des 3 méthodes.

Méthode	w	σ_h	$\sigma_{\bar{h}}$	$\sqrt{\sigma_h^2+\sigma_{\bar{h}}^2}$	$J(\mathbf{w})$
VQ_c	$(1.00\ 0.00\ 0.00)^T$	0.44	0.90	1.00	3.53
Conf	$(0.00\ 1.00\ 0.00)^T$	0.44	0.90	1.00	4.05
Pro	$(0.00\ 0.00\ 1.00)^T$	0.32	0.95	1.00	2.91
Fisher	$(0.43\ 0.72\ 0.54)^T$	0.50	1.34	1.43	4.99

Figure 7.4.b Dispersion et critère de Fisher

La figure 7.4.b montre le gain obtenu sur le critère $J(\mathbf{w})$ du discriminant linéaire de Fisher quand nous passons des méthodes individuelles (VQ_c, Conf, Pro) à leur combinaison (Fisher). On constate que, du point de vue de ce critère, la meilleure méthode indépendante est celle de la conformité; c'est aussi celle qui reçoit la pondération la plus élevée. L'usage conjoint des trois méthodes permet d'améliorer encore le critère retenu de 20% environ.

A) Combinaison de Fisher et base de données II

Les figures 7.4.c et 7.4.d donnent les matrices d'erreur obtenues dans le cas de la seconde base de données. Il découle de ces figures que le taux de fausse acceptation ρ_a vaut 3.5% et que celui de faux rejet ρ_r vaut 6.3%, le taux moyen d'erreur valant donc 4.9%.

$36 \cdot \rho_a$	XxXx	AnMa	IsMo	Made	MaSa	AnBe	BeZi	DaFi	FrLe	GePo	
AMSa	3	0	5	0	0	0	0	0	0	0	8
ChPa	4	1	6	0	0	0	0	0	0	0	11
ElGo	1	0	21	0	0	0	0	0	0	0	22
FrDC	0	0	9	1	2	0	0	0	0	0	12
MaHu	0	0	2	1	0	0	0	0	0	0	3
RuDr	0	0	0	8	0	0	0	2	0	0	10
FrKl	0	0	0	3	0	0	0	2	0	0	5
JaSE	0	0	0	2	0	0	5	36	0	2	45
JDBa	0	0	0	0	0	0	0	3	0	2	5
JJBe	0	0	0	0	0	0	1	18	0	3	22
LuMé	0	0	0	0	0	0	0	5	0	1	6
RoCa	0	0	0	0	0	0	0	1	0	0	1
	8	1	43	15	2	0	6	67	0	8	150
$\langle \rho_a \rangle$											3.5%

Figure 7.4.c Fausses acceptations

$180 \cdot \rho_r$	XxXx	AnMa	IsMo	Made	MaSa	AnBe	BeZi	DaFi	FrLe	GePo	
	12	0	0	52	11	8	14	7	1	9	114
$\langle \rho_r \rangle$											6.3%

Figure 7.4.d Faux rejets

Ce résultat conjoint est meilleur que n'importe lequel des résultats individuels observés jusqu'à maintenant pour chacune des 3 méthodes et répond à notre attente d'un gain du succès de reconnaissance pour une combinaison de méthodes judicieusement choisies.

Il est cependant surprenant de constater que ce n'est pas la méthode la plus efficace qui reçoit la plus grande pondération. La raison en est que le critère du discriminant linéaire de Fisher prend en compte la totalité des données dans la détermination du vecteur directeur de la droite de projection; il n'accorde pas une importance plus grande au nuage des distances critiques, proches des frontières de décision, qu'il n'en accorde à celui des distances éloignées de ces frontières. C'est d'ailleurs pour cette même raison que nous avons renoncé à faire usage du seuil théorique de l'équation A.4.7 et que nous y avons préféré l'usage d'un seuil pratique.

Il est toutefois encore plus important de se souvenir que la satisfaction optimale du critère de Fisher n'a pas pour conséquence nécessaire un pouvoir de discrimination maximal. En effet, le critère de Fisher ne correspond pas explicitement à celui de la minimisation du taux d'erreur des méthodes conjointes, qui est pourtant le critère que nous aimerions appliquer.

7.4.3 Combinaison par pouvoir discriminant

Nous allons tenter de pallier l'inconvénient du critère de Fisher, inhérent à une approche globale, en déterminant le vecteur directeur d'une autre façon. Celle que nous choisissons est très simple; nous décidons en effet de pondérer chaque méthode en accord avec sa propre efficacité, assimilée à l'inverse du taux d'échec observé lors des expériences précédentes. Il en résulte le vecteur directeur **w** de la figure 7.4.e, où l'on voit que nous avons accordé à chaque méthode le mérite qui lui revient.

Méthode	$\frac{1}{2}(\rho_a + \rho_r)$	**w**
VQ_c	5.7%	0.80
Conf	9.4%	0.49
Pro	12.9%	0.35

Figure 7.4.e Vecteur directeur du pouvoir discriminant

Bien que nous ayons abandonné l'espoir d'optimiser le critère du discriminant linéaire de Fisher, calculons tout de même la valeur qu'il prendrait pour le vecteur directeur de la figure 7.4.e et comparons aux conditions optimales de la figure 7.4.b le résultat présenté à la figure 7.4.f.

w	σ_h	$\sigma_{\bar{h}}$	$\sqrt{\sigma_h^2 + \sigma_{\bar{h}}^2}$	$J(\mathbf{w})$
$(0.80\ \ 0.49\ \ 0.35)^T$	0.51	1.31	1.41	4.80

Figure 7.4.f Dispersion et critère de Fisher

Nous constatons que, au sens du critère de Fisher, la qualité $J(\mathbf{w})$ de ce nouveau vecteur directeur n'est pas très éloignée de celle du vecteur optimal; en particulier, elle demeure supérieure à celles obtenues en considérant les méthodes de façon individuelle.

A) Combinaison par pouvoir discriminant et base de données II

Il nous reste encore à montrer l'essentiel, c'est à dire le succès de ce vecteur particulier qui dépend maintenant directement du pouvoir discriminant de chaque méthode individuelle. Pour cela, les figures 7.4.h et 7.4.i donnent les fausses acceptations et les faux rejets correspondants. Il découle de ces figures que le taux de fausse acceptation ρ_a vaut 2.6% et que celui de faux rejet ρ_r vaut 5.4%, le taux moyen d'erreur valant donc 4.0%.

$36\cdot\rho_a$	XxXx	AnMa	IsMo	Made	MaSa	AnBe	BeZi	DaFi	FrLe	GePo	
AMSa	0	0	3	0	0	0	0	0	0	0	3
ChPa	4	0	4	0	0	0	0	0	0	0	8
ElGo	0	0	13	0	0	0	0	0	0	0	13
FrDC	0	0	6	1	1	0	0	0	0	0	8
MaHu	0	0	1	5	0	0	0	0	0	0	6
RuDr	0	0	0	3	1	0	0	0	0	0	4
FrKl	0	0	0	1	0	0	0	1	0	0	2
JaSE	0	0	0	1	0	0	6	35	0	0	42
JDBa	0	0	0	0	0	0	0	3	0	0	3
JJBe	0	0	0	0	0	0	0	16	0	2	18
LuMé	0	0	0	0	0	0	0	6	0	1	7
RoCa	0	0	0	0	0	0	0	0	0	0	0
	4	0	27	11	2	0	6	61	0	3	114
$\langle\rho_a\rangle$											2.6%

Figure 7.4.h Fausses acceptations

$180\cdot\rho_r$	XxXx	AnMa	IsMo	Made	MaSa	AnBe	BeZi	DaFi	FrLe	GePo	
	2	3	1	39	10	2	17	13	1	10	98
$\langle\rho_r\rangle$											5.4%

Figure 7.4.i Faux rejets

La figure 7.4.j présente le diagramme de vérification des méthodes conjointes, où leur pondération a été déterminée selon le critère du pouvoir discriminant. On peut y déceler l'habituel décalage entre l'estimation du seuil a priori et le seuil a posteriori et constater, par comparaison avec la figure 6.1.l, la disparition des très grandes distances du domaine hétérogène. Par contre, les deux courbes

obtenues sont plus éloignées l'une de l'autre qu'elles ne l'étaient dans la figure citée, ce qui est le signe d'une meilleure robustesse.

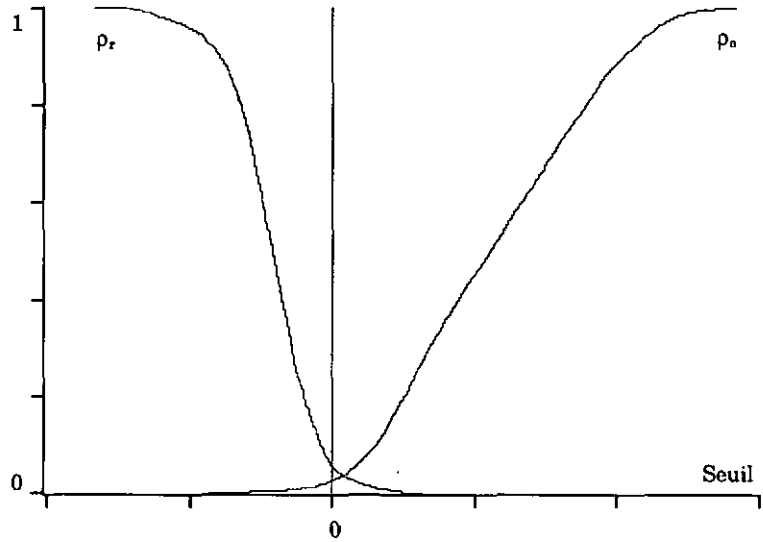

Figure 7.4.j Diagramme de vérification

Ce résultat est le meilleur que nous ayons observé sur notre seconde base de données. De ce point de vue pourtant, notre dernière façon d'établir le vecteur de pondération est intuitive et ne satisfait pas plus le critère de succès maximal que ne le faisait Fisher; elle ne satisfait d'ailleurs aucun critère explicite. Par conséquent, il nous faudrait maintenant entreprendre une optimisation de **w** selon un critère de minimisation du taux d'erreur, ce qui nous conduirait à déterminer ses 3 composantes en explorant un espace à deux paramètres seulement, parce que la norme du vecteur de pondération est une variable libre que nous avons arbitrairement décidée unitaire dans ce paragraphe et celui qui précède. Cette approche nous paraît une tâche besogneuse à laquelle nous ne nous livrerons pas, car nous pensons que le vecteur de la figure 7.4.e est déjà une bonne approximation de ce que nous pourrions trouver.

7.4.4 Récapitulation des résultats de trois méthodes conjointes

Nous avons construit une combinaison linéaire des distances issues de trois méthodes de reconnaissance de locuteur indépendante du texte dont deux sont de notre crû. Nous avons exploré deux façons de déterminer les valeurs de pondération en réalisant sur notre seconde base de données les expériences de reconnaissance dans le cadre d'une tâche de vérification; ces expériences montrent que le taux de succès qui résulte de ces combinaisons est supérieur à tous ceux rencontrés jusqu'à maintenant. La figure 7.4.k résume ces résultats; nous rappelons qu'ils ont été obtenus par l'usage de seuils a priori.

Combinaison	Base	ρ_a	ρ_r	$\frac{1}{2}(\rho_a + \rho_r)$
Fisher	11	3.5%	6.3%	4.9%
pouvoir discriminant	11	2.6%	5.4%	4.0%

Figure 7.4.k Résultats des méthodes conjointes

La littérature fait volontiers usage de seuils a posteriori qui offrent des résultats à la fois plus flatteurs et moins représentatifs des conditions réelles de reconnaissance; de sorte à permettre une confrontation à armes égales de nos résultats avec ceux que l'on y rencontre, nous allons aussi les donner sous cette forme. Sur notre seconde base de données et pour une méthode C, le taux moyen d'erreur équitable du pouvoir discriminant vaudrait 2.2% dans les conditions de l'expression 6.1.4 (les locuteurs de l'ensemble de test participent à la détermination des seuils, mais en revanche ne participent pas à la détermination des poids des diverses métriques ni à la construction du dictionnaire universel), et 1.2% dans celles de l'expression 6.1.3 (les locuteurs de l'ensemble de test participent à toutes les phases du processus d'apprentissage).

La matière rencontrée dans ce chapitre nous permet de conclure que l'indépendance attendue entre le filtre de synthèse et le résidu de l'analyse par prédiction linéaire existe bien dans la réalité, comme en témoigne la hiérarchie binaire indicée de la figure 7.3.o. Enfin, le taux d'erreur des méthodes conjointes de reconnaissance fondées sur l'usage de l'une ou de l'autre des familles de vecteurs caractéristiques s'étant montré inférieur à celui des méthodes individuelles, nous avons montré comment tirer parti de cette indépendance.

8 Spéculations

Nous avons exposé, au fil des chapitres précédents, toute une ménagerie de méthodes destinées à résoudre le problème de la reconnaissance de locuteurs indépendante du texte; certaines d'entre elles étaient déjà bien connues des spécialistes alors que d'autres étaient inédites. Cependant, comme nous l'avons déjà fait sentir à maintes occasions, le problème est loin d'être cerné et la multitude d'approches possibles rend le domaine presque inépuisable. Les deux difficultés majeures sont liées d'une part à l'état extrêmement diffus de l'information sur l'identité du locuteur, et d'autre part et surtout à la variabilité dont font preuve tous les locuteurs. Le mécanisme de production de la parole fait en effet non seulement appel à des structures de chair soumises à vieillissement, traumatisme, maladie bénigne ou grave, mais encore implique des processus de commande et de contrôle appris et modifiables, que ce soit consciemment ou inconsciemment. En outre, la parole possède une source inhérente de variabilité parce qu'elle est le support de l'information véhiculée par le langage. Dans ces conditions, trouver une méthode unique qui touche à tous les aspects du problème est un défi que nous ne saurions relever, pas plus que n'ont su le faire les autres chercheurs.

Le renoncement n'est pas non plus la bonne attitude; nous avons pu prouver dans cette thèse qu'il est possible de s'approcher de la solution du problème. Le but de ce chapitre est de présenter brièvement une extension possible de ce travail, que ce soit sous la forme de méthodes nouvelles ou la forme d'une réalisation palpable.

8.1 Utilisation du gain

Si nous examinons ce travail de thèse en tentant d'identifier les lacunes principales des méthodes nouvelles que nous y proposons, alors nous pouvons mettre en évidence deux défauts au moins. Le premier découle de l'observation suivante: nous avons analysé le signal de parole au moyen de la prédiction linéaire; des résultats de cette analyse, le filtre de synthèse et le résidu ont été

exploités. Par contre, le gain n'a fait l'objet d'aucune investigation de notre part! Il paraît donc légitime de vouloir examiner cette caractéristique complémentaire et de ne pas la rejeter aussi abruptement que nous l'avons fait, même si son usage est peu fréquent en reconnaissance de locuteurs, quoique supérieur à celui du résidu comme en témoigne [73LumR].

Le second défaut que nous voyons à notre approche est plus fondamental encore; il s'agit de se rendre compte que, de toutes les méthodes présentées ici, aucune ne considère l'écoulement du temps. Cela signifie par exemple que, même si nous permutions au hasard toutes les fenêtres découpées dans le signal de parole à la façon dont on bat un jeu de cartes, nous obtiendrions des résultats strictement identiques à ceux que nous avons déjà eus. Il est pourtant vraisemblable que, dans de telles conditions, un auditeur humain serait perturbé et diminuerait d'efficacité dans l'identification de locuteurs. La méthode des cepstres différentiels pourrait prétendre remédier à ce défaut, mais elle se contente en réalité de mesurer en chaque instant une pente temporelle, sans pour autant permettre de découvrir les caractéristiques qui se manifestent à une échelle de temps qu'elle est incapable d'appréhender.

Ce dernier défaut se rencontre dans presque la totalité des approches publiées de reconnaissance de locuteur indépendante du texte, à l'exclusion timide des méthodes faisant usage de modèles cachés de Markov, ou encore de réseaux neuromimétiques à délai temporel. Le qualificatif de timide est utilisé ici pour souligner le fait que la dépendance temporelle examinée par ces classes de méthodes est à relativement court terme, de l'ordre de la seconde, voire moins. Bien que ce ne soit pas le sujet de cette thèse, notons au passage que la reconnaissance de locuteur dépendante du texte exploite au contraire intensivement les informations disponibles dans l'évolution temporelle du mot de passe.

8.1.1 Mélange de voix

De sorte à comprendre mieux les mécanismes exposés ci-dessus, livrons-nous à une petite expérience informelle qui consiste à faire prononcer le même texte à deux locuteurs (i) et (j), à extraire des locutions $\lambda^{(i)}$ et $\lambda^{(j)}$ produites la suite temporelle de filtres de synthèse, de résidus et de gains issus de l'analyse par prédiction linéaire, et enfin à demander à quelques auditeurs de tenter d'identifier l'auteur d'une locution hybride $\lambda^{(i\wedge j)}$, fille bâtarde de $\lambda^{(i)}$ et $\lambda^{(j)}$, synthétisée après avoir permuté certaines suites des paramètres extraits. On peut aussi jouer sur le résidu et le remplacer par une excitation synthétique réalisée par

un train d'impulsions. La figure 8.1.a illustre les opérations de ce subterfuge que l'on trouve parfois dans la littérature [73MatH, 91SavM].

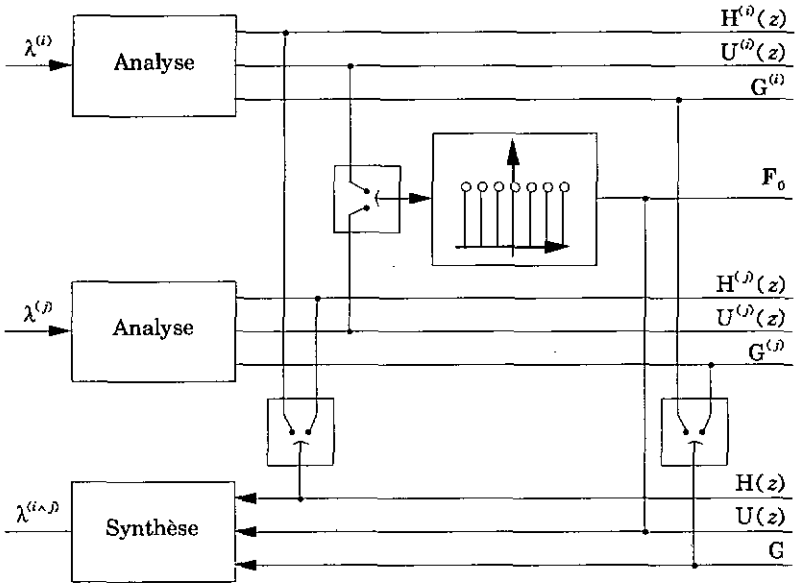

Figure 8.1.a Substitutions de paramètres

Les détails d'exécution exigent une correspondance temporelle grossière entre les deux locutions. La phrase prononcée est «Qui suis-je, d'où viens-je, où vais-je?» et sa durée est de 3 s environ. Son auteur devra être identifié par les 18 auditeurs participant à l'expérience, tous familiers des deux voix qui ne sont subjectivement ni particulièrement semblables ni particulièrement différentes; il s'agit de voix de locuteurs masculins.

De sorte à mieux mettre en évidence l'influence éventuelle du gain sur la reconnaissance humaine de locuteurs, nous avons ignoré la forme complète du résidu en le remplaçant par sa forme quantifiée selon le modèle de la figure 3.2.b. Ceci nous permet de construire 8 locutions manipulées en combinant tous les paramètres. Le travail des auditeurs est une tâche d'identification à $n = 2$; l'indécision n'est donc pas une réponse autorisée. La figure 8.1.b montre les *cas envisagés et les votes correspondants des auditeurs*, où H désigne le filtre de synthèse, F_0 le résidu quantifié, G le gain de l'analyse par prédiction

linéaire et où (*i*) et (*j*) sont les deux locuteurs. Les auditeurs n'avaient pas connaissance de la nature des manipulations, à l'exception des deux locuteurs qui participaient aussi à l'expérience en tant qu'auditeurs.

	H	F_0	G	(*i*)	(*j*)
1	(*i*)	(*i*)	(*i*)	17	1
2	(*i*)	(*i*)	(*j*)	14	4
3	(*i*)	(*j*)	(*i*)	11	7
4	(*i*)	(*j*)	(*j*)	10	8
5	(*j*)	(*i*)	(*i*)	4	14
6	(*j*)	(*i*)	(*j*)	0	18
7	(*j*)	(*j*)	(*i*)	5	13
8	(*j*)	(*j*)	(*j*)	0	18
Σ				61	83

Figure 8.1.b Manipulations réalisées et décisions des auditeurs

On constate d'emblée un déséquilibre en faveur du locuteur (*j*) qui reçoit plus de suffrages que l'autre locuteur; cependant, on constate aussi que le taux de confusion ρ_c est bas, puisqu'il ne vaut que 2.8% si l'on ne considère que les locutions où la seule manipulation intervenue est la quantification du résidu. Pour une analyse plus fine des résultats, reportons-nous à la figure 8.1.c qui intègre le nombre de votes basés sur chaque vecteur caractéristique. Par exemple, la locution d'indice 2 dans la figure 8.1.b contribue dans le tableau de la figure 8.1.c pour 14 votes basés sur H en faveur du locuteur (*i*), autant basés sur F_0, et pour 4 votes basés sur G en faveur du locuteur (*j*).

	(*i*)	(*j*)
H	37	48
F_0	35	46
G	62	63

Figure 8.1.c Critères de vote

Ce tableau met en évidence un fait surprenant: les auditeurs semblent avoir voté, inconsciemment, surtout en fonction du vecteur caractéristique que constitue le gain de l'analyse par prédiction linéaire! Nous leur avons alors demandé de se livrer à une analyse introspective des indices qui leur auraient permis d'identifier les locuteurs. Les réponses les mieux partagées désignent

l'accent régional de ces derniers comme facteur principal de reconnaissance; cet accent se traduit dans le cas particulier par des fins de mots plus traînantes pour l'un que pour l'autre. Cette observation subjective est cohérente avec l'importance objective assignée à chacun des critères de vote et donnée à la figure 8.1.c, puisque le gain G porte la majeure partie de l'information relative à l'évolution temporelle des modifications de l'énergie du signal, l'apport dû au filtre de synthèse étant peu variable du point de vue de l'énergie représentée.

8.1.2 Conclusion

Nous venons de voir que le gain de l'analyse par prédiction linéaire est un signal utile dans la caractérisation des locuteurs; en particulier, son évolution temporelle paraît prendre une importance capitale dans la petite expérience exposée au paragraphe 8.1.1. Nous admettons volontiers que cette expérience est peu représentative des conditions rencontrées ailleurs dans cette thèse. Par exemple, l'indépendance du texte n'y est pas respectée; en outre, nous ignorons si la paire de locuteurs qu'on y rencontre est banale ou au contraire extraordinaire. Cependant, nous croyons que le gain apporte une information complémentaire que seuls les insouciants peuvent se permettre de négliger.

8.2 Faisabilité en temps réel

La reconnaissance d'un locuteur par un auditeur humain est bien souvent de nature immédiate, même s'il arrive parfois que l'auditeur hésite ou tergiverse. En comparaison, il est intéressant de se poser la question de savoir quelle durée de traitement est nécessaire à un calculateur pour se livrer à un acte de reconnaissance d'une part, et d'autre part quelle est la durée de parole nécessaire pour rendre un jugement envers lequel il est possible d'accorder une confiance donnée. Pour ce second problème, l'examen des conditions d'exploitation de nos deux bases de données et les différents résultats que nous en obtenons permet de préciser au moins un point de fonctionnement (8 s pour 4.0%); nous n'avons cependant pas cherché à en trouver d'autres.

La réponse à la question de savoir si oui ou non un calculateur est capable d'effectuer une tâche de reconnaissance de locuteurs en temps réel est non seulement très dépendante des conditions techniques de réalisation des algorithmes, mais elle est encore liée à la définition même du qualificatif temps réel. Par exemple, pour un système échantillonné, une définition possible consiste à interdire l'usage de décimateurs, c'est à dire forcer la sortie du système à suivre

la cadence d'échantillonnage de son entrée, avec une tolérance libre ou imposée sur le délai entre le stimulus et la réponse. Dans le cas de la reconnaissance de locuteurs, cela signifie un acte de reconnaissance par échantillon de parole, ce qui paraît absurde car il est fort peu probable que des locuteurs se coupent la parole aussi fréquemment que cela, quelle que soit l'animation du débat. Une autre définition du temps réel se base sur la notion de besoin de celui qui recueille les résultats; cette notion, bien que floue, paraît la mieux appropriée dans tous les cas pratiques. Enfin, on peut encore appeler temps réel tout système répondant à une définition arbitraire telle que par exemple celle qui préconise que la durée de traitement ne doit pas dépasser la durée d'acquisition, sans préciser pour autant les conditions d'échantillonnage.

Nous avons choisi le cadre de cette dernière définition pour la reconnaissance de locuteurs indépendante du texte, où la sortie du système est une unique décision, quel que soit le nombre d'échantillons en entrée. Pour notre seconde base de données par exemple, cela signifie que la durée d'un traitement en rapport unitaire avec le temps réel vaut exactement 8 s et correspond à celle des fragments de test. Notons en passant que si nous parvenons à travailler en temps réel, alors les 6120 calculs de distance nécessaires à l'application de la méthodologie présentée à la figure 5.4.f impliquent une durée de traitement de E_t déjà supérieure à 13 h.

8.2.1 Temps de calcul VAX

Le calculateur à l'aide duquel nous avons réalisé le traitement automatique des données de ce travail de thèse est un `MicroVAX 3400`, de système d'exploitation `VAX/VMS V5.4-2`. Le langage de programmation utilisé est `Pascal` et le format que nous avons retenu code les nombres flottants sur 32 bit, les nombres irrationnels étant approchés au mieux par des nombres rationnels. Nous avons porté peu d'attention à la recherche de solutions informatiques rapides aux problèmes posés, pensant que l'intérêt principal de ce travail réside plus dans l'estimation de la qualité de reconnaissance des méthodes envisagées que dans l'optimisation en temps, en encombrement ou en régularité des algorithmes mis en œuvre.

Quelques cas concrets permettront de se fixer les idées. Par exemple, l'algorithme de classification des nuées dynamiques nécessite d'associer chaque échantillon à sa classe propre. Nous avons résolu ce problème par l'emploi d'autant de listes chaînées d'éléments que de classes en jeu; à chaque passage

d'un échantillon d'une classe dans l'autre nous l'avons extrait de la liste d'origine et nous l'avons incorporé à la liste de destination. Après quelques itérations de l'algorithme, l'ordre des éléments de ces listes ne correspond plus du tout à la séquence de leur adresse en mémoire. Or, en raison du nombre élevé d'échantillons utilisés, l'utilisation d'une mémoire virtuelle est indispensable; il s'ensuit que presque chaque accès finit par se traduire par un accès disque, ce qui pénalise lourdement la vitesse d'exécution. Cependant, seul l'aspect fonctionnel nous importe; par conséquent nous n'avons rien entrepris pour remédier à cet inconvénient.

Un autre exemple concerne le calcul du cepstre réel du résidu, qui requiert l'application de transformations de Fourier discrètes. Nous savons que des algorithmes de transformée rapide sont disponibles, non seulement pour les bases égales à des puissances de 2, mais encore pour de nombreux autres cas. Nous avons pourtant renoncé à les utiliser, par pur souci de simplicité. La seule concession que nous ayons faite à cet égard est de renoncer à des calculs en nombres complexes, grâce à une étape intermédiaire constituée d'une transformation de Hartley ne traitant que de nombres réels [89SteM].

Enfin, voici un dernier exemple de notre façon directe de traduire les algorithmes sous la forme de réalisations informatiques: en quantification vectorielle, il est indispensable de chercher le noyau le plus proche du vecteur à quantifier. Certains algorithmes permettent d'économiser une partie des comparaisons qu'exige une recherche exhaustive; notre approche a pourtant été de choisir cette dernière, de sorte à ne pas investir d'efforts de notre part dans les travaux préparatoires nécessaires à l'approche économe citée, mais à réserver nos forces pour le fond du problème.

Nous tenions à montrer par ces quelques exemples le goût et la couleur de notre approche de réalisation pratique des expériences dont nous avons parlé tout au long de cette thèse. C'est donc en gardant à l'esprit ces conditions qu'il faut considérer le tableau de la figure 8.2.a qui résume, pour certaines tâches importantes, le rapport entre l'estimation du temps d'exécution et le temps réel, au sens de la définition que nous y avons donnée.

Tâche	Facteur	Symbole
Analyse	270	$G \wedge c(n) \wedge v(n)$
Dictionnaire cepstral	25	Y
Proéminence	6	Pro
Conformité	5	Conf
Résidu moyen	2	$\langle v(n) \rangle$
Cepstre moyen	$\frac{1}{4}$	$\langle c(n) \rangle$
Quantification vectorielle	1	VQ_c

Figure 8.2.a Facteur temps réel des algorithmes réalisés sur VAX

Dans cette figure, nous voyons en première partie le facteur temps réel observé pour l'obtention de références ou de vecteurs caractéristiques prêts à être comparés. La seconde partie montre le facteur temps réel observé lors du calcul de distances; elle ne livre l'information que pour la quantification vectorielle parce que les autres méthodes ne consomment qu'un temps infime dans l'application d'une mesure de distance par rapport à celui utilisé pour l'établissement des vecteurs caractéristiques. On constate finalement que la charge en temps de calcul due à l'analyse par prédiction linéaire et à la transformation des vecteurs bruts qu'elle nous livre en vecteurs adaptés à la reconnaissance de locuteurs indépendante du texte est beaucoup plus grande que celle due à l'application des méthodes que nous avons envisagées.

D'un point de vue pratique, si l'on additionne les durées nécessaires à la création d'un vecteur représentatif pour la méthode conjointe du paragraphe 7.4, alors on obtient un facteur temps réel valant 306; le coût, en terme de facteur temps réel, d'un calcul d'une distance d'une locution par rapport à ce vecteur représentatif vaut encore 276. Ce coût tombe à 1 si l'on décide de comparer cette même locution par rapport à un autre vecteur représentatif existant, car il n'est plus alors besoin de traitements préliminaires. Quant à ces derniers, une analyse plus fine montre que l'analyse par prédiction linéaire est responsable d'un facteur temps réel valant 16, le complément à 270 étant à rechercher dans le calcul du cepstre réel du résidu qui consomme un facteur 125 environ par transformée de Fourier discrète. Notons que si nous faisions usage d'un algorithme rapide de transformation [82NusH], ce facteur ne vaudrait plus que 5 pour la conversion du résidu en son cepstre réel.

■ 8.2.2 Architecture matérielle

Au risque de nous répéter, rappelons que notre problème ne concerne que le principe de la reconnaissance; l'amélioration des conditions pratiques de réalisation n'a jamais été notre but principal. Toutefois, comme cet aspect revêt une importance technologique évidente [88AttJ, 88BigB], nous allons brièvement mentionner dans ce paragraphe une voie possible pour la construction d'un dispositif réel de reconnaissance, en l'état actuel de la technique. De façon concrète, tentons d'organiser grossièrement du matériel commercialement disponible de sorte à réaliser la même fonction que le logiciel que nous avons écrit, avec pour but un gain de vitesse tel que le facteur temps réel lié à un acte de reconnaissance soit proche de l'unité, ou inférieur.

La figure 8.2.b récapitule les étapes de prétraitement par lesquelles nous sommes passé et qu'il nous faut impérativement réaliser. La figure 8.2.c résume sous forme graphique les méthodes que nous voulons utiliser; nous y avons reporté en gris les bases de données qui contiennent l'infonnation représentative d'un locuteur. Ces figures ne sont valides que pour de la reconnaissance; les étapes nécessaires à l'apprentissage y sont exclues car il est presque toujours acceptable de le réaliser en temps différé.

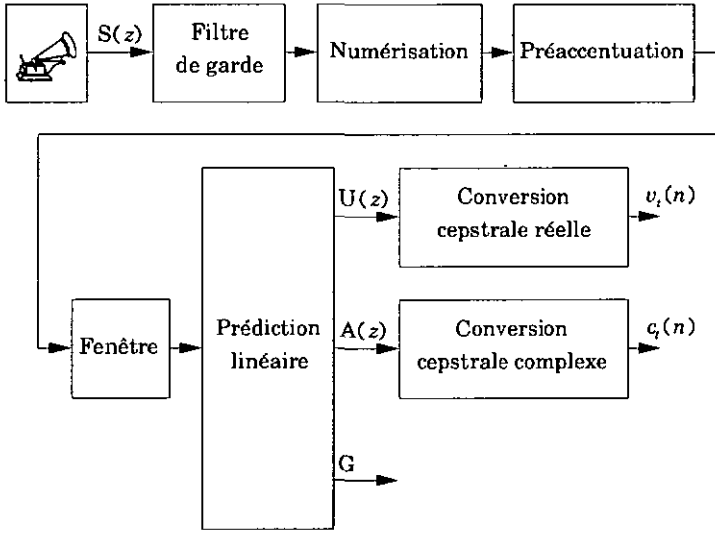

Figure 8.2.b Prétraitement

En fonction des commentaires du paragraphe 8.2.1, il est maintenant facile de repérer et de répartir les tâches lourdes. Par exemple, un premier processeur de traitement de signal pourrait s'occuper des détails de l'acquisition, de la prédiction linéaire et des deux conversions cepstrales qui consomment principalement deux transformées de Fourier. Un second processeur serait chargé de la quantification vectorielle (autant pour la méthode de la conformité que pour celle de l'erreur moyenne de quantification), de la proéminence et de la combinaison des résultats. Si nécessaire, il est encore possible d'adapter le taux de recouvrement de sorte à réduire le nombre d'opérations.

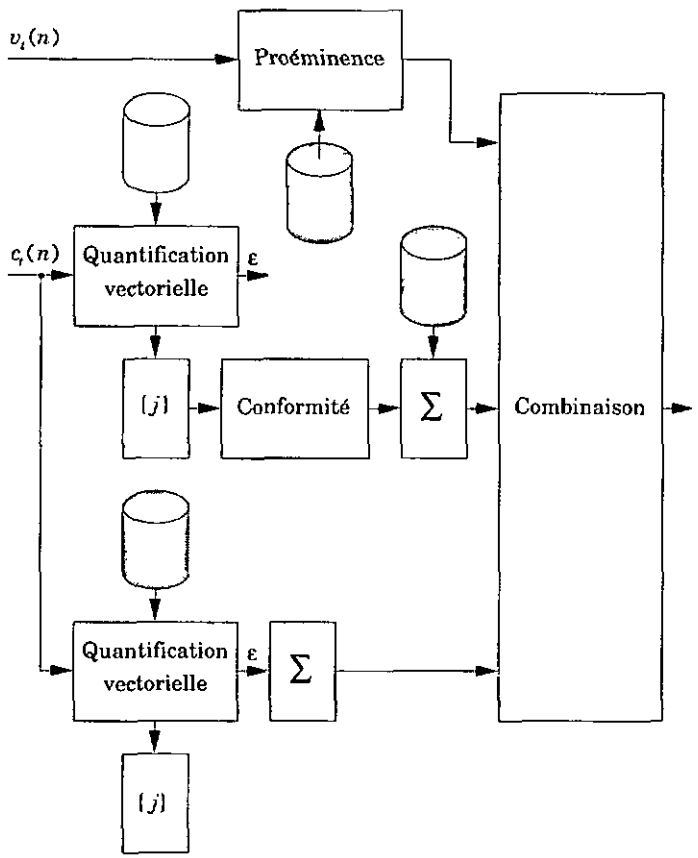

Figure 8.2.c Reconnaissance

La charge de calcul de la prédiction linéaire est principalement liée au calcul de la matrice d'autocorrélation; or, les opérations qui y correspondent sont très régulières et se prêtent bien à une optimisation sur un processeur spécialisé; il en va de même pour les transformées de Fourier, pour lesquelles de nombreux algorithmes beaucoup plus futés que celui que nous avons utilisé sont disponibles. Dans ces conditions, le hut d'un facteur unitaire envers le temps réel ne paraît pas inaccessible à la réalisation des éléments de la figure 8.2.b.

La charge de calcul de la quantification vectorielle est principalement liée aux nombreux calculs de distance euclidienne pondérée effectués en vain dans la détermination du noyau le plus proche voisin du vecteur à quantifier. Constatons là encore que le calcul d'une distance de ce type est une opération qui montre une bonne régularité; on peut donc espérer un gain important pour ces opérations si l'on utilise un processeur d'architecture adéquate. De plus, certains algorithmes astucieux de quantification vectorielle autorisent une économie du nombre de comparaisons effectives, parfois au prix d'une approximation dans le choix du plus proche voisin, et parfois non [84JayN, 86DavG]. Dans ces conditions, nous considérons comme possible la prise en charge simultanée des deux tâches de quantification vectorielle par un seul processeur. Le calcul de la proéminence et la combinaison des distances constitue une charge négligeable et n'est pas considéré ici.

■ 8.2.3 Applications potentielles

Mais à quoi donc pourrait bien servir un dispositif de vérification d'identité d'un locuteur semblable à celui que nous venons de décrire? Pour nous aider dans la formulation de la réponse à cette question, rappelons ici sa caractéristique essentielle: l'indépendance du texte. Quant à son taux de succès, souvenons-nous qu'il atteint 96%, ce qui signifie tout de même en moyenne 1 erreur tous les 25 essais. Ce dernier chiffre permet d'exclure dès l'abord toute application de nature judiciaire, car nous considérons que la condamnation d'un innocent est inacceptable, fût-ce même si ce moyen permet aussi de relaxer 24 autres honnêtes hommes [79BunE, 85HolH, 85TosO, 89NodH, 90MatP].

L'indépendance du texte est une médaille à deux faces. D'un côté elle offre une liberté plus grande au locuteur qui désire se faire reconnaître du système et qui n'a pas besoin de faire l'effort de se souvenir d'un mot de passe et de la façon précise de le prononcer. Elle rend aussi plus difficile le travail d'un imposteur volontaire, parce que ce dernier doit maintenant imiter la façon de parler de

son sujet de manière complète, alors que la dépendance du texte lui permettait de ne se concentrer que sur la seule émission du mot de passe. Mais, d'un autre côté, cette liberté est frustrante; si l'on essaie d'utiliser un tel système comme cerbère, nous pensons que les locuteurs n'auront que peu d'intérêt à dépenser une énergie imaginative dans le seul but de lui offrir un texte différent à chaque tentative. Dans de telles conditions, l'usage d'un système dépendant du texte paraît mieux approprié; de plus, celui-ci est généralement considéré comme plus performant que celui-là [75RosA, 80FurS, 82NeyH, 86BirM1, 86NaiJ, 88AttJ, 88VelG, 89NaiJ, 89RocC, 89ZalJ, 91GagD]. L'application typique d'un système de reconnaissance de locuteurs dépendante du texte est par conséquent le contrôle d'accès, que ce soit un accès physique [86BirM2] ou un accès à des informations [89GreS].

Alors, à quoi bon chercher à rendre indépendant du texte un système de reconnaissance de locuteurs? Nous répondrons finalement que l'application la plus prometteuse de la reconnaissance de locuteurs indépendante du texte est dans le contrôle continu d'une transaction orale:

- on peut imaginer par exemple un serveur téléphonique d'informations confidentielles où l'entrée serait gardée par un système dépendant du texte qui demanderait à la fois un code d'identification et un mot de passe pour en autoriser l'accès, puis qui céderait le pas à notre système indépendant du texte dont le travail serait de vérifier ensuite la concordance entre l'identité reconnue et la voix formulant les diverses requêtes nécessaires à l'interrogation du serveur. Notre système deviendrait alors un système de surveillance ne nécessitant aucune intervention de la part du locuteur;
- on peut encore imaginer de mettre en parallèle quelques systèmes de ce type, chacun d'eux étant adapté à un des locuteurs prenant part à une discussion. L'objet du travail consisterait alors à comptabiliser les temps de parole de façon automatique;
- on peut aussi utiliser un dispositif de ce genre pour une recherche automatique dans une banque de données sous forme orale. Le but serait de retrouver et d'extraire les citations d'un locuteur donné. L'utilisateur de ce dispositif pourrait par exemple fouiller les archives sonores d'une entreprise de radiodiffusion;
- on peut de plus incorporer un système de reconnaissance de locuteurs à un système de reconnaissance de la parole dépendant du locuteur, sous la forme d'un module de surveillance qui puisse avertir le système soit d'une dérive des caractéristiques vocales, et donc de la nécessité d'une mise à jour des références du locuteur en cours, soit simplement d'un changement de locuteur.

Dans ce cas, la réalisation d'une tâche d'identification pourrait servir à commuter automatiquement le jeu de références personnalisées;
- en outre, on peut utiliser ce système pour une caractérisation objective de la qualité d'un canal de transmission de voix où la conservation de l'identité des locuteurs serait une exigence des utilisateurs [84PapP].

Finalement, la reconnaissance de locuteurs automatique et indépendante du texte est bien adaptée à toutes les tâches de surveillance; par rapport à d'autres techniques, l'avantage principal de celle-ci est sa très grande discrétion. En effet, il suffit que le locuteur se manifeste par une émission de parole pour que la surveillance ait lieu; l'indépendance du texte libère le locuteur de toute contrainte et il n'a pas à se soucier de satisfaire le système par une action particulière. La surveillance peut même s'opérer à son insu, ce que chacun peut considérer comme un bien ou comme un mal selon son éthique personnelle et selon l'application envisagée.

9 Conclusions

La fin de ce mémoire de thèse approchant, il est d'usage de présenter le bilan des connaissances acquises, des échecs, des réussites, des espoirs déçus et des découvertes inattendues. Le chapitre des conclusions est aussi le lieu traditionnel où l'on trouve la synthèse des résultats. Nous avons choisi de présenter tout ceci dans un ordre chronologique dans le but de montrer le chemin parcouru entre notre point de départ et notre point d'arrivée, et d'imaginer comment notre pérégrination au royaume de la vérification de locuteurs indépendante du texte pourrait se poursuivre.

9.1 Point de départ

En premier lieu, nous avons désiré nous assurer de la concordance de nos résultats avec ceux de la littérature du domaine de sorte à nous assurer de l'efficacité de nos outils d'analyse (prédiction linéaire), de la validité des méthodes classiques de reconnaissance (méthode du cepstre complexe moyen du filtre de synthèse, celle de son erreur moyenne de quantification vectorielle), et de la justesse de nos critères de mesure (taux de faux rejet et de fausse acceptation d'une tâche de vérification). Nous avons ensuite cherché à combler les lacunes des méthodes de reconnaissance existantes, en nous fondant sur l'hypothèse que toutes les caractéristiques, sans exception, portent à divers degrés une manifestation de l'identité du locuteur. Il s'ensuit que nous avons introduit le cepstre réel du résidu de l'analyse par prédiction linéaire comme nouveau vecteur caractéristique ainsi qu'une nouvelle méthode générale de reconnaissance appelée conformité. Enfin, nous avons cherché à combiner les résultats issus de plusieurs méthodes de sorte à améliorer les performances de la reconnaissance.

9.2 Parcours

Ce travail s'articule en quatre éléments principaux. Le premier s'applique à la construction de deux bases de données qui puissent nous permettre de mener

les expériences de vérification du locuteur indépendante du texte. Le deuxième concerne la répétition de deux expériences classiques de reconnaissance et la confrontation des résultats que nous observons sur nos bases de données avec ceux de la littérature. Le troisième s'enquiert des performances accessibles en tâche de vérification par nos deux innovations, et compare ces résultats à ceux des expériences classiques. Le quatrième et dernier se rapporte au succès que nous pouvons obtenir de l'utilisation conjointe des certaines méthodes ainsi qu'à la justification du choix des méthodes à combiner.

9.3 Point d'arrivée

La construction de deux bases de données contenant la voix de nombreux locuteurs en quantité suffisante et propices à la reconnaissance de locuteurs indépendante du texte nous est apparue très tôt comme nécessaire; comme corollaire, nous avons dû nous livrer aux choix de la stratégie de leur exploitation. La première base de données, de petite taille, de vocabulaire restreint et ne contenant qu'une unique session par locuteur, convient bien à un travail préparatoire d'exploration de méthodes de reconnaissance. En particulier, elle se prête avec bonheur à une exploitation par une méthode C. La seconde base de données contient des locutions constituées de parole plus naturelle ou spontanée que celles de la première; ces locutions ont été récoltées dans des conditions acoustiques variables, ce qui accroît la difficulté du défi posé à la reconnaissance. Elle est en outre plus complète, autant en raison d'un nombre élevé de locuteurs que par le fait que nous disposons de plusieurs sessions pour beaucoup d'entre eux. La méthodologie d'utilisation que nous y appliquons se distingue de celle rencontrée chez d'autres auteurs par une robustesse plus élevée dans l'estimation des taux d'erreur, en raison du grand nombre de distances calculées. Nous réservons l'usage d'une méthode U à cette seconde base de données que nous acceptons volontiers de mettre à disposition d'autres chercheurs.

Nous rencontrons deux types de difficulté dans la confrontation de nos résultats avec ceux de la littérature: le premier est lié à la nature de la tâche de reconnaissance (comparaison de résultats d'identification avec des résultats de vérification), tandis que le second concerne la stratégie d'exploitation des bases de données (méthode C ou méthode U). Toutefois, en tenant compte de ces deux facteurs, nous constatons que les résultats que nous obtenons pour la méthode du cepstre complexe moyen et celle de son erreur moyenne de quantification

vectorielle sont comparables à ceux de la littérature, ce qui signifie que notre premier but est atteint.

Nous sommes encore en mesure de compléter la méthode classique de l'erreur moyenne de quantification vectorielle qui ne considérait jusqu'alors qu'une partie fragmentaire des données exploitables. Nous avons nommé méthode de la conformité notre nouvelle approche; seule, cette méthode produit des résultats meilleurs que ceux obtenus par la méthode du cepstre moyen. De plus, nous sommes en mesure de proposer une façon d'exploiter un vecteur caractéristique nouveau, laissé en jachère par les autres auteurs, qui n'est autre que le résidu de l'analyse par prédiction linéaire. Ce vecteur caractéristique offre l'avantage d'une orthogonalité certaine par rapport au filtre de synthèse intensivement utilisé dans le traitement de la parole; l'exploration de ses potentialités pour la reconnaissance de locuteurs indépendante du texte nous a permis de mettre à jour une méthode particulière que nous nommons proéminence. Seule, cette méthode produit des résultats comparables à ceux qu'obtient la méthode de la fréquence fondamentale moyenne, tout en exigeant un nombre plus petit d'ajustements de paramètres de calcul.

Nous sommes parvenu à combiner fructueusement trois méthodes de reconnaissance de locuteurs indépendantes du texte; ces méthodes sont l'erreur moyenne de quantification vectorielle, notre méthode de la conformité et celle de la proéminence. Nous avons pu montrer que le succès lié à l'utilisation conjointe des méthodes citées provient d'une complémentarité importante des caractéristiques du locuteur traitées séparément par chacune. En combinant ces méthodes, pour une méthodologie U, plusieurs sessions et seuils a priori, nous observons un taux d'erreur moyen valant 4%, ce que nous considérons comme excellent. Des résultats encore plus flatteurs peuvent s'obtenir si l'on accepte une méthodologie C, une seule session et des seuils a posteriori, car alors le taux moyen d'erreur équitable observé diminue encore et atteint la valeur minuscule de 1.2%.

9.4 Futur

A l'issue de ce travail, de très nombreuses voies s'ouvrent encore. Ainsi, pour mieux cerner la place des deux éléments nouveaux que nous proposons, nous pourrions encore les comparer, voire les utiliser, dans les multiples autres méthodes de reconnaissance de locuteur indépendante du texte que nous avons

ignorées ou à peine citées: quantification matricielle [90JuaB, 93CheM], modèles cachés de Markov [88ZheY, 90RosA, 90SavM, 91CarM, 91RosA] ou réseaux neuromimétiques [89RogC, 90BenY, 90OglJ, 90PfiB, 91BenY, 91OglJ, 91RudL, 92PakM], par exemple.

Nous pourrions aussi tenter d'améliorer encore l'adéquation entre la transformation du matériau brut constitué par les vecteurs caractéristiques eux-mêmes (fréquences d'apparition pour la méthode de la conformité, résidus pour celle de la proéminence) et la métrique qui les compare. Rappelons que du seul bonheur de ce mariage dépend le succès de l'extraction de l'information pertinente disponible dans le signal de parole.

Nous pourrions encore, dans une tâche de vérification, chercher à ne pas considérer seulement et uniquement la qualité de l'homogénéité entre une voix et une prétention d'identité au moyen de la seule référence associée à cette dernière, mais encore à tenir compte de la probabilité d'émission de cette voix par d'autres locuteurs [91HigA2]. Ainsi, les variations du locuteur dont on cherche à vérifier l'identité ne seraient dommageables que si elles tendent à le rendre semblable au modèle de ses contradicteurs, alors que dans l'état actuel ces variations sont dommageables dans tous les cas.

Par ailleurs, une réalisation concrète de nos algorithmes par du matériel efficace permettrait de tester leur validité en vraie grandeur. Ce n'est parfois que par ce genre d'approche que les faiblesses ou les avantages de ce qui n'apparaît que comme un détail pendant la phase exploratoire d'une technique prennent une importance insoupçonnée [92ReyD1].

Enfin et surtout, la voie la plus prometteuse à nos yeux nous emmène vers la découverte de l'aspect le plus négligé jusqu'à maintenant: l'évolution temporelle du signal de parole [86HigA, 90LhoE]. Nous pourrions par exemple nous intéresser à la vitesse d'élocution du locuteur où à sa façon d'utiliser les pauses dans le signal de parole. Dans cette voie, l'exploitation du signal de gain paraît le chemin le plus aisé à suivre car il s'agit d'une grandeur scalaire facile à manipuler et d'interprétation aisée; de plus, l'aspect complémentaire de ce dernier par rapport aux vecteurs caractéristiques dont nous avons déjà fait usage est évident. N'oublions cependant pas non plus que l'analyse de l'évolution temporelle du filtre de synthèse et de celle du résidu permettraient certainement aussi de dévoiler une part de l'information concernant l'identité du locuteur.

Annexes

A.1 Expériences réalisées

A.1.1 Fréquence fondamentale

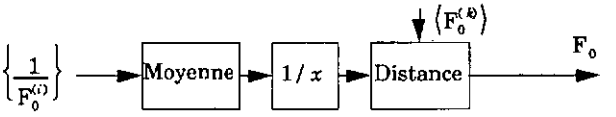

Figure A.1.a Valeur moyenne

A.1.2 Cepstre réel du résidu

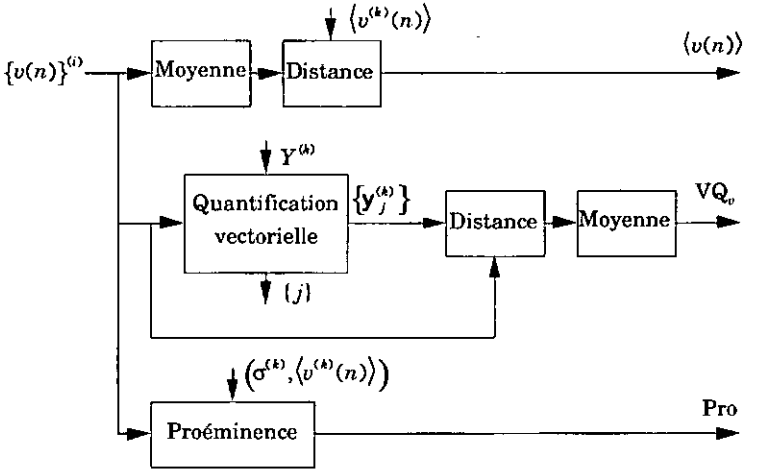

Fig. A.1.b Valeur moyenne, err. moyenne de quant. vectorielle, proéminence

■ A.1.3 Cepstre complexe du filtre de synthèse

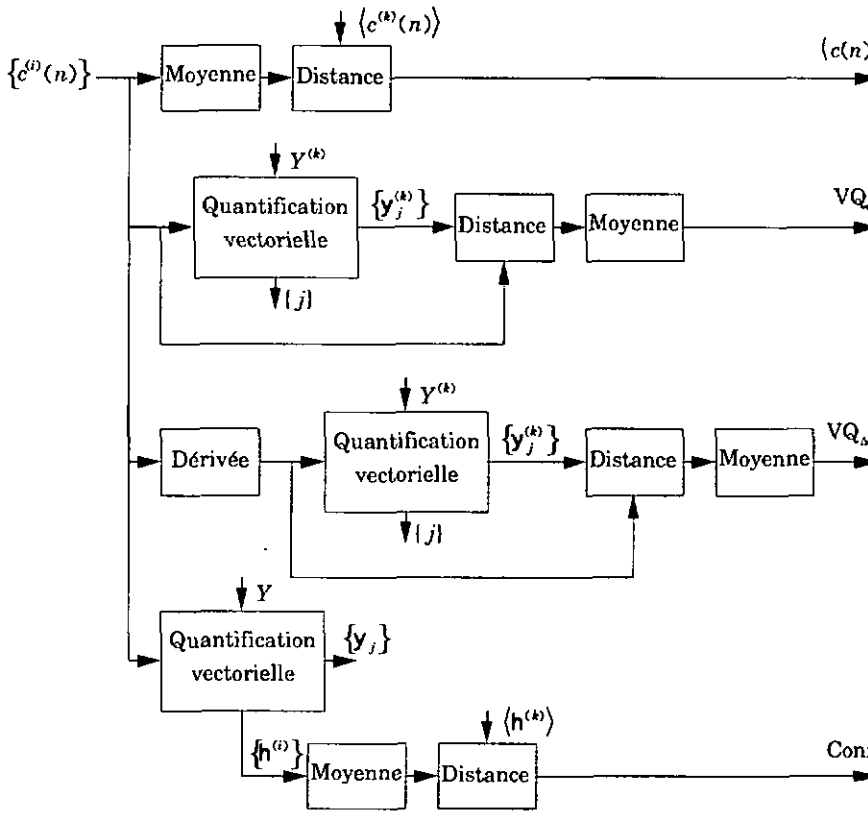

Figure A.1.c Valeur moyenne, erreur moyenne de quantification vectorielle immédiate puis différentielle, conformité

A.2 Taux d'erreur observés

A.2.1 Base de données I

Méthode	Distance	ρ_a	ρ_r	$\frac{1}{2}(\rho_a+\rho_r)$
$\langle F_0 \rangle$		10.2%	9.3%	9.7%
$\langle v(n) \rangle$	d_2	9.3%	9.1%	9.2%
VQ_c	d_2	22.7%	23.8%	23.2%
$\langle c(n) \rangle$	d_2	5.1%	4.3%	4.7%
VQ_c	d_2	2.6%	2.3%	2.5%
$VQ_{\Delta c}$	d_2	42.6%	42.7%	42.6%
Conf	d_2	3.1%	2.9%	3.0%

Figure A.2.a Seuils a posteriori

A.2.2 Base de données II

Méthode	Distance	ρ_a	ρ_r	$\frac{1}{2}(\rho_a+\rho_r)$
$\langle v(n) \rangle$	d_2	16.3%	14.8%	15.6%
	$d_{2,\text{Pond}}$	14.0%	18.6%	16.3%
	$d_{2,\text{Maha}}$	0.4%	84.6%	42.5%
Pro		11.4%	14.3%	12.9%
$\langle c(n) \rangle$	d_2	39.8%	55.9%	47.9%
	$d_{2,\text{Pond}}$	10.4%	21.9%	16.2%
	$d_{2,\text{Maha}}$	6.2%	25.7%	15.9%
VQ_c	d_2	4.6%	10.3%	7.5%
	$d_{2,\text{Pond}}$	3.3%	8.2%	5.7%
$VQ_{\Delta c}$	d_2	42.5%	25.2%	33.8%
Conf	d_2	9.0%	10.7%	9.9%
	$d_{2,\text{Pond}}$	7.3%	11.6%	9.4%
	$d_{2,\text{Maha}}$	0.3%	60.9%	30.6%
$\{\text{Pro}, VQ_c, \text{Conf}\}$	Fisher	3.5%	6.3%	4.9%
	pouvoir discr.	2.6%	5.4%	4.0%

Figure A.2.b Seuils a priori

A.3 Corrélation normalisée

Supposons que plusieurs méthodes de reconnaissance soient disponibles; dans ce cas, pour les mêmes ensembles d'apprentissage et de test E_a et E_t, chacune est susceptible de fournir des résultats différents. De façon plus détaillée, nous devons nous attendre à ce que les distances entrant dans le processus de décision soient dépendantes non seulement des données, mais encore de la méthode. Traduisons ce fait par formalisme approximatif, en reprenant les notations du paragraphe 2.4.4 où m désigne la méthode de reconnaissance utilisée dans une tâche de vérification

A.3.1 $$\frac{\partial d(\lambda_j^{(i)}, \omega_k, m)}{\partial m} \neq 0 \qquad \lambda_j^{(i)} \in L \ \wedge \ \omega_k \in \Omega^*$$

Il reste à déterminer si les méthodes impliquées sont corrélées. En effet, pour des présentations de couples (*voix*, *identité*) identiques, si une première méthode m_x livre des distances d_x dont les valeurs ont trop tendance à se calquer sur celles d'une seconde méthode m_y, même si elles en diffèrent au sens de l'expression A.3.1, alors nous affirmons que la combinaison de ces méthodes est infructueuse. Ce degré de corrélation se mesure par la valeur de covariance normalisée, ou coefficient de corrélation ρ_{xy} entre la méthode x et la méthode y. Soit σ_x la racine carrée de la variance des distances propres à la méthode x

A.3.2 $$\sigma_x = \sqrt{\mathbf{E}\left(d_x - \mathbf{E}(d_x)\right)}$$

Dans cette expression, l'estimateur **E** retourne l'espérance mathématique de son paramètre d'entrée. Par exemple, si la variable d_x est discrète et si le cardinal de l'ensemble de ses échantillons disponibles est N_x, alors

A.3.3 $$\sigma_x = \sqrt{\frac{\sum_{n=0}^{N_x-1} d_x^2(n) - \frac{1}{N_x} \cdot \left(\sum_{n=0}^{N_x-1} d_x(n) \right)^2}{N_x - 1}}$$

Soit encore C_{xy} la covariance entre les deux méthodes

A.3.4 $$C_{xy} = \mathbf{E}\left(\left(d_x - \mathbf{E}(d_x)\right) \cdot \left(d_y - \mathbf{E}(d_y)\right)\right)$$

Dans un cas discret, on suppose qu'il existe autant d'éléments pour la méthode x que pour la méthode y

A.3.5
$$C_{xy} = \frac{\sum_{n=0}^{N-1} d_x(n) \cdot d_y(n) - \frac{1}{N} \cdot \left(\sum_{n=0}^{N-1} d_x(n) \right) \cdot \left(\sum_{n=0}^{N-1} d_y(n) \right)}{N-1}$$

Finalement, le coefficient de corrélation ρ_{xy} est

A.3.6
$$\rho_{xy} = \frac{C_{xy}}{\sigma_x \cdot \sigma_y}$$

Il découle de l'inégalité de Schwarz que les valeurs prises par ce coefficient sont bornées. En particulier, une valeur unitaire équivaut à une dépendance absolue, tandis que si les méthodes sont indépendantes, alors le coefficient de corrélation est nul. Il est néanmoins important de remarquer que le contraire n'est pas nécessairement vrai, car un coefficient de corrélation nul ne signifie qu'une indépendance linéaire entre les variables considérées; tout autre type de dépendance reste admis

A.3.7
$$|\rho_{xy}| \leq 1$$

A.4 Discriminant linéaire de Fisher

Ce chapitre, comme le précédent, a pour but de montrer une technique exploitant l'information contenues dans les distances $d(\lambda_j^{(i)}, \omega_k, m)$; cette technique a pour nom discriminant linéaire de Fisher. Elle permet de déterminer une pondération linéaire optimale entre les méthodes de reconnaissance, au sens d'un critère objectif particulier, et nécessite la connaissance de ω_i l'identité vraie du locuteur. Il s'ensuit que cette technique se rencontre exclusivement dans la phase d'apprentissage.

Dans ce qui suit, nous assimilerons les distances, obtenues des diverse méthodes de reconnaissance, aux coordonnées d'échantillons plongés dans un espace U muni d'autant de dimensions que de méthodes. Nous nommerons domaine homogène X_h les distances calculées dans le cas où $\omega_i = \omega_k$, et domaine hétérogène $X_{\bar{h}}$ celles pour lesquelles $\omega_i \neq \omega_k$. Le but du discriminant linéaire de Fisher est de trouver un plan de séparation dans U tel que chaque domaine

occupe son propre demi-espace; il s'agit donc d'un problème de classification. Dans ce cas, le problème posé paraît être le même que celui déjà résolu par les nuées dynamiques pour $\mathbf{K} = 2$ classes, puisque tous deux ont pour but de découper l'espace U en deux régions dont la frontière soit un plan; la particularité du discriminant linéaire de Fisher est toutefois de placer ce plan de sorte à séparer les étiquettes associées aux échantillons, alors que l'algorithme des nuées dynamiques se contente de séparer les échantillons eux-mêmes.

A.4.1 Principe

Le principe du discriminant linéaire de Fisher est de projeter tous les échantillons sur une droite passant par l'origine et de ne considérer que la densité des points projetés sur cette droite. L'action menée est de faire tournoyer cette dernière autour de l'origine jusqu'à lui trouver une orientation telle que les projections des deux domaines soient à la fois compactes et distantes. Le plan de séparation est alors choisi perpendiculaire à cette droite, et sa position par rapport à l'origine est donnée par l'algorithme. Bien que ce processus soit déterministe et direct, la figure A.4.a en donne une vision itérative pour mieux en illustrer le principe.

La partie gauche de la figure A.4.a montre quelques échantillons plongés dans un espace à deux dimensions, ainsi que leur étiquette associée constituée d'un disque soit gris soit blanc. La partie supérieure de cette figure présente un cas où le plan de séparation est normal à l'axe x; la partie supérieure droite montre comment se distribuent les échantillons étiquetés pour cette orientation particulière de la droite de projection. La partie médiane propose une deuxième orientation de cette droite, et la partie inférieure une troisième encore. Les distributions correspondantes sont données dans la partie droite de la figure, où ξ est une grandeur proportionnelle à l'éloignement de l'origine du point projeté.

On constate, par l'examen de la partie droite de la figure A.4.a, que la compacité et la distance des deux ensembles concernés augmente entre la partie supérieure et la partie inférieure de la figure, au point qu'il devient possible de trouver un seuil de décision sur ξ tel qu'il n'en résulte plus que 3 erreurs de classification (2 disques gris et 1 blanc seraient mal classés).

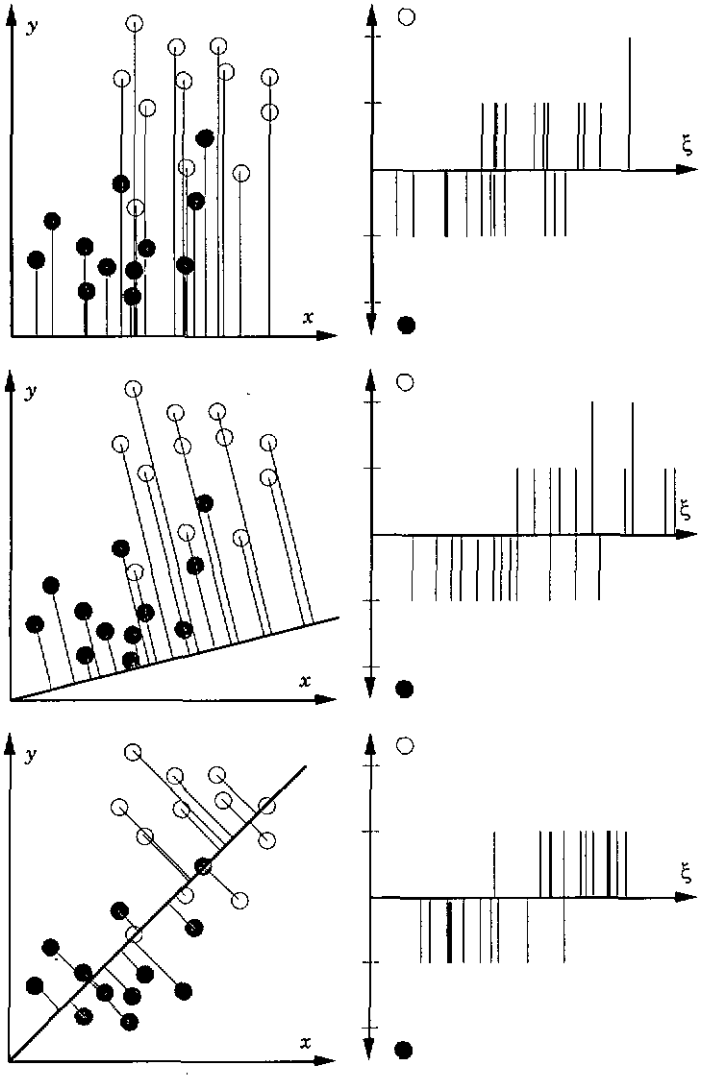

Figure A.4.a Principe du discriminant linéaire de Fisher

■ A.4.2 Critère d'optimisation

La technique du discriminant linéaire de Fisher mesure la distance entre les deux ensembles par le carré de l'écart entre les projections de leur centre de

gravité, la mesure de compacité globale étant remplacée par une mesure de dispersion, donnée par la somme des variances des projections des échantillons de l'un des deux ensembles d'une part, et de ceux de l'autre ensemble d'autre part. Le critère à maximiser est le rapport entre le carré de la distance entre les ensemble et leur dispersion. Formellement, soient X_h et $X_{\bar{h}}$ les deux ensembles d'échantillons $\mathbf{x}$; si $\mathbf{w}$ est un vecteur qui donne la direction de la droite de projection, alors l'éloignement de l'origine est proportionnel à ξ

A.4.1 $$\xi = \mathbf{w}^T \cdot \mathbf{x} \qquad \forall \mathbf{x} \in X_h \cup X_{\bar{h}}$$

Les centres de gravité des projections sont identiques aux projections des centres de gravité

A.4.2 $$\begin{cases} m_h = \dfrac{\mathbf{w}^T}{\mathrm{Card}(X_h)} \cdot \displaystyle\sum_{\forall \mathbf{x} \in X_h} \mathbf{x} = \mathbf{w}^T \cdot \mathbf{m}_h \\ m_{\bar{h}} = \dfrac{\mathbf{w}^T}{\mathrm{Card}(X_{\bar{h}})} \cdot \displaystyle\sum_{\forall \mathbf{x} \in X_{\bar{h}}} \mathbf{x} = \mathbf{w}^T \cdot \mathbf{m}_{\bar{h}} \end{cases}$$

Les dispersions sont données par

A.4.3 $$\begin{cases} \sigma_h^2 = \dfrac{1}{\mathrm{Card}(X_h) - 1} \cdot \displaystyle\sum_{\forall \mathbf{x} \in X_h} \left(\mathbf{w}^T \cdot \mathbf{x} - m_h\right)^2 \\ \sigma_{\bar{h}}^2 = \dfrac{1}{\mathrm{Card}(X_{\bar{h}}) - 1} \cdot \displaystyle\sum_{\forall \mathbf{x} \in X_{\bar{h}}} \left(\mathbf{w}^T \cdot \mathbf{x} - m_{\bar{h}}\right)^2 \end{cases}$$

Le critère à maximiser est

A.4.4 $$J(\mathbf{w}) = \frac{(m_{\bar{h}} - m_h)^2}{\sigma_{\bar{h}}^2 + \sigma_h^2}$$

Nous ne développerons pas ici les étapes conduisant à l'obtention du vecteur $\mathbf{w}$ satisfaisant la minimisation du critère exprimé en A.4.4. Bornons nous à souligner le fait que ce vecteur est déterminé à une constante d'échelle près, puisque seule sa direction nous importe.

■ A.4.3 Solution

Soient les matrices de dispersion $\mathbf{S}_h$ et $\mathbf{S}_{\bar{h}}$ données par

A.4.5
$$\begin{cases} \mathbf{S}_h = \dfrac{1}{\mathrm{Card}(X_h)-1} \cdot \sum_{\forall \mathbf{x} \in X_h} (\mathbf{x}-\mathbf{m}_h)\cdot(\mathbf{x}-\mathbf{m}_h)^{\mathrm{T}} \\ \mathbf{S}_{\bar{h}} = \dfrac{1}{\mathrm{Card}(X_{\bar{h}})-1} \cdot \sum_{\forall \mathbf{x} \in X_{\bar{h}}} (\mathbf{x}-\mathbf{m}_{\bar{h}})\cdot(\mathbf{x}-\mathbf{m}_{\bar{h}})^{\mathrm{T}} \end{cases}$$

La droite de projection optimale est alignée sur le vecteur **w**

A.4.6
$$\mathbf{w} = (\mathbf{S}_{\bar{h}} + \mathbf{S}_h)^{-1} \cdot (\mathbf{m}_{\bar{h}} - \mathbf{m}_h)$$

Le seuil de décision μ correspondant est aussi livré par le discriminant linéaire de Fisher

A.4.7
$$\mu = \frac{(\mathbf{m}_{\bar{h}} - \mathbf{m}_h)^{\mathrm{T}} \cdot (\mathbf{S}_{\bar{h}} + \mathbf{S}_h)^{-1} \cdot (\sigma_h^2 \cdot \mathbf{m}_{\bar{h}} + \sigma_{\bar{h}}^2 \cdot \mathbf{m}_h)}{\sigma_{\bar{h}}^2 + \sigma_h^2}$$

A.5 Partition initiale des nuées dynamiques

A.5.1 Base de données I

Nous montrons ici comment construire une bonne partition initiale Y_0 en utilisant la totalité de notre base de données. Nous y parvenons en trois étapes qui consistent respectivement en la recherche d'une contribution individuelle pour chaque locution, suivie d'une sélection des éléments représentatifs de ces contributions, et terminée par l'établissement d'une contribution globale.

La première étape fournit une contribution indépendante pour chaque locution $\lambda_j^{(i)}$. Nous commençons en effet par établir une partition initiale $Y_0^{(i)}$ simplement en considérant les K premiers vecteurs de la locution concernée, puis nous calculons une partition finale $Y^{(i)}$ pour cette locution selon l'algorithme donné en 2.2.16. La métrique d utilisée dans la mesure d'agrégation des classes selon 2.2.13 est euclidienne.

La deuxième étape établit une contribution globale pour toutes les locutions $\lambda_j^{(i)}$, dont nous réalisons la quantification vectorielle par leur dictionnaire personnalisé $Y^{(i)}$ obtenu lors de la première étape. De façon indépendante pour chaque locution, nous ne conservons de chaque classe qu'une fraction des données quantifiées; cette fraction vaut par exemple 10%, ce qui garantit la dis-

parition des noyaux trop peu représentatifs. Nous amalgamons alors les vecteurs conservés de sorte à créer une unique grande locution finale. Réduire ainsi le volume des données de 80 locutions d'une durée de 15 s à 1 locution d'une durée de 120 s est souhaitable pour deux raisons: d'une part, la tâche de classification s'en trouve accélérée d'autant, et, d'autre part, certains noyaux peu fréquentés du dictionnaire disparaissent, laissant la place à une représentation plus fine des classes les plus importantes.

La troisième étape consiste à mener le travail de classification sur l'amalgame, en utilisant comme partition initiale ses premiers vecteurs. La partition obtenue après convergence sert enfin de partition initiale unique Y_0, identique pour chaque locution, car nous admettons que cette partition est de bonne qualité et propre à générer pour chaque locution une partition finale $Y^{(i)}$ proche du minimum absolu.

A.5.2 Base de données II

Nous avons simplifié la construction des dictionnaires de notre seconde base de données en renonçant à la procédure compliquée énoncée au paragraphe A.5.1. Nous nous sommes contentés de les établir en faisant jouer le rôle de la partition initiale aux K premiers vecteurs du premier de la paire de fragments servant à la génération d'une référence. La métrique d utilisée dans la mesure d'agrégation des classes selon 2.2.13 est euclidienne ou euclidienne pondérée, en rapport avec l'expérience à mener.

A.6 Caricature des résultats de la littérature

Nous présentons ci-dessous un tableau de résultats donnant une vision très caricaturale des performances de nombreux systèmes de reconnaissance du locuteur décrits dans la littérature. Chaque colonne tente de tirer un des traits de ce portrait sommaire; les catégories que nous y avons fixées sont arbitraires et ne constituent qu'un reflet infidèle de la réalité. Dans de nombreux cas, la valeur du taux d'erreur correspond au meilleur taux observé par un auteur sur sa base de données, tandis que dans d'autres cas il s'agit d'un taux typique. Les valeurs ont parfois été tirées des diagrammes fournis par les auteurs de l'article considéré et n'ont qu'une faible précision numérique; dans d'autres cas, nous avons dû recalculer le taux d'erreur en fonction des informations disponibles.

■ A.6.1 Légende

1) Système de reconnaissance
H Humain
A Automatique

2) Degré de dépendance du texte
D Dépendance du texte
R Indépendance restreinte
I Indépendance

3) Signal considéré
T Tout le signal
P Signal de parole sans les pauses
L Voix laryngée exclusivement
A Autre

4) Méthode d'analyse
H Filtre de synthèse de l'analyse par prédiction linéaire
U Résidu de l'analyse par prédiction linéaire
B Banc de filtres
F0 Fréquence fondamentale
A Autre

5) Tâche de reconnaissance
I* Identification 1 à n
I Identification 1 à n+1
Vc Vérification en méthode C
Vu Vérification en méthode U

6) Taux d'erreur
% Taux typique ou meilleur taux

A.6.2 Taux d'erreur

Référence	1)	2)	3)	4)	5)	6)
[62KerL]	H	D	T	B	I*	1.0
[63ComA]	H	D	L	?	I*	35.0
[72FurS]	A	?	?	?	I*	9.0
[72FurS]	A	?	?	?	V?	7.0
[72WolJ]	A&	R	A	B	Vc	2.0
[73LumR]	A	D	T	A	Vc	1.2
[74FurS]	A	?	?	H	V?	1.0
[75RosA]	A	D	T	H	Vc	3.6
[79BunE]	A	I	?	A	I*	0.5
[79BunE]	A	I	?	A	Vu	1.0
[79MarJ]	A	I	L	H	I*	1.9
[79MarJ]	A	I	L	H	Vc	4.3
[80FurS]	A	D	T	H	Vc	0.1
[80FurS]	A	D	T	H	Vu	0.8
[81FurS1]	A	D	T	H	Vc	0.1
[81FurS1]	A	D	T	H	Vu	0.8
[81FurS2]	A	D	L	H	I*	5.0
[81FurS2]	A	D	L	H	Vc	3.0
[81ShrM]	A	I	L	H	I*	3.5
[81ShrM]	A	I	L	H	Vc	3.3
[82MohN]	A	I	P	H	I	0.8
[82MohN]	A	I	P	H	Vu	1.0
[82NeyH]	A	D	T	B	I*	1.0
[82NeyH]	A	D	T	B	Vc	2.5
[82SchR]	A	I	T	H	I*	2.5
[82ShrM]	A	I	L	H	I*	2.6
[82ShrM]	A	I	L	H	Vc	5.0
[83HunM]	A	I	L	H	I*	11.0
[83LiK]	A	I	P	H	I*	4.0
[83LiK]	A	I	P	H	I	5.1
[83ShrM]	A	I	L	H	I*	12.5
[84KraM]	A	I	?	H	I*	31.0
[84PapP]	H	D	T	?	I*	14.0
[85GisH]	A	I	?	H	I*	17.0
[85SooF]	A	R	T	H	I*	1.5

Référence	1)	2)	3)	4)	5)	6)
[85Wol]]	A	I	P	H	I*	5.0
[86BirM1]	A	D	T	H	Vc	1.9
[86BirM2]	A	D	T	H	Vc	1.9
[86GisH]	A	I	?	H	I*	5.0
[86HigA]	A	I	?	H	I*	0.0
[86HigA]	A	I	?	H	Vc	4.0
[86NaiJ]	A	D	T	H	Vc	0.4
[86RosA]	A	D	P	H	Vc	0.3
[86RosA]	A	R	P	H	Vc	3.1
[86SooF]	A	R	P	H	I*	7.2
[88AttJ]	A	R	P	H	Vu	0.9
[88AttJ]	A	I	P	H	Vu	1.9
[88BigB]	A	D	T	H	Vc	0.3
[88LiK]	A	I	?	H	I*	11.7
[88LiK]	A	I	?	H	Vc	11.2
[88NakH]	A	I	P	A	I*	0.0
[88NodH]	A	R	L	H	Vc	9.1
[88SooF]	A	R	P	H	I*	7.2
[88VelG]	A	R	P	H	Vc	4.8
[88ZheY]	A	D	T	H	I*	6.3
[89MasJ]	A	R	?	?	I*	6.4
[89NaiJ]	A	D	T	H	Vc	2.3
[89NodH]	A	R	L	H	I*	14.5
[89NodH]	A	R	L	H	Vc	4.0
[89RocC]	A	D	T	H	Vc	3.0
[89XuL1]	A	R	P	H	I*	2.0
[89XuL2]	A	R	P	H	I*	2.0
[89ZalJ]	A	R	L	H	?	2.5
[90BenY]	A	D	P	A	I*	3.0
[90JuaB]	A	R	P	H	I*	0.5
[90OglJ]	A	R	?	H	I*	8.0
[90PfiB]	A	D	T	H	Vc	15.0
[90RosA]	A	R	P	H	Vc	1.8
[90SavM]	A	?	L	H	Vu	9.6
[91BenY]	A	I	T	H	I*	2.0
[91CarM]	A	R	P	B	Vu	4.5
[91GagD]	A	D	T	H	Vu	1.1

Référence	1)	2)	3)	4)	5)	6)
[91HigA1]	A	R	T	H	Vc	1.7
[91KreJ]	H	I	T	?	I*	21.5
[91KreJ]	H	I	T	?	Vu	15.5
[91MatT]	A	I	?	H	I*	2.0
[91OglJ]	A	R	P	B	Vc	4.5
[91RosA]	A	R	P	H	Vc	0.3
[91RosR]	A	I	?	B	I*	0.5
[91RudL]	A	I	P	B	I*	0.0
[92ChaH]	A	D	L	H	Vc	11.0
[92GonY]	A	D	L	H	I*	8.0
[92HatH]	A	I	?	B	I*	0.0
[92KaoY]	A	I	?	H	I*	6.7
[92MatT]	A	?	?	H	I*	4.5
[92NetL]	A	D	T	H	Vc	4.3
[92PakM]	A	?	L	H	I*	1.5
[92ReyD1]	A	I	P	B	I*	11.7
[92ReyD2]	A	I	?	?	I*	21.2
[92SavM]	A	D	L	?	Vc	7.0
[92TseB]	A	R	P	H	Vc	4.0
[92TseB]	A	R	P	H	I*	1.7
[93CheM]	A	R	P	H	I*	0.0
[93ThéP]	A	I	T	A	Vc	1.2
[93ThéP]	A	I	T	A	Vu	4.0

Figure A.6.a Caricature des taux d'erreur

A.6.3 Pionniers des principales méthodes de reconnaissance du locuteur indépendantes du texte

Nous donnons ci-dessous une liste des méthodes les plus représentatives de l'état de l'art en reconnaissance du locuteur indépendante du texte. Chaque méthode citée est complétée par une référence ayant pour propriété d'être une des premières à l'introduire. Nous ne nous sommes cependant livré à aucune recherche historique approfondie; nous ne prétendons donc nullement donner ici la liste des inventeurs de ces méthodes, le terme de pionnier convenant mieux aux auteurs cités.

A) Reconnaissance humaine	
Audition	[91KreJ]
Sonagrammes	[62KerL]
B) Reconnaissance automatique	
Valeur moyenne	[72FurS]
Sélection de caractéristiques	
Manuelle	[72WolJ]
Par modèles cachés de Markov	[90SavM]
Densité de probabilité explicite	[82SchR]
Quantification vectorielle instantanée	
Densité explicite des erreurs de quantification	[83LiK]
Erreur moyenne de quantification	[85SooF]
Noyaux constitués d'unités phonétiques	[86HigA]
Quantification vectorielle différée	[90JuaB]
Quantification matricielle instantanée	[93CheM]
Modèles cachés de Markov	[90RosA]
Réseaux neuromimétiques	
Perceptron, 1 modèle par locuteur	[90OglJ]
1 modèle par sexe puis réseau global, multifenêtre	[91BenY]
Compétition du locuteur et du monde, Markov	[91CarM]
Décision en arbre	[91RudL]
Analyse par prédiction non linéaire	[92HatH]
Proéminence	[93ThéP]
Conformité	[93ThéP]

Apophtegme

Tout au long de ce travail, nous avons cherché à améliorer à la fois l'état de nos connaissances et la qualité de la reconnaissance de locuteurs. Nous l'avons fait parce que nous croyons que le Mieux est l'ennemi du Bien.

Dans ce combat, le Mieux l'emporte.

Toujours.

Et par définition.

Neuchâtel, mai 1993

Remerciements

En tout premier lieu, je tiens à remercier chaleureusement mes parents, non seulement pour m'avoir mis au monde, mais encore pour tout ce qu'ils m'ont permis de réaliser jusqu'à aujourd'hui ainsi que pour l'aide exceptionnelle qu'il m'ont accordée sous forme d'affection, d'encouragements et de soutien.

Je remercie encore mes coreligionnaires, qu'ailleurs on appelle collègues, endossant ici le sacerdoce qui d'un doctorat, qui d'une charge d'assistant, qui du maintient de l'infrastructure nécessaire à un tel travail. Tous m'ont aidé par leur entrain et par leurs compétences. En particulier, je remercie l'équipe du centre de calcul qui a su tolérer mes frasques et ma boulimie de ressources informatiques, ainsi que mon directeur de thèse dans son rôle multiple d'égide, de guide et de garde-fou.

Je remercie aussi tous ceux qui ont accepté de me livrer un peu de leur âme, sous la forme d'échantillons de parole. Ceux-là, j'espère les reconnaître à jamais...

Je remercie enfin les experts et j'applaudis leur acte de bravoure qui consiste à se lancer dans la lecture et l'examen de ce document.

Cette thèse a été menée à l'institut de microtechnique de l'université de Neuchâtel, en partie dans le cadre du projet AGEN-56 intitulé "Identification de la voix de l'usager dans des services vocaux à accès privé", sous la responsabilité du Dr. H. Hügli.

Bibliographie

Nous donnons ici une liste de documents ayant un rapport parfois proche, parfois lointain avec le sujet traité. La clef de tri utilisée est l'année de parution, suivie du nom abrégé de son auteur premier cité. L'initiale de son prénom ferme la marche du mot clef et permet d'éviter les confusions. Le corps du texte de cette thèse se réfère à ces documents par leur mot clef embrassé par des parenthèses carrées [—]. Les abréviations des titres de revues ou des actes de conférences sont communes et nous jugeons inutile de préciser ici leur développement intégral. Les titres de livres et de films cinématographiques sont guillemetés «—» et les titres d'articles sont donnés entre doubles apostrophes "—". Nos propres références ferment cette liste.

[62KerL] L. G. Kersta, *"Voiceprint Identification"*, Nature, Vol. 196, 1962, pp. 1253-1257

[63ComA] A. J. Compton, *"Effects of Filtering and Vocal Duration upon the Identification of Speakers, Aurally"*, J. Acoust. Soc. Am., Vol. 35, Nº 11, 1963, pp. 1748-1752

[64VolW] W. D. Voiers, *"Perceptual Bases of Speaker Identity"*, J. Acoust. Soc. Am., Vol. 36, Nº 6, 1964, pp. 1065-1073

[68KubS] S. Kubrick, A. C. Clarke, *«2001: A Space Odissey»*, Metro Goldwin Meyer, 1968, 141'

[68OppA] A. V. Oppenheim, R. W. Schafer, *"Homomorphic Analysis of Speech"*, IEEE Trans. Audio and Electroacoustics, Vol. 16, Nº 2, 1968, pp. 221-226

[68SteK] K. N. Stevens, *"Speaker Authentification and Identification: A Comparison of Spectrographic and Auditory Presentations of Speech Material"*, J. Acoust. Soc. Am., Vol. 44, Nº ?, 1968, pp. 1596-1607

[70FanG] G. Fant, *«Acoustic Theory of Speech Production»*, Mouton, 1970, 328 p.

[72FlaJ] J. L. Flanagan, *«Speech Analysis Synthesis and Perception»*, Springer-Verlag, 1972, 444 p.

[72FukK] K. Fukunaga, *«Introduction to Statistical Pattern Recognition»*, Academic Press, 1972, 369 p.

[72FurS] S. Furui, F. Itakura, S. Saito, *"Talker Recognition by the Longtime Averaged Speech Spectrum"*, Trans. IECE, Vol. 55, Nº 9, 1972, pp. 11-13

[72WolJ] J. J. Wolf, *"Efficient Acoustic Parameters for Speaker Recognition"*, J. Acoust. Soc. Am., Vol. 51, Nº 6, Part 2, 1972, pp. 2044-2056

[73LumR] R. C. Lummis, *"Speaker Verification by Computer Using Speech for Temporal Registration"*, IEEE Trans. Audio and Electroacoustics, Vol. 21, Nº 2, 1973, pp. 80-89

[73MatH] H. Matsumoto, S. Hiki, T. Sone, T. Nimura, *"Multidimensional Representation of Personal Quality of Vowels and its Acoustical Correlates"*, IEEE Trans. Audio and Electroacoustics, Vol. 21, Nº 5, 1973, pp. 428-436

[74FurS] S. Furui, *"An Analysis of Long-Term Variation of Feature Parameters of Speech and its Application to Talker Recognition"*, Trans. IECE, Vol. 57, Nº 1, 1974, pp. 9-10

[75FanG] G. Fant, M. A. A. Tatham, *«Auditory Analysis and Perception of Speech»*, Academic Press, 1975, 564 p.

[75MakJ] J. Makhoul, *"Linear Prediction: A Tutorial Review"*, Proc. IEEE, Vol. 63, N° 4, 1975, pp. 561-580

[75RosA] A. E. Rosenberg, M. R. Sambur, *"New Techniques for Automatic Speaker Verification"*, IEEE Trans. ASSP, Vol. 23, N° 2, 1975, pp. 169-176

[76AtaB] B. S. Atal, *"Automatic Recognition of Speakers from Their Voices"*, Proc. IEEE, Vol. 64, N° 4, 1976, pp. 460-475

[76RosA] A. E. Rosenberg, *"Automatic Speaker Verification: A Review"*, Proc. IEEE, Vol. 64, N° 4, 1976, pp. 475-486

[76MarJ] J. D. Markel, A. H. Gray, *«Linear Prediction of Speech»*, Communication and Cybernetics 12, Springer-Verlag, 1976, 288 p.

[76SamM] M. R. Sambur, *"Speaker Recognition Using Orthogonal Linear Prediction"*, IEEE Trans. ASSP, Vol. 24, N° 4, 1976, pp. 283-289

[77MarJ] J. D. Markel, B. T. Oshika, A. H. Gray, *"Long-Term Feature Averaging for Speaker Recognition"*, IEEE Trans. ASSP, Vol. 25, N° 4, 1977, pp. 330-337

[77MonR] R. B. Monsen, A. M. Engebretson, *"Study of Variations in the Male and Female Glottal Wave"*, J. Acoust. Soc. Am., Vol. 62, N° 4, 1977, pp. 981-993

[78JesP] P. Jesorsky, *"Principles of Automatic Speaker Recognition"*, Speech Communication with Computers, Carl Hanser Verlag, 1978, pp. 93-137

[78OppA] A. V. Oppenheim, *«Applications of Digital Signal Processing»*, Prentice-Hall, 1978, 499 p.

[78RabL] L. R. Rabiner, R. W. Schafer, *«Digital Processing of Speech Signals»*, Prentice-Hall, 1978, 512 p.

[79BunE] E. Bunge, *"Forensic Voice Identification by Computers"*, Int. Crim. Pol. Rev., N° 34, 1979, pp. 254-270

[79MarJ] J. D. Markel, S. B. Davis, *"Text-Independent Speaker Recognition from a Large Linguistically Unconstrained Time-Spaced Data Base"*, IEEE Trans. ASSP, Vol. 27, N° 1, 1979, pp. 74-82

[79WarG] G. H. Warfel, *«Identification Technologies»*, Thomas Books, 1979, 188 p.

[80BuzA] A. Buzo, A. H. Gray Jr., R. M. Gray, J. D. Markel, *"Speech Coding Based Upon Vector Quantization"*, IEEE Trans. ASSP, Vol. 28, N° 5, 1980, pp. 562-574

[80FurS] S. Furui, A. E. Rosenberg, *"Experimental Studies in a New Automatic Speaker Verification System Using Telephone Speech"*, Proc. ICASSP, 1980, pp. 1060-1062

[80LinY] Y. Linde, A. Buzo, R. M. Gray, *"An Algorithm for Vector Quantization Design"*, IEEE Trans. Communications, Vol. 28, N° 1, 1980, pp. 84-95

[81CorP] P. Corsi, *"Speaker Recognition: A Survey"*, Proc. Second NATO Advanced Study Institute on Speech Processing, Bonas, 1981, pp. 277-308

[81FurS1] S. Furui, *"Cepstral Analysis Technique for Automatic Speaker Verification"*, IEEE Trans. ASSP, Vol. 29, N° 2, 1981, pp. 254-272

[81FurS2] S. Furui, *"Comparison of Speaker Recognition Methods Using Statistical Features and Dynamic Features"*, IEEE Trans. ASSP, Vol. 29, N° 3, 1981, pp. 342-350

[81ShrM] M. Shridhar, N. Mohankrishnan, M. Baraniecki, *"Text-Independent Speaker Recognition Using Orthogonal Linear Prediction"*, Proc. ICASSP, Atlanta, 1981, pp. 197-200

[82AtaB] B. S. Atal, J. R. Remde, *"A New Model of LPC Excitation for Producing Natural-Sounding Speech at Low Bit Rates"*, Proc. ICASSP, Paris, 1982, pp. 614-617

[82DidE] E. Diday, J. Lemaire, J. Pouget, F. Testu, *«Éléments d'analyse de données»*, Bordas, Paris, 1982, 462 p.

[82MohN] N. Mohankrishnan, M. Shridhar, M. A. Sid-Ahmed, *"A Composite Scheme for Text-Independent Speaker Recognition"*, Proc. ICASSP, Paris, 1982, pp. 1653-1656

[82NeyH] H. Ney, R. Gierloff, *"Speaker Recognition Using a Feature Weighting Technique"*, Proc. ICASSP, Paris, 1982, pp. 1645-1648

[82NusH] H. J. Nussbaumer, *«Fast Fourier Transform and Convolution Algorithms»*, Springer-Verlag, 1982, 276 p.

[82SchR] R. Schwartz, S. Roucos, M. Berouti, *"The Application of Probability Density Estimation to Text-Independent Speaker Identification"*, Proc. ICASSP, Paris, 1982, pp. 1649-1652

[82ShrM] M. Shridhar, N. Mohankrishnan, *"Text-Independent Speaker Recognition: A Review and Some New Results"*, Speech Comm., Vol. 1, Nº 3-4, 1982, pp. 257-267

[83CoxR] R. V. Cox, R. E. Crochiere, J. D. Johnston *"Real-Time Implementation of Time Domain Harmonic Scaling of Speech for Rate Modification and Coding"*, IEEE Trans. ASSP, Vol. 31, Nº 1, 1983, pp. 258-272

[83FerM] M. Ferreti, F. Cinare, *«Synthèse, reconnaissance de la parole»*, Editests, 1983, 282 p.

[83HunM] M. J. Hunt, *"Further Experiments in Text-Independent Speaker Recognition Over Communication Channels"*, Proc. ICASSP, Boston, 1983, pp. 563-566

[83LiK] K. P. Li, E. H. Wrench Jr., *"An Approach to Text-Independent Speaker Recognition with Short Utterances"*, Proc. ICASSP, Boston, 1983, pp. 555-558

[83ShrM] M. Shridhar, N. Mohankrishnan, M. A. Sid-Ahmed, *"A Comparison of Distance Measures for Text-Independent Speaker Identification"*, Proc. ICASSP, Boston, 1983, pp. 559-562

[84BerM] M. Berouti, H. Garten, P. Kabal, P. Mermelstein, *"Efficient Computation and Encoding of the Multipulse Excitation for LPC"*, Proc. ICASSP, 1984, pp. 10.1.1-10.1.4

[84CouF] F. de Coulon, *«Théorie et traitement des signaux»*, Traité d'électricité Vol. VI, J. Neirynck, Presses Polytechniques Romandes, Lausanne, 1984, 548 p.

[84JayN] N. S. Jayant, P. Noll, *«Digital Coding of Waveforms»*, Prentice-Hall, 1984, 688 p.

[84KraM] M. Krasner, J. Wolf, K. Karnofsky, R. Schwartz, S. Roucos, H. Gish, *"Investigation of Text-Independent Speaker Identification Techniques under Conditions of Variable Data"*, Proc. ICASSP, 1984, pp. 18B.5.1-18B.5.4

[84KunM] M. Kunt, *«Traitement numérique des signaux»*, Traité d'électricité Vol. XX, J. Neirynck, Presses Polytechniques Romandes, Lausanne, 1984, 402 p.

[84PapP] P. E. Papamichalis, G. R. Doddington, *"A Speaker Recognizability Test"*, Proc. ICASSP, 1984, pp. 18B.6.1-18B.6.4

[84SinS] S. Singhal, B. S. Atal, *"Improving Performance of Multi-Pulse LPC Coders at Low Bit Rates"*, Proc. ICASSP, 1984, pp. 1.3.1-1.3.4

[85DodG] G. R. Doddington, *"Speaker Recognition—Identifying People by their Voices"*, Proc. IEEE, Vol. 73, Nº 11, 1985, pp. 1651-1664

[85GisH] H. Gish, K. Karnofsky, M. Krasner, S. Roucos, R. Schwartz, J. Wolf, *"Investigation of Text-Independent Speaker Identification Over Telephone Channels"*, Proc. ICASSP, Tampa, 1985, pp. 379-382

[85HolH] H. Hollien, *"Natural Speech Vectors for Speaker Identification"*, Speech Tech'85, pp. 331-334

[85MakJ] J. Makhoul, S. Roucos, H. Gish, *"Vector Quantization in Speech Coding"*, Proc. IEEE, Vol. 73, Nº 11, 1985, pp. 1551-1588

[85MaxJ] J. Max, L. Audaire, D. Berthier, R. Bigret, J.-C. Carré, H. Chevalier, B. Escudié, A. Hellion, J.-L. Lacoume, M. Martin, R. Miquel, P. Peltié, M. Trottot, S. Valette, R.Vergne, *«Méthodes et techniques de traitement du signal et applications aux mesures physiques»*, Masson, 1985, 354 p.

[85MokA] A. Mokeddem, *«Analyse de la parole: reconnaissance multilocuteur de mots isolés pour les systèmes miniaturisés»*, Thèse présentée à la faculté des sciences de Neuchâtel, Imprimerie Centrale Neuchâtel, 1985, 175 p.

[85SooF] F. K. Soong, A. E. Rosenberg, L. R. Rabiner, B. H. Juang, *"A Vector Quantization Approach to Speaker Recognition"*, Proc. ICASSP, Tampa, 1985, pp. 387-390

[85TosO] O. Tosi, *"Voice Identification Legal Application"*, Speech Tech'85, pp. 335

[85WolJ] J. J. Wolf, H. Gish, K. Karnofsky, M. Krasner, S. Roucos, R. Schwartz, *"Text-Independent Speaker Identification under Variable Conditions"*, Speech Tech'85, pp. 327-330

[86AtaB] B. S. Atal, *"High-Quality Speech at Low Bit Rates: Multi-Pulse and Stochastically Excited Linear Predictive Coders"*, Proc. ICASSP, Tokyo, 1986, pp. 1681-1684

[86BerM] M. Berouti, J. Jachner, D. Sloan, P. Mermelstein, *"Reducing Signal Delay in Multipulse Coding at 16Kb/s"*, Proc. ICASSP, Tokyo, 1986, pp. 3043-3046

[86BirM1] M. Birnbaum, R. W. Bossemeyer, L. A. Cohen, F. X. Welsh, *"Using Cepstral Features in Speaker Verification"*, Speech Tech'86, pp. 287-290

[86BirM2] M. Birnbaum, L. A. Cohen, F. X. Welsh, *"A Voice Password System for Access Security"*, AT&T Technical Journal, Vol. 65, N° 5, 1986, pp. 68-74

[86CamJ] J. P. Campbell, T. E. Tremain, *"Voiced/Unvoiced Classification of Speech with Application to the U.S. Government LPC-10E Algorithm"*, Proc. ICASSP, Tokyo, 1986, pp. 473-476

[86DavG] G. Davidson, A. Gersho, *"Complexity Reduction Methods for Vector Excitation Coding"*, Proc. ICASSP, Tokyo, 1986, pp. 3055-3058

[86FurS] S. Furui, *"Research on Individuality Features in Speech Waves and Automatic Speaker Recognition Techniques"*, Speech Comm., Vol. 5, N° 2, 1986, pp. 183-197

[86GisH] H. Gish, M. Krasner, W. Russel, J. Wolf, *"Methods and Experiments for Text-Independent Speaker Recognition Over Telephone Channels"*, Proc. ICASSP, Tokyo, 1986, pp. 865-868

[86HigA] A. L. Higgins, R. E. Wohlford, *"A New Method of Text-Independent Speaker Recognition"*, Proc. ICASSP, Tokyo, 1986, pp. 869-872

[86NaiJ] J. M. Naik, G. R. Doddington, *"High Performance Speaker Verification Using Principal Components"*, Proc. ICASSP, Tokyo, 1986, pp. 881-884

[86OzaK] K. Ozawa, T. Araseki, *"Low Bit Rate Multi-Pulse Speech Coder with Natural Speech Quality"*, Proc. ICASSP, Tokyo, 1986, pp. 457-460

[86RosA] A. E. Rosenberg, F. K. Soong, *"Evaluation of a Vector Quantization Talker Recognition System in Text-Independent and Text-Dependent Modes"*, Proc. ICASSP, Tokyo, 1986, pp. 873-876

[86RosM] M. Rossi, *«Électroacoustique»*, Traité d'électricité Vol. XXI, J. Neirynck, Presses Polytechniques Romandes, Lausanne, 1986, 561 p.

[86SooF] F. K. Soong, A. E. Rosenberg, *"On the Use of Instantaneous and Transitional Spectral Information in Speaker Recognition"*, Proc. ICASSP, Tokyo, 1986, pp. 877-880

[86VerW] W. Verhelst, O. Steenhaut, *"A New Model for the Short-Time Complex Cepstrum of Voiced Speech"*, IEEE Trans. ASSP, Vol. 34, N° 1, 1986, pp. 43-51

[87ElID] D. F. Elliott, *«Handbook of Digital Signal Processing, Engineering Applications»*, Academic Press, 1987, 999 p.

[87ShaD] D. O'Shaughnessy, *«Speech Communication (Human and Machine)»*, Addison-Wesley, 1987, 568 p.

[88AtiJ] J. B. Atili, M. Savic, J. P. Campbell Jr., *"A TMS3220-Based Real Time, Text-Independent, Automatic Speaker Verification System"*, Proc. ICASSP, New York City, 1988, pp. 599-602

[88BigB] B. Bigelow, *"Speaker Verification Over the Telephone: A Technical Perspective of Practical Applications"*, Speech Tech'88, pp. 119-122

[88ChiD] D. G. Childers, K. Wu, K. S. Bae, D. M. Hicks, *"Automatic Recognition of Gender by Voice"*, Proc. ICASSP, New York City, 1988, pp. 603-606

[88LiK] K. P. Li, J. E. Porter, *"Normalizations and Selection of Speech Segments for Speaker Recognition Scoring"*, Proc. ICASSP, New York City, 1988, pp. 595-598

[88NakH] H. Nakasone, C. Melvin, *"Computer Assisted Voice Identification System"*, Proc. ICASSP, New York City, 1988, pp. 587-590

[88NodH] H. Noda, *"Frequency-Warped Spectral Distance Measures for Speaker Verification in Noise"*, Proc. ICASSP, New York City, 1988, pp. 576-579

[88SooF] F. K. Soong, A. E. Rosenberg, *"On the Use of Instantaneous and Transitional Spectral Information in Speaker Recognition"*, IEEE Trans. ASSP, Vol. 36, Nº 6, 1988, pp. 871-879

[88VelG] G. Velius, *"Variants of Cepstrum Based Speaker Identity Verification"*, Proc. ICASSP, New York City, 1988, pp. 583-586

[88WilJ] J. Willbur, F. J. Taylor, *"Consistent Speaker Identification via Wigner Smoothing Techniques"*, Proc. ICASSP, New York City, 1988, pp. 591-594

[88ZheY] Y.-C. Zheng, B.-Z. Yuan, *"Text-Dependent Speaker Identification Using Circular Hidden Markov Models"*, Proc. ICASSP, New York City, 1988, pp. 580-582

[89BasM] M. Basseville, *"Distance Measures for Signal Processing and Pattern Recognition"*, Signal Processing, Nº 18, 1989, pp. 349-369

[89FedA] A. Federico, G. Ibba, A. Paolini, N. De Sario, B. Saverione, *"Comparison between Automatic Methods and Human Listeners in Speaker Recognition Tasks"*, Eurospeech, Paris, 1989, Vol. 1, pp. 279-282

[89FeuT] T. C. Feustel, G. A. Velius, R. J. Logan, *"Human and Machine Performance on Speaker Identity Verification"*, Speech Tech'89, pp. 169-170

[89FitK] K. Fitzgerald, *"The Quest for Intruder-Proof Computer Systems"*, IEEE Spectrum, August 1989, pp. 22-26

[89GiaA] A. Giannini, M. Pettorino, U. Cinque, *"Speaker's Identification by Voice"*, Eurospeech, Paris, 1989, Vol. 1, pp. 283-286

[89GreS] S. Green, T. C. Feustel, G. A. Velius, *"Field-Accessible Training Secured by Speaker Identity Verification"*, Speech Tech'89, pp. 166-168

[89MasJ] J. S. Mason, J. Oglesby, L. Xu, *"Codebooks to Optimise Speaker Recognition"*, Eurospeech, Paris, 1989, Vol. 1, pp. 267-270

[89NajJ] J. M. Naik, L. P. Netsch, G. R. Doddington, *"Speaker Verification Over Long Distance Telephone Lines"*, Proc. ICASSP, Glasgow, 1989, pp. 524-527

[89NodH] H. Noda, *"On the Use of the Information on Individual Speaker's Position in the Parameter Space for Speaker Recognition"*, Proc. ICASSP, Glasgow, 1989, pp. 516-519

[89ProJ] J. G. Proakis, D. G. Manolakis, *«Introduction to Digital Signal Processing»*, Macmillan, 1989, 944 p.

[89RocC] C. Rocchi, E. Mumolo, *"A New Method for Performing Weighted Distances for Speaker Authentification"*, Eurospeech, Paris, 1989, Vol. 1, pp. 275-278

[89RogC] C. Rogers, D. Chien, M. Featherston, K. Min, *"Neural Network Enhancement for a Two Speaker Separation System"*, Proc. ICASSP, Glasgow, 1989, pp. 357-360

[89SteM] M. C. Steckner, D. J. Drost, *"Fast Cepstrum Analysis Using the Hartley Transform"*, IEEE Trans. ASSP, Vol. 37, Nº 8, 1989, pp. 1300-1302

[89XuL1] L. Xu, J. S. Mason, *"Instantaneous and Transitional Perceptually-Based Features in Speaker Identification"*, Eurospeech, Paris, 1989, Vol. 1, pp. 271-274

[89XuL2] L. Xu, J. Oglesby, J. S. Mason, *"The Optimization of Perceptually-Based Features for Speaker Identification"*, Proc. ICASSP, Glasgow, 1989, pp. 520-523

[89YonG] G. Yong, J. S. Mason, *"Speaker Normalization via a Linear Transformation on a Perceptual Feature Space and its Benefits in ASR Adaptation"*, Eurospeech, Paris, 1989, Vol. 1, pp. 258-261

[89ZalJ] J. Zalewski, *"Text-Dependent Speaker Recognition in Noise"*, Eurospeech, Paris, 1989, Vol. 1, pp. 287-289

[90BenY] Y. Bennani, F. Fogelman Soulie, P. Gallinari, *"A Connectionist Approach for Automatic Speaker Identification"*, Proc. ICASSP, Albuquerque, 1990, pp. 265-268

[90BoiR] R. Boîte, *"La reconnaissance de la parole et la vérification du locuteur"*, AGEN Mitteilungen, N° 52, 1990, pp. 5-13

[90WheC] C. Wheddon, R. Linggard, *«Speech and Language Processing»*, Chapman and Hall, 1990, 339 p.

[90EatJ] J. Eatock, J. S. Mason, *"Speaker-Dependent Classification in Speaker Recognition"*, ESCA Proc. Speaker Characterization in Speech Technology, Edimburgh, 1990, pp. 94-97

[90FurS] S. Furui, *"Speaker-Dependent Feature Extraction, Recognition and Processing Techniques"*, ESCA Proc. Speaker Characterization in Speech Technology, Edimburgh, 1990, pp. 10-27

[90JuaB] B.-H. Juang, F. K. Soong, *"Speaker Recognition based on Source Coding Approaches"*, Proc. ICASSP, Albuquerque, 1990, pp. 613-616

[90LhoE] E. Lhote, L. Abou Haidar, *"Speaker Verification by a Vocal Proxemy Cue"*, ESCA Proc. Speaker Characterization in Speech Technology, Edimburgh, 1990, pp. 149-154

[90MatP] P. Maturi, *"Speaker Identification in Forensics: A Simulation Experiment"*, ESCA Proc. Speaker Characterization in Speech Technology, Edimburgh, 1990, pp. 155-160

[90MonA] A. I. C. Monaghan, D. R. Ladd, *"Speaker-Dependent and Speaker-Independent Parameters in Intonation"*, ESCA Proc. Speaker Characterization in Speech Technology, Edimburgh, 1990, pp. 167-174

[90NeuP] P. J. Neufeld, N. Colman, *"When Science Takes the Witness Stand"*, Scientific American, Vol. 262, N° 5, 1990, pp. 18-25

[90OglJ] J. Oglesby, J. S. Mason, *"Optimization of Neural Models for Speaker Identification"*, Proc. ICASSP, Albuquerque, 1990, pp. 261-264

[90PfiB] B. Pfister, *"Sprecherverifikation mit einem neuronalen Netz"*, AGEN Mitteilungen, N° 52, 1990, pp. 47-52

[90RosA] A. E. Rosenberg, C.-H. Lee, F. K. Soong, *"Sub-Word Unit Talker Verification Using Hidden Markov Models"*, Proc. ICASSP, Albuquerque, 1990, pp. 269-272

[90SavM] M. Savic, S. K. Gupta, *"Variable Parameter Speaker Verification System based on Hidden Markov Modeling"*, Proc. ICASSP, Albuquerque, 1990, pp. 281-284

[90SkvJ] J. Skvarc, M. Miletic, *"Speaker Sex Estimation"*, ESCA Proc. Speaker Characterization in Speech Technology, Edimburgh, 1990, pp. 181-186

[91BasC] C. Basztura, *"Experiments of Automatic Speaker Recognition in Open Sets"*, Speech Comm., Vol. 10, N° 2, 1991, pp. 117-127

[91BenY] Y. Bennani, P. Gallinari, *"On the Use of TDNN-Extracted Features Information in Talker Identification"*, Proc. ICASSP, Toronto, 1991, pp. 385-388

[91CarM] M. J. Carey, E. S. Parris, J. S. Bridle, *"A Speaker Verification System Using Alpha-Nets"*, Proc. ICASSP, Toronto, 1991, pp. 397-400

[91FusJ] J. W. Fussel, *"Automatic Sex Identification from Short Segments of Speech"*, Proc. ICASSP, Toronto, 1991, pp. 409-412

[91GagD] D. A. Gaganelis, E. D. Frangoulis, *"A Novel Approach to Speaker Verification"*, Proc. ICASSP, Toronto, 1991, pp. 373-376

[91HigA1] A. L. Higgins, L. Bahler, J. Porter, *"Speaker Verification Using Randomized Phrase Prompting"*, Digital Signal Processing, N° 1, 1991, pp. 89-106

[91HigA2] A. L. Higgins, L. G. Bahler, *"Text-Independent Speaker Verification by Discriminator Counting"*, Proc. ICASSP, Toronto, 1991, pp. 405-408

[91KreJ] J. Kreiman, G. Papcun, *"Comparing Discrimination and Recognition of Unfamiliar Voices"*, Speech Comm., Vol. 10, N° 3, 1991, pp. 265-275

[91MatT] T. Matsui, S. Furui, *"A Text-Independent Speaker Recognition Method Robust Against Utterance Variations"*, Proc. ICASSP, Toronto, 1991, pp. 377-380

[91OglJ] J. Oglesby, J. S. Mason, *"Radial Basis Function Networks for Speaker Recognition"*, Proc. ICASSP, Toronto, 1991, pp. 393-396

[91RosA] A. E. Rosenberg, C.-H. Lee, S. Gokcen, *"Connected Word Talker Verification Using Whole Word Hidden Markov Model"*, Proc. ICASSP, Toronto, 1991, pp. 381-384

[91RosR] R. C. Rose, J. Fitzmaurice, E. M. Hofstetter, D. A. Reynolds, *"Robust Speaker Identification in Noisy Environments Using Noise Adaptive Speaker Models"*, Proc. ICASSP, Toronto, 1991, pp. 401-404

[91RudL] L. Rudasi, S. A. Zahorian, *"Text-Independent Talker Identification with Neural Networks"*, Proc. ICASSP, Toronto, 1991, pp. 389-392

[91SavM] M. Savic, I.-H. Nam, *"Voice Personality Transformation"*, Digital Signal Processing, Nº 1, 1991, pp. 107-110

[91WeiS] S. M. Weiss, C. A. Kulikowski, *«Computer Systems That Learn»*, Morgan Kaufmann, 1991, 223 p.

[92ChaH] H. M. Chang, *"Augmented Phonetic Map for Voice Verification"*, Proc. ICASSP, San Francisco, 1992, pp. II.169-II.172

[92GonY] Y. Gong, J.-P. Haton, *"Non-Linear Vectorial Interpolation for Speaker Recognition"*, Proc. ICASSP, San Francisco, 1992, pp. II.173-II.176

[92HatH] H. Hattori, *"Text-Independent Speaker Recognition Using Neural Networks"*, Proc. ICASSP, San Francisco, 1992, pp. II.153-II.156

[92KaoY] Y.-H. Kao, P. K. Rajasekaran, J. S. Baras, *"Free-Text Identification Over Long Distance Telephone Channel Using Hypothesized Phonetic Segmentation"*, Proc. ICASSP, San Francisco, 1992, pp. II.177-II.180

[92MatT] T. Matsui, S. Furui, *"Comparison of Text-Independent Speaker Recognition Methods Using VQ-Distortion and Discrete/Continuous HMMS"*, Proc. ICASSP, San Francisco, 1992, pp. II.157-II.160

[92NetL] L. P. Netsch, G. R. Doddington, *"Speaker Verification Using Temporal Decorrelation Post-Processing"*, Proc. ICASSP, San Francisco, 1992, pp. II.181-II.184

[92PakM] M. R. Pakravan, *"A New Way for Implementing a Speaker Identification Systems"*, Int. Conf. Signal Processing Applications and Technology, Boston, 1992, pp. 1035-1041

[92ReyD1] D. A. Reynolds, R. C. Rose, M. J. T. Smith, *"PC-Based TMS320C30 Implementation of the Gaussian Mixture Model Text-Independent Speaker Recognition System"*, Int. Conf. Signal Processing Applications and Technology, Boston, 1992, pp. 967-973

[92ReyD2] D. Reynolds, R. C. Rose, *"An Integrated Speech-Background Model for Robust Speaker Identification"*, Proc. ICASSP, San Francisco, 1992, pp. II.185-II.188

[92SavM] M. Savic, J. Sorensen, *"Phoneme Based Speaker Verification"*, Proc. ICASSP, San Francisco, 1992, pp. II.165-II.168

[92SiuM] M.-H. Siu, G. Yu, H. Gish, *"An Unsupervised, Sequential Learning Algorithm for the Segmentation of Speech Waveforms with Multiple Speakers"*, Proc. ICASSP, San Francisco, 1992, pp. II.189-II.192

[92TscB] B. L. Tseng, F. K. Soong, A. E. Rosenberg, *"Continuous Probabilistic Acoustic Map for Speaker Recognition"*, Proc. ICASSP, San Francisco, 1992, pp. II.161-II.164

[93CheM] M.-S. Chen, P.-H. Lin, H.-S. Wang, *"Speaker Identification Based on a Matrix Quantization Method"*, IEEE Trans. ASSP, Vol. 41, Nº 1, 1993, pp. 398-403

[90ThéP1] P. Thévenaz, H. Hügli, *"Combining Four Text-Independent Speaker Recognition Methods"*, ESCA Proc. Speaker Characterization in Speech Technology, 1990, Edimburgh, pp. 187-191

[90ThéP2] P. Thévenaz, *"Reconnaissance de locuteurs indépendante du texte"*, AGEN Mitteilungen, Nº 52, 1990, pp. 35-45

[92ThéP] P. Thévenaz, H. Hügli, *"A Residue-Based Approach to Text-Independent Speaker Recognition"*, Proc. First Swiss Symposium on Pattern Recognition and Computer Vision, Lausanne, 1992, pp. 35-41