Etat actuel
de la recherche en reconnaissance du locuteur
et des applications en criminalistique

Olivier CAPPÉ
cappe@sig.enst.fr

Ecole Nationale Superieure des Telecommunications,
département Signal

janvier 1995

Table des Matières

Avant propos

Ce rapport bibliographique a pour but, à la fois, de fournir une vue d'ensemble sur les recherches concernant les techniques de reconnaissance du locuteur, et de décrire la manière dont est appliquée la reconnaissance du locuteur dans un domaine particulier, celui de la criminalistique. Ces deux aspects constituent des parties relativement distinctes du rapport qui peuvent être lues séparément.

Le rapport débute par un premier chapitre qui constitue une introduction générale au domaine de la reconnaissance du locuteur, notamment pour tout ce qui concerne la terminologie. Le chapitre 2 fournit une description aussi détaillée que possible des caractéristiques des différentes techniques de reconnaissance du locuteur étudiées dans la littérature, sans pour autant aller jusqu'à une description complète des algorithmes. Le chapitre 3 tente de rendre compte des nombreux problèmes spécifiques posés par la reconnaissance du locuteur dans le cadre judiciaire, tout en faisant le point sur les méthodes utilisées à l'heure actuelle. Les techniques évoquées dans ce chapitre sont parfois décrites de manière peu précise car les publications dans le domaine sont plus rares[1], et souvent, peu exploitables (d'une manière générale, les travaux menés en criminalistique sont très appliqués et, le plus souvent, à vocation confidentielle).

En ce qui concerne le chapitre 2, tout en veillant à présenter de manière (à peu près) exhaustive les techniques de reconnaissance décrites dans les publications citées en bibliographie, j'ai privilégié les aspects liés à la vérification en mode indépendant du texte ainsi qu'à la robustesse vis à vis des variations du canal de transmission. Par ailleurs, j'ai préféré ne pas évoquer les techniques pour lesquelles je disposais de très peu de références bibliographiques (celles-ci sont simplement mentionnées pour mémoire au paragraphe 2.2.2.g).

Les références bibliographique citées dans ce document ont été obtenues par une recherche relativement systématique à partir de la base de données INSPEC (sur les termes *speaker recognition* et *speaker identification*). On peut donc espérer que la bibliographie soit raisonnablement complète, au moins en ce qui concerne les articles de revue antérieurs à 1993. Pour les articles de congrès, la bibliographie est certainement moins représentative car ils sont d'une manière générale moins bien répertoriés dans les bases de données. Il manque aussi sûrement de nombreuses références européennes ou japonaises, celles-ci étant peu représentées dans la base de données INSPEC.

Dans la version actuelle du rapport (janvier 1995), j'ai essayé de tenir à jour les références bibliographiques, au moins en ce qui concerne les articles de revue parus en 1994, en y incluant [Assaleh 94], [Gish 94], [Quatieri 94], [Reynolds 94a] et [Sankar 94].

Pour aborder le domaine de la reconnaissance du locuteur, on peut conseiller les deux ar-

[1] Ceci est particulièrement vrai en France, où contrairement à ce qui se passe dans plusieurs autres pays, ces questions dites de *forensic speaker identification* ou de *forensic phonetics* font l'objet d'assez peu d'études scientifiques. Par contre, comme dans tous les autres pays où ce type de considérations est pris en compte dans le cadre d'expertises judiciaires, ces questions ne manquent pas de susciter périodiquement des controverses justifiées eu égard à la gravité des conséquences de l'expertise dans ce contexte.

ticles [Doddington 85] et [OShaughnessy 86] qui offrent une vision globale du domaine, ainsi que [Furui 94] et [Rosenberg 91] qui font le point sur plusieurs développements techniques très récents. Pour se faire une idée plus complète sur les recherches en cours, il est intéressant de consulter les actes du congrès [ESCA 94]. On trouve par ailleurs de nombreuses informations concernant les modèles utilisés en reconnaissance du locuteur dans [Rabiner 93] (bien que cet ouvrage soit consacré plus spécifiquement à la reconnaissance de la parole). Enfin, les publications les plus complètes sur les applications dans le domaine de la criminalistique sont [Hollien 90] et [Kunzel 94].

Chapitre 1

Introduction à la reconnaissance du locuteur

1.1 Caractérisation des applications de reconnaissance du locuteur

1.1.1 Reconnaissance, vérification et identification

Traditionnellement, le terme de **reconnaissance du locuteur** (*speaker recognition*) désigne de manière générique toutes les applications où l'on cherche à obtenir des renseignements concernant l'identité d'une personne à partir d'un enregistrement de sa voix [Atal 76], [Calliope 89, Chap. XIX], [Doddington 85], [OShaughnessy 86], [Rosenberg 91]. Pour qualifier plus précisément les différentes applications entrant dans le cadre de la reconnaissance du locuteur, on distingue en général deux types de tâches :

Vérification du locuteur Lorsque l'on cherche à décider si l'identité revendiquée par un locuteur est compatible avec sa voix. Dans ce type d'applications, il s'agit donc de trancher entre les deux hypothèses : soit le locuteur est bien le **locuteur autorisé**[1], c'est à dire celui dont l'identité est revendiquée, soit nous avons affaire à un **imposteur** qui cherche à se faire passer pour un locuteur autorisé. Les applications classiquement envisagées pour la vérification de locuteur correspondent donc à l'idée de "serrure vocale" qui peut être utilisée, par exemple, pour valider des transactions bancaires effectuées par téléphone, ou pour compléter un dispositif d'accès (à un bâtiment, un système informatique, etc.).

Identification du locuteur Quand il s'agit de déterminer, parmi un ensemble de N locuteurs potentiels, à quel locuteur correspond un enregistrement vocal. En identification, la réponse apportée n'est plus de type binaire (acception ou rejet) comme dans le cas de la vérification puisqu'il est nécessaire de désigner un locuteur parmi un groupe. On distingue encore deux sous problèmes d'identification selon que l'on est sûr ou non du fait que l'enregistrement provient bien d'un des membres du groupe de locuteurs potentiels :

[1] Le terme de **client** est aussi souvent utilisé du fait de la nature commerciale de la plupart des applications envisagées.

- Si l'on a affaire à un **groupe fermé** (*closed-set*), c'est à dire si l'enregistrement provient forcément d'un des membres du groupe, il suffit de désigner un des N locuteurs du groupe.
- Dans le cas d'un **groupe ouvert** (*open-set*), la possibilité que le locuteur ne figure pas dans le groupe des locuteurs connus est prise en compte. Dans ce cas, il existe une alternative supplémentaire : soit on désigne un des locuteurs connus membres du groupe, soit on décide qu'il s'agit d'un autre locuteur, extérieur au groupe.

Notons que l'identification avec un groupe ouvert est plus difficile qu'avec un groupe fermé, puisqu'il faut prendre en compte l'existence d'imposteurs. D'autre part, la vérification peut être considérée comme un cas particulier d'identification avec groupe ouvert pour lequel la taille du groupe est réduite à $N = 1$ locuteur.

Pour une application de reconnaissance ou d'identification, il est nécessaire de disposer d'une base de données contenant des enregistrements de référence correspondant à chacun des locuteurs autorisés. En pratique, on ne conserve pour chaque locuteur que les paramètres utiles pour la reconnaissance extraits de (ou de ses) enregistrement(s) de référence. Ces informations constituent les **données de référence** (*reference templates*) du locuteur. L'étape préliminaire qui consiste à bâtir les données de référence propres à chaque locuteur est appelée phase d'**apprentissage**.

1.1.2 Remarque concernant les termes d'identification et de vérification

La présentation adoptée ci-dessus correspond à l'acceptation de loin la plus courante des termes d'identification et de vérification. En particulier, ces termes doivent être systématiquement interprétés dans ce sens lorsqu'ils sont employés dans le milieu du traitement du signal et de la parole. Toutefois, dans certaines références écrites par des auteurs plutôt issus du milieu des phonéticiens, et notamment dans la plupart des publications traitant spécifiquement de l'application en criminalistique (par exemple [Federico 93], [Hollien 90] et [Kunzel 94]), ces termes sont utilisés dans un sens totalement opposé : la vérification désigne toutes les applications où le locuteur est **coopérant** et où l'enregistrement est réalisé dans de bonnes conditions. Ce qui inclut, en particulier, toutes les applications envisagées précédemment. Par contre, l'identification correspond aux cas où les conditions d'enregistrements ne sont pas maîtrisées et où le locuteur ne cherche pas à se faire connaître. Le terme d'identification désigne donc en particulier les applications en criminalistique [Hollien 90], [Kunzel 94].

Cette ambiguïté concernant les termes de vérification et d'identification peut être la source de confusions (particulièrement pour les publications consacrées la criminalistique). Pour notre part, nous nous tiendrons à la présentation effectué au paragraphe 1.1.1 qui reflète l'interprétation la plus courante. Nous aurons l'occasion de préciser la terminologie que nous utiliserons pour décrire les différentes applications en criminalistique au paragraphe 3.1.

1.1.3 Modes dépendant et indépendant du texte

On parle de reconnaissance de locuteur en mode **dépendant du texte** (*text-dependent* ou *fixed-text*) lorsque le texte prononcé par le locuteur est fixé et connu à l'avance. A l'opposé, lorsque le texte prononcé par le locuteur n'est pas connu a priori, on parle de mode **indépendant du texte** (*text-independent* ou *free-text*). Cette distinction entre deux modes de fonctionnement des applications de reconnaissance du locuteur est très importante car les techniques utilisées,

ainsi que les performances obtenues, dans les deux cas sont très différentes [Doddington 85], [OShaughnessy 86].

1.1.4 Caractéristiques et variabilité

A la base de toutes les techniques de reconnaissance de locuteur, on commence par mesurer des **caractéristiques** (*features*) du signal de parole [Atal 76], [Doddington 85], [Hollien 90], [OShaughnessy 86], [Sambur 75]. Ces caractéristiques doivent (idéalement) être faciles à déterminer à partir du signal de parole, robustes par rapport aux conditions d'enregistrement, et surtout, fournir le plus de renseignements possible concernant l'identité du locuteur. Nous aurons l'occasion de détailler plus loin les caractéristiques mesurées par les différents systèmes de reconnaissance du locuteur. Pour l'instant, il faut retenir qu'à l'heure actuelle on dispose de peu de connaissance concernant la nature des caractéristiques "optimales" pour la reconnaissance du locuteur. La principale raison en est que le signal de parole résulte d'un ensemble complexe de mécanismes très différents de par leur nature, et dont la plupart restent mal connus [Calliope 89].

Une question importante est celle de la **variabilité** des caractéristiques mesurées. On parle de **variabilité intra-locuteur** lorsque l'on s'intéresse à la variation des paramètres mesurés pour un même locuteur, et de **variabilité inter-locuteur** si on considère des locuteurs différents [Doddington 85], [Hollien 90], [Rosenberg 91]. Pour la reconnaissance de locuteur, on cherche à extraire des caractéristiques du signal de parole qui présentent une forte variabilité inter-locuteur (pour pouvoir différencier les locuteurs entre eux) et une faible variabilité intra-locuteur (pour garantir la robustesse du système).

1.2 Mesures de l'efficacité de la reconnaissance du locuteur

En général, l'efficacité d'une méthode de reconnaissance de locuteur est quantifiée par la donnée de la probabilité (ou taux) d'erreur. Plus cette dernière est faible, plus la technique utilisée pour la reconnaissance peut être considérée comme fiable. Toutefois, il est très important de distinguer plusieurs types d'erreurs différentes qui ne peuvent pas toujours être résumées par un chiffre unique. De plus, les mesures de fiabilité utilisées ne sont pas les mêmes selon que l'on a affaire à un problème de vérification ou d'identification [Calliope 89, Chap. XIX], [OShaughnessy 86], [Furui 81a], [Soong 88].

1.2.1 Pour la vérification

Dans une application de vérification du locuteur, la tâche consiste à trancher entre les deux hypothèses (1) "le locuteur correspond à l'identité revendiquée" (acceptation) et (2) "le locuteur est un imposteur" (rejet). On distingue donc deux types d'erreurs :

Taux de rejet erroné (sous entendu de locuteur autorisé) Qui est la probabilité de détecter un imposteur alors que l'on a affaire au véritable locuteur. En général, cette probabilité d'erreur est désignée par les termes de *false rejection rate* et parfois de *type I error rate* (erreur de première espèce, terme utilisé dans la théorie de la décision), ou de *false alarm rate* (taux de fausses alarmes, ce qui correspond à la terminologie en usage dans les problèmes de détection).

Taux d'acceptation erronée (sous entendu d'imposteur) Qui est la probabilité d'accepter le locuteur alors que l'on a affaire à un imposteur (*false acception rate*, *type II error rate* ou *miss rate*).

En général, l'attitude choisie consiste à régler les paramètres de détection de telle façon que ces deux probabilités d'erreur soient égales. Dans ce cas, que le locuteur soit ou non un imposteur, la probabilité d'erreur est la même. Celle-ci peut être résumé par un chiffre unique baptisé **taux d'erreur équiprobable** (*equal-error rate*). Toutefois, cette attitude ne constitue pas une obligation, pour certaines applications, il peut être souhaitable de chercher à minimiser un type d'erreur plutôt qu'un autre. Ainsi, dans le cas où le système de vérification de locuteur est utilisé pour contrôler l'accès à des installations ultra-confidentielles, le plus important est de réduire au maximum le taux d'acceptation erronée. Dans cet exemple, il est beaucoup plus grave d'accepter un imposteur que de rejeter un locuteur autorisé. D'une manière plus générale, il appartient à l'utilisateur d'un système de reconnaissance de *décider* quel doit être le compromis entre ces deux types d'erreurs en fonction des caractéristiques de l'application envisagée. En criminalistique ce problème se pose avec d'autant plus d'acuité compte tenu de la gravité d'éventuelles décisions erronées dans ce domaine [Federico 93], [Kunzel 94].

Il est très important de garder en mémoire le fait que les taux d'erreur définis ci-dessus n'ont aucune raison d'être identiques pour tous les locuteurs autorisés. Il est par exemple possible que le taux d'acceptation erronée soit beaucoup plus important pour un des locuteurs autorisé parce que sa voix est facile à imiter. Inversement si un des locuteurs fait preuve de mauvaise volonté, et ne s'astreint pas à parler de manière relativement uniforme, il est fort probable que le taux de rejet erroné sera beaucoup plus important pour ce locuteur particulier. De telles variations de l'efficacité de la reconnaissance selon les locuteurs ont été constatées sur plusieurs systèmes [Doddington 85], [Furui 81a], [Oglesby 94], [Soong 87] (cf. paragraphe 2.3). Ceci pose d'ailleurs une question importante dans le cadre de la criminalistique qui est celle de l'estimation de la fiabilité d'un résultat considéré individuellement [Noda 89].

Il convient donc de distinguer des taux d'erreur **individuels**, c'est à dire valables pour un locuteur particulier, et des taux d'erreur globaux obtenus par moyenne des données correspondant à tous les locuteurs. Rappelons que si les taux d'erreur individuels ne sont pas répartis de manière "suffisamment uniforme", l'utilisation de taux d'erreur globaux risque de masquer des réalités importantes (cf. [Doddington 85]).

1.2.2 Pour l'identification

Pour une application d'identification du locuteur, il s'agit de décider entre N locuteurs (avec éventuellement une alternative supplémentaire qui ne correspond à aucun des locuteurs si l'on a affaire à un groupe ouvert).

Une description très précise des performances d'une méthode d'identification consiste à fournir la **matrice des confusions** qui regroupe sous forme matricielle la probabilité de désigner le locuteur j alors que le texte est prononcé par le locuteur i (où $i = 1, \ldots, N$, $j = 1, \ldots, N$ et $j \neq i$). Cette description devient cependant très lourde lorsque le nombre N de locuteurs est important. On se contente en général de donner les chiffres plus synthétiques que sont les N **taux d'erreur** (ou taux de confusion) qui correspondent à la probabilité qu'un texte prononcé par le locuteur i ($i = 1, \ldots, N$) soit attribué à un autre locuteur. Par référence, à la terminologie utilisée dans la théorie de la décision, le taux d'erreur correspondant au locuteur i est parfois appelé erreur de $i^{ème}$ espèce. Le plus souvent, et surtout pour les systèmes destinés à fonctionner sur un groupe de taille importante, on simplifie encore en utilisant le taux d'erreur moyen, indépendant du locuteur, obtenu par moyenne des taux d'erreur individuels. Comme

dans le cas de la vérification, le taux d'erreur global peut dissimuler des disparités éventuelles existant entre les différents locuteurs.

Enfin, notons que si l'on travaille avec un groupe ouvert, il est nécessaire de prendre en compte des causes d'erreur supplémentaires analogues à celles que nous avons vu pour la vérification : acceptation erronée lorsqu'un locuteur extérieur au groupe a été confondu avec un des locuteurs du groupe, et rejet erroné lorsqu'un des locuteurs du groupe a été déclaré imposteur.

1.3 Méthodologie d'évaluation pour la reconnaissance du locuteur

1.3.1 Principe de l'évaluation

Dans le domaine de la reconnaissance du locuteur, une des principales difficultés réside dans l'évaluation de l'efficacité des techniques employées. D'une manière générale, la phase d'évaluation est souvent plus coûteuse, en termes de moyens techniques et de quantité de travail nécessaires, que la phase de mise au point. Trois types d'arguments peuvent être avancés afin de comparer les performances de différentes techniques de reconnaissance :

1. Une première possibilité consiste à mettre au point un modèle théorique du fonctionnement de la technique de reconnaissance utilisée. En pratique, de tels modèles ont une utilité assez limitée car il ne correspondent que de très loin au fonctionnement en situation réelle. Toutefois, ils peuvent permettre de dégager des grandes tendances [Doddington 85]. Le principal apport de ce type d'arguments a été de montrer que l'identification est une tâche plus difficile que la vérification [Doddington 85], [OShaughnessy 86], et ce d'autant plus que le nombre de locuteurs autorisés augmente [Doddington 85]. Un élément de comparaison important à conserver en mémoire est le fait que la plus mauvaise technique possible (la réponse au hasard !) fournit un taux d'erreur constant de 50% pour la vérification alors qu'elle correspond à un taux d'erreur croissant de $100(1 - 1/N)\%$ (où N désigne la taille du groupe) pour l'identification [OShaughnessy 86].

2. Sans aller jusqu'à une modélisation complète du fonctionnement, il est parfois possible de comparer des techniques sur la base d'arguments théoriques. Ce type de démarche a notamment été utilisé pour sélectionner les caractéristiques du signal de parole les plus appropriés pour la reconnaissance [Atal 76], [Cheung 78], [Das 71], [Sambur 75]. Malheureusement, il est souvent impossible de progresser très loin dans ce domaine sans utiliser des modèles [Atal 76], [Bricker 71] ou émettre des hypothèses [Mohn 71] qui ne correspondent qu'imparfaitement à la réalité.

3. Enfin, il est possible d'évaluer la fiabilité d'une technique par une démarche **empirique** en constituant une **base de données** d'enregistrements de parole, puis en effectuant des tests systématiques.

La seconde démarche (arguments théoriques) a surtout été exploitée durant les années 70, puis, avec l'apparition de nouvelles techniques dans les années 80, la troisième démarche (évaluation empirique) s'est très nettement imposée. Cette évolution s'explique en partie par le fait qu'il est devenu beaucoup plus difficile de mettre en évidence des arguments théoriques compte tenu de la complexité des techniques actuelles de reconnaissance du locuteur. De plus, les progrès de l'informatique ont rendu l'évaluation empirique possible même avec les ordinateurs les plus courants.

L'évaluation empirique constitue une méthode de validation très satisfaisante car elle permet d'obtenir directement une estimation de la fiabilité en situation réelle. Cette stratégie est en ce sens beaucoup plus efficace que les arguments théoriques qui ne peuvent être utilisés que pour comparer différentes méthodes entre elles. Toutefois, nous aurons l'occasion de voir au paragraphe 2.3 que le caractère empirique de l'évaluation ne va pas sans poser de sérieux problèmes, notamment en ce qui concerne l'interprétation et le domaine de validité des résultats .

1.3.2 Pratique de l'évaluation empirique

Il faut bien avoir conscience du fait que l'évaluation empirique est, en général, une démarche très lourde car l'estimation des performances n'est significative que si le nombre d'enregistrements disponibles est très important. Le dimensionnement et la composition de la base de données utilisée pour l'évaluation empirique doivent en effet vérifier un ensemble de contraintes qui sont liées, soit à des considérations statistiques, soit à la nature du signal de parole.

Contraintes statistiques Tout résultat obtenu à partir d'une série d'expériences ne représente qu'une *estimation* : il est nécessairement entaché d'une certaine incertitude. Supposons par exemple que l'on teste un système de reconnaissance de locuteurs 20 fois de suite et que l'on obtienne qu'une seule erreur. Il serait faux d'affirmer que le taux d'erreur du système est de 5%. La valeur de 5% correspond simplement à la meilleure estimation que l'on puisse faire du taux d'erreur. Pour quantifier l'incertitude de cette estimation, on utilise souvent la notion d'*intervalle de confiance*, par exemple, à 90% (défini comme l'intervalle dans lequel se situe, avec une probabilité de 90%, la valeur exacte de la quantité à estimer). En supposant que la probabilité d'erreur est la même pour tous les locuteurs et que les différents essais se font de manière indépendante, on montre [Ventsel 73, §14.5.] que l'intervalle de confiance à 90% correspondant à l'exemple précédent est 1.5% — 14%. Pour confirmer la valeur de 5% il serait donc nécessaire d'organiser un plus grand nombre d'essais.

Pour évaluer le nombre d'expérience nécessaires, on peut utiliser le résultat suivant : avec les mêmes hypothèses que précédemment, *l'estimation peut être considérée comme précise (intervalle de confiance à 90% limité environ à* $\pm 20\%$ *de la valeur estimée) dès que*

$$N > \frac{5000}{t_e}$$

où t_e désigne le taux d'erreur *exprimé en pour-cent* et N le nombre d'expériences nécessaires. Pour l'exemple précédent, afin de confirmer que le taux d'erreur est bien de 5%, il faudrait donc organiser au moins 1000 essais. Il est particulièrement important de noter que le nombre d'expériences nécessaires N est inversement proportionnel aux taux d'erreur t_e. Pour des systèmes appelés à être très fiable (taux d'erreur de quelques pour-cent), il est donc nécessaire d'organiser un très grand nombre d'essais (de l'ordre de 1000 à 5000). On peut déjà noter à ce stade que la constitution de la base de données implique l'enregistrement et le stockage d'un grand nombre d'enregistrements de parole ce qui constitue déjà en soi en travail considérable.

Contraintes liées au signal de parole Pour rendre compte de la variabilité des caractéristiques de la parole il est nécessaire d'enregistrer chaque locuteur en plusieurs occasions afin d'intégrer la variabilité intra-locuteur. Il est fortement conseillé d'enregistrer chaque locuteur lors d'au moins quatre à cinq séances distinctes séparées dans le temps du plus grand délai possible (sur une période de plusieurs mois). Ces contraintes ne doivent pas être sous-estimée car la variabilité intra-locuteur des caractéristiques mesurées influe très notablement sur les performances de la reconnaissance [Furui 81a], [Furui 81b], [Soong 87]. En particulier, si chaque

locuteur est enregistré au cours d'une seule séance, les performances de la reconnaissance se trouvent artificiellement sur-évaluées [Rosenberg 91].

Il est par ailleurs indispensable d'enregistrer un nombre suffisant de locuteur. Ce point est à rapprocher de ce qui a été dit au paragraphe 1.2 concernant les taux d'erreur individuels. Le résultat de l'évaluation empirique est d'autant plus significatif que le nombre de locuteurs est assez important et couvre un ensemble suffisamment représentatif de voix.

Enfin selon l'application envisagée, il est nécessaire d'enregistrer les locuteurs dans des conditions plus ou moins particulières (à travers le téléphone, en présence de bruit, etc.) ou en simulant artificiellement ces conditions.

Dimensionnement de la base de données Pour terminer cette partie, nous allons donner une idée du nombre minimal d'enregistrements à effectuer en vue de l'évaluation empirique.

Supposons que l'on dispose de $n = n_l \times n_e$ enregistrements, où n_l désigne le nombre de locuteurs enregistrés et n_e le nombre d'enregistrements de chaque locuteur. Il est théoriquement possible d'effectuer $n_l n_e(n_e - 1)$ tests intra-locuteur et $n_l(n_l - 1)n_e^2$ tests entre locuteurs différents (en supposant que le test de reconnaissance n'est pas symétrique), soit un total de l'ordre de n^2 tests. On arrive donc assez rapidement à l'ordre de grandeur du nombre de tests nécessaires pour obtenir des résultats statistiquement fiables (de l'ordre de quelques milliers de tests) : une base de données constituée de $n_l = 10$ locuteurs enregistrés en $n_e = 5$ occasions différentes est déjà satisfaisante du point de vue statistique. On trouve d'ailleurs de nombreux exemples pour lesquels la base de données utilisée est de cet ordre de grandeur minimal : [Cheung 78], [Furui 81b], [Gish 86], [Matsui 91], [Rose 90], [Rose 91], [Soong 88].

Toutefois, la généralisation des résultats obtenus lors de ce type d'évaluation à des applications "réelles" est très délicate. Il est clair que si on désire obtenir des résultats représentatifs pour une *grande population* (voire, comme en criminalistique pour l'ensemble des locuteurs possibles), il est indispensable de tester les performances de la reconnaissance avec un nombre beaucoup plus important de locuteurs [Bimbot 94a], [Naik 94]. C'est notamment le cas de [Furui 81a], [Reynolds 94b], [Soong 87] qui utilisent des bases de données contenant 100 à 150 locuteurs, ainsi que de [Noda 89] (étude consacrée à la criminalistique avec plus de 500 locuteurs) et [Webb 93] (évaluation d'un système d'identification en mode dépendant du texte avec près de 1000 locuteurs). [Boves 94] présente par ailleurs un exemple pratique détaillé de la constitution d'une base de données de grande taille destinée à la reconnaissance du locuteur.

Chapitre 2

Techniques de reconnaissance

2.1 Historique

On parle de technique *automatique* lorsque l'intégralité de la tâche de reconnaissance, y compris l'étape de décision, est réalisée sans aucune intervention humaine. D'un point de vue économique, ce sont surtout les techniques totalement automatiques qui présentent un intérêt potentiel. C'est donc naturellement principalement sur ce type de techniques que ce sont portés la quasi-totalité des efforts de recherche et d'évaluation.

Les premiers essais de reconnaissance automatique du locuteur datent de la seconde moitié des années 1960 [Bricker 71]. Toutefois durant toutes les années 1970, les travaux sur le sujet restent peu nombreux et émanent essentiellement de grands centres de recherche, surtout américains, liés à des constructeurs informatique tels que IBM [Das 71] et Texas Instruments [Doddington 85], [Rosenberg 76] ou bien à des compagnies de télécommunications comme AT&T [Bricker 71], [Rosenberg 75], [Rosenberg 76], [Sambur 75] et NTT [Furui 86] . Dans les années 1980, les progrès de l'informatique ainsi que l'apparition de nouvelles techniques ont stimulé la recherche dans le domaine du traitement de la parole. A l'heure actuelle, de nombreux centres de recherche à travers le monde travaillent sur les problèmes liés à la reconnaissance du locuteur (voir par exemple les actes du colloque récent consacré au sujet [ESCA 94]).

D'une manière générale, un système de reconnaissance automatique du locuteur comprend les éléments suivants [OShaughnessy 86], [Furui 94], [Rosenberg 91] :

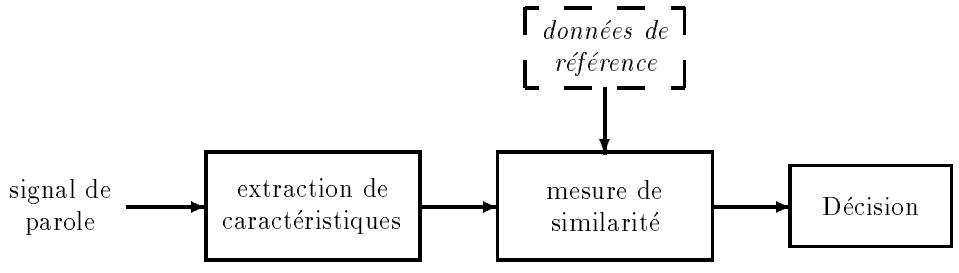

Figure 2.1: Schéma synoptique d'un système de reconnaissance automatique du locuteur

Dans ce schéma, la première étape correspond à la mesure, à partir du signal de parole, des caractéristiques utiles pour la reconnaissance. La seconde étape recouvre l'ensemble des méthodes mises en œuvre afin de comparer les différents jeux de caractéristiques mesurées. La dernière étape correspond à la décision finale de reconnaissance. Le mécanisme de décision ainsi que les données de référence qu'il est nécessaire de faire intervenir dépendent surtout du type de tâche de reconnaissance considéré (identification ou vérification) [Furui 94].

De manière un peu schématique, il est possible de diviser les travaux de recherches concernant la reconnaissance du locuteur en deux groupes relativement homogènes : dans les années 1970, c'est surtout la première étape (recherche de caractéristiques pertinentes) qui est approfondie tandis que l'étape de comparaison reste en général très simple ; dans les années 1980, il se dégage un relatif consensus sur la nature des caractéristiques du signal à mesurer et c'est la partie comparaison des caractéristiques qui devient de plus en plus complexe.

2.1.1 Extraction de caractéristiques

Au cours des années 1970, l'objectif principal des premiers travaux consacrés à la reconnaissance du locuteur consiste à déterminer quelles sont les caractéristiques du signal de parole qui permettent de discriminer différents locuteurs. En général, cette sélection de caractéristiques pertinentes s'effectue en ayant recours à des arguments de type statistique [Bricker 71], [Cheung 78], [Markel 77], [Mohn 71], [Rosenberg 76], [Sambur 75].

Pour comparer les différentes caractéristiques entre elles, la mesure d'efficacité le plus souvent utilisée est le *F ratio* (technique d'analyse de variance) [Bricker 71], [Goldstein 75], [Mohn 71], [Wolf 72]. Cette mesure permet de sélectionner les caractéristiques qui présentent à la fois une faible variabilité intra-locuteur et une forte variabilité inter-locuteur [Calliope 89, Chap. XIX], [OShaughnessy 86]. Une des limitations de cette mesure est qu'elle devient peu significative dès lors que les caractéristiques sont corrélés entre elles [Atal 76], [Mohn 71]. Dans de tels cas, il est nécessaire de considérer plusieurs caractéristiques simultanément. Une généralisation du *F ratio* applicable à des données multidimensionnelles est la mesure dite de *divergence* [Atal 76], [Cheung 78], [OShaughnessy 86].

En général, la démarche adoptée consiste à partir d'un grand nombre de caractéristiques possibles et à ne conserver que les plus efficaces. Pour ce faire, il est possible d'avoir recours à l'analyse discriminante [Bricker 71], [Mohn 71] qui permet d'extraire un sous ensemble de caractéristiques possédant un *F ratio* maximal. Ce type de démarche pose toutefois des problèmes car elle suppose des hypothèses qui ne sont pas forcément vérifiées par les caractéristiques considérées [Mohn 71]. Le plus souvent, la sélection des caractéristiques se fait par une procédure, en général itérative, faisant directement référence aux performances de reconnaissance. La procédure la plus utilisée est le *knock out* (élimination) qui consiste, à chaque étape, à éliminer la caractéristique qui contribue le moins aux performances de la reconnaissance [Sambur 75], [OShaughnessy 86], [Rosenberg 75]. Cette procédure très simple ne garantit pas l'optimalité du choix des caractéristiques (un exemple de procédure de sélection plus complexe est présentée dans [Cheung 78]).

Les caractéristiques du signal considérées sont le plus souvent issues soit d'études de type psychoacoustique visant à déterminer les facteurs intervenant dans la reconnaissance par des auditeurs humains (comme par exemple [LaRiviere 75]), soit de considérations concernant les mécanismes physiques de production de la parole [Atal 76], [Goldstein 75], [Sambur 75], [Wolf 72]. Les caractéristiques étudiées sont fréquemment déterminées par des procédures semi-automatiques basées sur une segmentation du signal de parole par un expert humain [Goldstein 75], [Sambur 75], [Wolf 72]. D'autre part, dans la quasi totalité des articles cités, l'exploitation qui est faite des

caractéristiques mesurées reste très simple :

- En mode dépendant du texte, les caractéristiques sélectionnées sont mesurées à différents instants caractéristiques. On obtient donc un profil (en anglais *contour*) qui décrit l'évolution temporelle de chaque caractéristique considérée. Par la suite, les profils mesurés pour les différents locuteurs sont comparés en calculant une distance moyenne [Rosenberg 75]. Cette méthode nécessite de recourir à une procédure d'alignement temporel qui vise à corriger les décalages temporels pouvant exister entre deux répétitions d'un même texte [Das 71], [Rosenberg 76].
- En mode indépendant du texte, seule la valeur moyenne des caractéristiques sélectionnées, calculée sur l'ensemble du texte prononcé par chaque locuteur, est utilisée pour la reconnaissance [Atal 76], [Cheung 78], [Markel 77], [Markel 79], [Shridhar 82]. D'autre part, la conviction la plus répandue est que seules doivent être prises en compte les parties voisées du signal de parole [Cheung 78], [Markel 77], [Shridhar 82].

2.1.2 Exploitation globale des données

Dès le début des années 1980, les travaux de recherche consacrés à la reconnaissance du locuteur connaissent un changement radical d'orientation caractérisé par deux éléments :

1. Il se dégage un consensus sur le type de caractéristiques du signal de parole à mesurer dans le cadre de la reconnaissance du locuteur : actuellement, la quasi totalité des systèmes travaille à partir de paramètres dérivés du spectre à court-terme du signal de parole. Un point important est que ce type de paramètres est mesuré "continûment" tout au long du signal. Les méthodes basées sur la localisation explicite d'événements acoustiques dans le signal de parole ont quasiment disparu [Doddington 85], [Rosenberg 91] (à de rares exceptions près tel que [Fatokakis 93]).
2. Le point central de la recherche se déplace vers l'étude de méthodes efficaces permettant d'exploiter l'ensemble des caractéristiques mesurées. Plutôt que de chercher à mettre explicitement en évidence des éléments spécifiques au locuteur, la démarche suivie consiste à élaborer des méthodes qui prennent en compte l'ensemble des paramètres mesurés afin de mesurer la similarité existant entre deux jeux de mesures. Ceci se traduit par le fait que les progrès obtenus dans le domaine viennent essentiellement de l'introduction de méthodes avancées d'analyse de données multidimensionnelles.

Cette évolution de la recherche suit ce qui s'est passée dans le domaine de la reconnaissance de la parole où l'approche dite acoustique-phonétique, basée sur l'extraction et la reconnaissance explicite des événements acoustiques correspondant aux éléments phonétiques, s'est avérée moins efficace, dans le cadre d'applications pratiques, que les méthodes de type reconnaissance de forme [Rabiner 93]. En ce qui concerne la reconnaissance du locuteur, l'avis de S. Furui, éminent spécialiste du domaine, résume bien cette évolution [Furui 94][1] : "les progrès récents obtenus dans le domaine de la reconnaissance du locuteur sont principalement dus à l'amélioration des techniques utilisées pour modéliser et décrire les caractéristiques mesurées pour chaque locuteur. Ces progrès n'ont pas forcément permis d'accroître ou d'améliorer nos connaissances en ce qui concerne les particularités propres à chaque locuteur, et la manière de les extraire du signal de parole" (traduit de l'anglais).

[1] Curieusement, la conclusion de S. Furui reprend, quasiment dans les mêmes termes, la conclusion A. Rosenberg et F. Soong (autres spécialistes du domaine) dans [Rosenberg 91].

2.1.3 Repères bibliographiques

Compte tenu du nombre relativement important de références bibliographiques, il nous a semblé utile de fournir quelques repères permettant de distinguer les principaux "groupes" de publications. Cette présentation (non exhaustive !) sous forme de court résumé, a été limitée aux équipes de recherche pour lesquelles nous disposions d'un nombre de suffisant de publications :

AT&T Bell Labs

Auteurs : A. Rosenberg, F. Soong, ...

Domaine : Reconnaissance du locuteur à partir d'enregistrements téléphoniques en mode indépendant du texte avec "vocabulaire contraint" (limité à des codes formées de chiffres, voire à des chiffres isolés).

Techniques : Caractérisation fondée sur la quantification vectorielle [Soong 85], [Soong 87], [Rosenberg 86], éventuellement avec des paramètres variationnels [Soong 88]. Modélisation par un modèle de Markov caché [Rosenberg 90], [Webb 93]. Modélisation sous la forme de mélange de densités gaussiennes multidimensionnelles [Tseng 92].

NTT Human Interface Laboratories

Auteurs : S. Furui, T. Matsui

Domaine : Reconnaissance du locuteur de manière générale. Techniques robustes vis à vis de la variabilité (due au canal de transmission ou au locuteur).

Techniques : Reconnaissance en mode dépendant du texte avec alignement temporel par programmation dynamique [Furui 81a], [Furui 81b]. Caractérisation par quantification vectorielle avec étiquetage automatique préliminaire [Matsui 91]. Comparaison entre les techniques basées sur la quantification vectorielle et celles utilisant un modèle de Markov caché [Matsui 92], [Matsui 94a]. Vérification simultanée du locuteur et du texte (pour les applications de type *texte dicté*) en utilisant des modèles de Markov cachés [Matsui 93], [Furui 94]. Méthodes de *normalisation de la mesure de similarité* dans les applications de vérification du locuteur [Matsui 93], [Matsui 94b], [Furui 94].

MIT Lincoln Lab.

Auteurs : D. Reynolds, C. Rose, ...

Domaine : Reconnaissance du locuteur en mode indépendant du texte (parole naturelle, enregistrée par téléphone, éventuellement en milieu bruité).

Techniques : Modélisation sous la forme de mélange de densités gaussiennes multidimensionnelles [Rose 90], [Rose 91], [Reynolds 92], [Reynolds 94b]. Prise en compte du bruit de fond dans le modèle [Rose 91], [Rose 94].

BBN

Auteurs : H. Gish, ...

Domaine : Reconnaissance du locuteur en mode indépendant du texte : techniques robustes pour les enregistrements téléphoniques. Séparation de locuteurs en train de dialoguer.

Techniques : Modélisation sous la forme d'une densité gaussienne multidimensionnelle [Krasner 84], [Gish 85], [Gish 86], [Gish 90], [Gish 91], [Yu 93]. Modélisation statistique de l'effet du canal de transmission [Gish 85], [Gish 86]. Modification du classificateur gaussien pour accroître sa robustesse vis à vis des variations du canal de transmission [Gish 86], [Gish 90]. Application du modèle gaussien à la séparation de locuteurs [Gish 91], [Yu 93].

2.2 Description des techniques de reconnaissance

2.2.1 Caractéristiques utilisées

Actuellement, la grande majorité des systèmes de reconnaissance du locuteur travaillent à partir de représentations paramétriques du spectre de puissance à court-terme du signal de parole [Rosenberg 91]. Le point de départ est donc une vision synthétique (décrite par un faible nombre de paramètres) du contenu spectral de portions successives du signal délimitées par une fenêtre temporelle glissante.

En étant plus précis, on peut dire que les paramètres utilisés sont presque toujours de type "cepstraux" (pour plus d'information concernant les propriétés du cepstre, se reporter à [Rabiner 78, §7]). Les figures 2.2, 2.3 et 2.4 présentent les trois possibilités les plus classiques de calcul des paramètres cepstraux à partir du signal de parole.

2.2.1.a Méthodes de calcul des paramètres cepstraux

L'approche de la figure 2.2, fondée sur la transformée de Fourier à court-terme nécessite deux calculs de transformée de Fourier discrète, directe (TFD) et inverse (TFDI). C'est l'approche de calcul la plus classique du cepstre [Rabiner 78]. Compte tenu du fait que le spectre à court-terme de puissance (en échelle logarithmique) est une grandeur réelle, la transformée de Fourier inverse peut être remplacée par une transformée en cosinus discrète (*Discrete Cosine Transform* ou DCT) [Rabiner 93, Eq. (4.90)], [Soong 88].

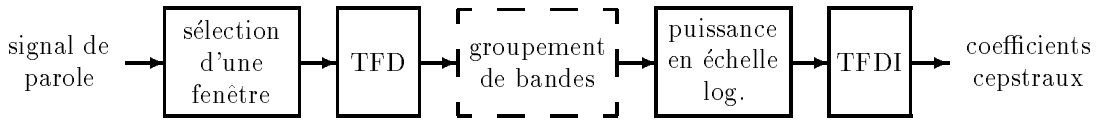

Figure 2.2: Calcul des paramètres cepstraux : approche par transformée de Fourier à court-terme

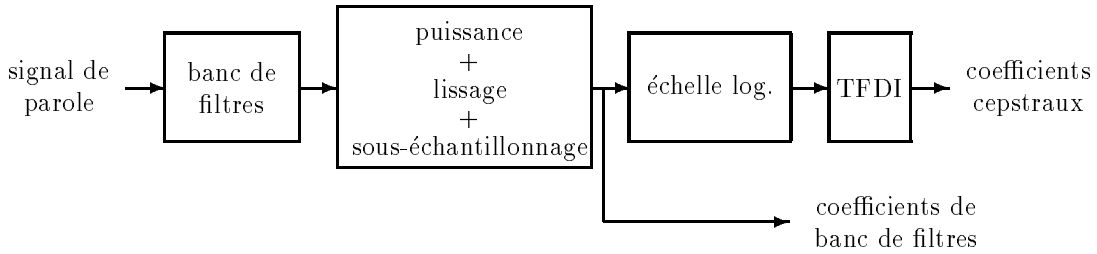

Figure 2.3: Calcul des paramètres cepstraux : approche par banc de filtres

Sur la figure 2.3, le principe est quasiment le même à la différence près que l'étape de transformée de Fourier à court-terme directe (partie gauche de la figure) a été remplacée par un banc de filtres avec calcul de puissance dans chaque voie. De fait, ces deux approches sont extrêmement proches car la transformée de Fourier à court-terme (figure 2.2) peut être formulée sous la forme d'un banc de filtres uniforme (avec décimation des signaux de sous-bande) [Rabiner 93]. La possibilité de grouper les bandes de la transformée de Fourier à court-terme (en pointillés

sur la figure 2.2) permet de simuler un banc de filtres à largeur de bande non-uniforme ce qui rend les deux approches quasiment semblables. Le terme de groupement de bande correspond à un cumul du périodogramme du signal à court-terme (module au carré de la TFD du signal à court-terme) sur plusieurs points fréquentiels, éventuellement en utilisant une fonction de fenêtrage (pour plus de détail, voir le chapitre 2 de [Reynolds 92] ou [Davis 80]). Il faut noter que le groupement de bandes de la transformée de Fourier à court-terme ne correspond pas à un véritable banc de filtres, il permet simplement d'obtenir une estimation de la puissance dans les voies d'un banc de filtres équivalent. S'il est possible de contrôler la largeur de bande de chaque voie ainsi que l'allure approximative des filtres du banc de filtre équivalent, il est par contre impossible de spécifier la forme exacte des réponses temporelles ou fréquentielles des filtres de sous-bande comme dans l'approche de la figure la figure 2.3 [Rabiner 93].

Sur la figure 2.3, on a représenté le fait que dans certains systèmes, les paramètres utilisés correspondent directement aux valeurs de la puissance du signal présent dans chaque voie du banc de filtre.

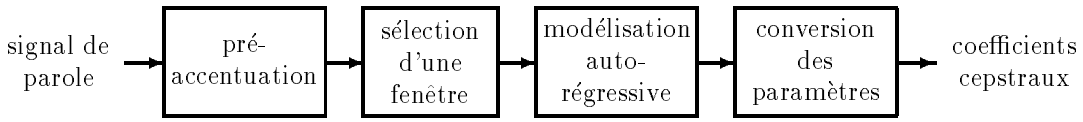

Figure 2.4: Calcul des paramètres cepstraux : approche par modélisation autorégressive

L'approche de la figure 2.4, qui est la plus classique, utilise un modèle autorégressif (AR) du signal qui fournit une première description paramétrique du contenu spectral du signal. Une formule de conversion permet ensuite de passer directement des paramètres du modèle AR au coefficients cepstraux [Rabiner 78, §8.8], [Rabiner 93, §3.3.7]. Dans cette approche, la préaccentuation est nécessaire pour garantir une adéquation satisfaisante du modèle AR au signal [Rabiner 93, §3.3.7]. Cette préaccentuation n'a pas d'influence sur la suite du traitement, particulièrement pour les systèmes utilisant la *normalisation cepstrale* (cf. paragraphe 2.2.4).

En pratique, on ne conserve que les premiers coefficients cepstraux ce qui a pour effet de lisser la représentation spectrale équivalente [Rabiner 78, §7]. Compte tenu de cette modification, les approches des figures 2.2 et 2.4 donnent en fait des résultats extrêmement proches, la structure de la figure 2.2 étant plus coûteuse en termes de charge de calcul [Furui 81a].

Il est très important de limiter les caractéristiques mesurées à la bande passante des enregistrements traités (voir par exemple les difficultés rencontrées dans [Gish 90]). Dans le cas d'enregistrements téléphoniques, il est indispensable de se restreindre à la bande 300 — 3000 Hz (soit par filtrage initial, soit par calcul des paramètres cesptraux uniquement dans cette bande de fréquence). En général, on considère que compte tenu du type de mesures de similarité utilisé, il est bénéfique de conserver les paramètres correspondant à toutes les fenêtres où le signal est présent, que celui-ci soit voisé ou non [Soong 87]. Par contre, il peut être utile d'exclure les fenêtres où le signal de parole est absent (pauses, hésitations, etc.). Ce point ne doit pas être négligé car le coefficient cepstral correspondant au niveau absolu du signal n'est en général pas utilisé, ce qui signifie que des paramètres correspondant au seul bruit de fond présent sur l'enregistrement sont susceptibles d'être traités exactement comme des paramètres mesurés alors que le locuteur parle [Soong 87]. Cette élimination des fenêtres de "silence" est surtout nécessaire pour les applications où le canal de transmission est de mauvaise qualité (bruit de fond important), et éventuellement, susceptible de varier [Gish 90], [Naik 89], [Reynolds 94b].

Un autre point important est le fait que tous les coefficients cepstraux ne présentent pas des variations intra-locuteur de même ampleur : la variance des coefficients cepstraux décroît avec leur ordre [Soong 88]. Pour tenir compte de cette propriété, on utilise généralement la distance dite de Mahalanobis, qui se réduit ici à une distance euclidienne pondérée du fait de la décorrélation des coefficients cepstraux. Cette distance est beaucoup plus efficace dans le cadre de la reconnaissance du locuteur car elle permet de réduire l'influence des coefficients qui présentent une forte variabilité intra-locuteur [Soong 88]. Compte tenu des propriétés du signal de parole, les poids à appliquer aux coefficients cepstraux pour le calcul de la distance pondérée peuvent être approximés par une progression linéaire en fonction de l'ordre des coefficients cepstraux (on obtient alors la distance dite *root powers sum*) [Juang 87], [Soong 88], [Xu 89b].

2.2.1.b Paramètres dérivés

Une modification, introduite récemment, consiste à calculer les paramètres cepstraux issus d'un banc de filtres dit "en bandes critiques" qui s'inspire de résultats concernant le fonctionnement de l'audition humaine [Rabiner 93, §4.5.6], [Botte 88]. Les paramètres obtenus sont appelés coefficients cepstraux en échelle fréquentielle Mel (*Mel-Frequency Ceptsrum Coefficients* ou MFCC) [Davis 80]. Cette modification peut être mise en œuvre très simplement dans le cadre de l'implémentation de la figure 2.3 [Rabiner 93, §3.2] (et de manière approximative dans celle de la figure 2.2). Il est également possible de simuler cette modification à partir des paramètres cepstraux par l'approche de la figure 2.4 au moment du calcul de distance [Rabiner 93, §4.5.6]. Ces paramètres MFCC semblent être légèrement plus efficaces que les paramètres cepstraux classiques [Reynolds 94b], [Reynolds 94a]. En ce qui concerne la représentation spectrale équivalente la différence entre les deux jeux de paramètres est essentiellement que les MFCC fournissent une vision plus précise en base fréquence (en dessous de 1000 Hz) et moins détaillée en haute fréquence.

Une autre modification plus conséquente consiste à utiliser des paramètres variationnels (dits Δ-cepstraux) qui caractérisent les variations des paramètres cepstraux dans les fenêtres proches de la fenêtre à court-terme courante [Furui 81a], [Soong 88]. Ces paramètres variationnels sont nettement moins efficaces que les paramètres cepstraux dans le cadre de la reconnaissance du locuteur lorsqu'ils sont utilisés seuls. Par contre, utilisés avec les paramètres cepstraux "instantanés" il conduisent à une amélioration substantielle des performances [Furui 81a], [Tseng 92], [Soong 88]. Les paramètres variationnels présentent en outre l'intérêt d'être insensibles à des variations linéaires du canal de transmission entre deux enregistrements [Soong 88]. Toutefois, les paramètres variationnels ne sont réellement significatifs que dans les cas où il est possible de les comparer dans un même contexte, c'est à dire dans les applications où le texte prononcé par chaque locuteur est fixé [Soong 88]. En mode indépendant du texte l'utilisation de paramètres, Δ-cepstraux, en plus des paramètres cepstraux, ne semble pas améliorer sensiblement les performances de reconnaissance (et ce même dans les cas où le vocabulaire est fortement contraint comme dans [Tseng 92]).

2.2.1.c Avantages des paramètres cepstraux

Le succès des paramètres cepstraux (au sens large, en y incluant paramètres mentionnés ci-dessus) s'explique par le fait qu'ils cumulent un grand nombre d'avantages :

- Le cepstre est une représentation de l'information spectrale pour laquelle une opération de filtrage linéaire se traduit par une modification additive (c'est le principe du traitement de signal homomorphique) [Rabiner 78, §7]. Par conséquent, il est théoriquement

simple de compenser dans le domaine cepstral l'effet d'un filtrage linéaire (même si le filtre est inconnu). C'est le principe de la *normalisation cepstrale* [Furui 81a] qui est, à ce jour, le moyen le plus efficace pour contrer les variations du canal de transmission (cf. paragraphe 2.2.4). Cette propriété est donc en particulier indispensable pour les enregistrements téléphoniques [Furui 94], [Gish 86] et [Gish 90], [Reynolds 94b], [Rose 90].

- Le cepstre fournit naturellement un moyen très simple de réduire de manière pertinente la quantité d'information. On sait en effet que le lissage de la représentation spectrale obtenu en ne conservant que les premiers coefficients cepstraux conserve l'enveloppe spectrale caractéristique du signal de parole [Rabiner 78, §7], [Soong 88].

- La famille la plus simple de distances (distance euclidienne, éventuellement pondérée) correspond, lorsqu'elle est appliquée aux paramètres cepstraux, à une mesure de distance sur les spectres à court-terme dont on sait qu'elle est pertinente dans le cas de signaux de parole [Rabiner 93, §4.5]. Cette propriété vient à bout d'un défaut commun à plusieurs systèmes plus anciens pour lesquels les mesures de similarité calculées étaient difficilement interprétables (c'est en particulier le cas pour les système qui utilisaient des distances euclidiennes appliquées directement aux coefficients du modèle AR comme [Rosenberg 75] ou [Shridhar 82]).

- Les coefficients cepstraux sont statistiquement très faiblement corrélés [Soong 88]. Cette propriété rend d'une part inutile les procédures d'orthogonalisation souvent employées avec d'autres jeux de paramètres, d'autre part, elle simplifie la mesure de la distance entre deux jeux de paramètres cepstraux [Soong 88]. Par ailleurs, cette propriété simplifie notablement la structure des modèles utilisant des mélanges de densités gaussiennes [Matsui 92], [Reynolds 94b]. Notons que la propriété de non-corrélation est une conséquence de la transformée de Fourier inverse (dernière opération à droite sur les figure 2.2 et 2.3) [Soong 88]. Par conséquent, les coefficients de banc de filtres (même si la puissance dans chaque voie est exprimée sur une échelle logarithmique) ne possèdent pas cette propriété.

- Enfin, de manière plus anecdotique, les paramètres cepstraux fournissent un moyen très simple de s'affranchir de la différence de niveau qui peut exister entre deux spectres identiques : il suffit de ne pas prendre en compte le premier coefficient cepstral pour travailler sur des paramètres normalisés en niveau (par rapport au niveau moyen sur une échelle logarithmique) [Rabiner 93, §4.5.2].

A de rares exceptions près, les caractéristiques utilisées dans les systèmes actuels de reconnaissance sont donc soit les paramètres cepstraux standards (en général calculés par l'approche de la figure 2.4) soit des paramètres de type MFCC, éventuellement complétés par leurs paramètres variationnels respectifs (Δ-cepstre ou Δ-MFCC) pour les applications en mode dépendant du texte. On trouvera des résultats (ainsi que des références) concernant plusieurs autres types de caractéristiques proposés récemment notamment dans [Assaleh 94], [Kao 93], [Naik 94], [Openshaw 93], [Reynolds 94a], [Xu 89a] et [Xu 89b]. [Quatieri 94] présente une étude originale (et très récente) visant à intégrer, dans la reconnaissance du locuteur, des caractéristiques du signal de parole qui ne soient pas uniquement "spectrales".

2.2.2 Mesures de similarité

Conformément à ce qui a été dit au paragraphe 2.1.2, dans la plupart des systèmes actuels de reconnaissance du locuteur, ce sont les méthodes mises en œuvre pour évaluer la similarité entre

plusieurs jeux de paramètres mesurés qui constituent la partie centrale du traitement. Dans ce paragraphe, sont présentées les principales mesures de similarité utilisées. L'ordre de présentation reflète à peu près la complexité croissante du modèle sous jacent, ce qui correspond aussi environ à l'ordre chronologique dans lequel ces techniques ont été étudiées.

Avant de débuter cette présentation, il est nécessaire de préciser quelques points de terminologie qui permettent de mieux comprendre le principe des méthodes utilisées. Nous avons vu au paragraphe précédent que les paramètres mesurés sont des coefficients cepstraux qui reflètent les variations du spectre à court-terme du signal de parole. En général, on conserve une dizaine de coefficients cepstraux, mesurés une fois toutes les 8 à 15 ms (disons 10 coefficients toutes les 10 ms, pour simplifier). Une vision commode consiste à considérer que toutes les 10 ms on mesure en fait une *quantité vectorielle* appartenant à un espace vectoriel (de dimension 10 dans notre exemple). Chaque coefficient cepstral constitue donc une coordonnée du vecteur mesurée. Cette interprétation est ici particulièrement justifiée du fait que les coefficients cepstraux sont des quantités homogènes.

Si on suppose que l'on analyse un signal de parole d'une durée de 20 s, on obtient donc 2000 mesures successives de cette grandeur vectorielle. Le principe retenu consiste à caractériser, à partir de ces mesures, la répartition (ou, de manière plus précise la *distribution*) de cette quantité vectorielle. Les techniques utilisées peuvent donc être définies comme étant des méthodes d'analyse statistique (à partir d'observation) de données multidimensionnelles. Formulé de cette façon, la mesure de similarité se ramène à un problème classique de *reconnaissance de forme* [Duda 73], [Rabiner 93].

On distingue deux grand types de techniques reconnaissance de forme : *méthodes paramétriques* lorsqu'on suppose a priori connue la forme de la distribution de la quantité vectorielle et qu'il s'agit, à partir des mesures, de déterminer des paramètres inconnus de cette distribution ; *méthodes non-paramétriques* lorsqu'on ne dispose pas de modèle permettant de décrire la forme de la distribution [Duda 73]. Enfin, dans le cadre de la reconnaissance du locuteur, il est très important de distinguer les *méthodes séquentielles* où l'ordre dans lequel sont mesurés les vecteurs d'observations est pris en compte et les *méthodes globales* où cet ordre n'est pas considéré comme significatif.

2.2.2.a Discrimination par la valeur moyenne

Technique non-paramétrique globale.
Références : [Doddington 85], [Furui 86] ainsi que [Cheung 78], [Markel 77], [Markel 79], [Shridhar 82].

Cette approche, étudiée surtout à la fin des années 1970 et au début des années 1980, est la plus simple envisageable : il s'agit de caractériser la distribution des caractéristiques vectorielles mesurées uniquement par leur valeur moyenne [Cheung 78], [Markel 77], [Markel 79], [Shridhar 82]. La similarité entre deux jeux de mesures se réduit donc à un calcul de distance entre leurs valeurs moyennes. Compte tenu du fait que les paramètres utilisés sont liés au spectre à court-terme, ce type de démarche est souvent désignée par le terme de *Long Term (Averaged) Spectrum* (spectre moyen à long-terme), et parfois par les abréviations LAS, LTAS ou LTS [Doddington 85], [Doherty 78], [Doherty 76], [Furui 86], [Kunzel 94].

Malheureusement, le spectre moyen à long-terme est très sensible aux variations du canal de transmission, ce qui le rend inutilisable dès que les locuteurs ne sont pas enregistrés dans un environnement complètement contrôlé, et en particulier, lorsqu'il s'agit d'enregistrements téléphoniques [Doddington 85], [Furui 81a] (voir aussi plus loin le paragraphe concernant le classificateur gaussien). De plus, le spectre moyen à long terme ne semble pas constituer une carac-

téristique stable du locuteur. D'une part, il varie notablement selon le style de parole adoptée par le locuteur (parole chuchotée, voix forte, etc.) [Doddington 85]. D'autre part, si le spectre moyen à long-terme est relativement stable sur de courtes périodes (moins d'une dizaine de jours), il présente une variabilité très importante pour des enregistrements d'un même locuteur réalisés à plusieurs mois d'intervalle [Furui 86].

En pratique, dans les cas où le texte prononcé est suffisamment long, qui sont les seuls où l'approche du spectre moyen à long-terme est applicable [Markel 79], l'ensemble des chercheurs s'entendent pour affirmer que ce sont les variations autour du spectre moyen qui sont les plus significatives : le spectre moyen à long-terme étant trop sensible à la variabilité inter-session des caractéristiques du locuteur et aux variations du canal de transmission, il ne doit pas être pris en compte mais plutôt utilisé comme un élément de normalisation [Furui 94], [Rosenberg 91] (voir le paragraphe 2.2.4 concernant la normalisation cepstrale).

2.2.2.b Calcul de distance avec alignement temporel

Technique non-paramétrique séquentielle, applicable en mode dépendant du texte.
Références : [Furui 81a], [Furui 81b], [OShaughnessy 86] ainsi que [Naik 89].

En mode dépendant du texte, une idée assez naturelle consiste à considérer les paramètres dans l'ordre où ils ont été mesurés et à calculer la distance, fenêtre par fenêtre, entre les paramètres correspondant aux différents locuteurs à comparer. La difficulté posée par cette approche est qu'il est nécessaire de compenser les différences de vitesse d'élocution pouvant exister entre plusieurs répétitions d'un même texte. Cet *alignement temporel* est réalisé par une technique de programmation dynamique (algorithme dit de *Dynamic Time Warping* ou DTW) [Furui 94], [OShaughnessy 86].

Dans le domaine de la reconnaissance du locuteur comme dans celui de la reconnaissance de la parole, ce type de technique a été peu à peu abandonné au profit des modèles séquentiels statistiques (comme les modèles de Markov cachés) qui sont moins "rigides" et donc plus robustes vis à vis de la variabilité inhérente au signal de parole [Furui 94], [Naik 89], [Rabiner 93].

2.2.2.c Classificateur gaussien

Technique paramétrique globale.
Références : [Krasner 84], [Gish 85], [Gish 86], [Gish 90], [Gish 94] ainsi que [Bimbot 93], [Bimbot 94b].

Pour décrire un ensemble de mesures plus précisément que par une simple valeur moyenne, l'approche la plus courante consiste à supposer que les mesures suivent une répartition gaussienne. Il s'agit ici d'une distribution gaussienne multidimensionnelle car nous avons affaire à des mesures d'une quantité vectorielle. Les paramètres à estimer à partir des données sont donc, d'une part, le vecteur moyen (identique à celui qui est utilisé dans les approches dites à long-terme), et d'autre part, la matrice de covariance des données [Duda 73, §2.7].

L'intérêt de cette modélisation gaussienne est qu'elle permet de calculer facilement une famille de tests optimaux visant à déterminer si un nouvel ensemble de mesures est compatibles avec le modèle. Cette famille de test est fondée sur *le rapport de vraisemblance*, la vraisemblance des observations se calculant facilement à partir de la forme analytique de la distribution gaussienne lorsqu'on admet l'hypothèse que les mesures sont statistiquement indépendantes [Duda 73, §2], [Gish 85], [Gish 86].

Un des problèmes posés par cette approche est que le calcul de vraisemblance fait intervenir à la fois la moyenne et la matrice de covariance des mesures [Gish 85]. En effet, dans le cadre d'une application où les enregistrements sont réalisés par le téléphone, le terme lié à la moyenne des mesures est peu significatif (voir plus haut le paragraphe concernant la caractérisation par la moyenne). Devant les difficultés posées par une modélisation de l'effet du canal de transmission en l'absence d'information concernant celui-ci, les auteurs de [Krasner 84], [Gish 85], [Gish 86] et [Gish 90] en sont venu à préconiser une modification du classificateur gaussien qui consiste à ne considérer que la partie de la vraisemblance qui dépend de la matrice de covariance des données. Cette modification du classificateur gaussien conduit à une approche où l'on cherche à mesurer la *similarité existant entre les matrices de covariance*. Diverses variantes de cette approche sont présentées dans [Gish 90], [Gish 94] et [Bimbot 93], [Bimbot 94b]. Ces différentes techniques font, au moins implicitement, référence au modèle gaussien multidimensionnel dans ce sens qu'elles supposent que la matrice de covariance permet de rendre efficacement compte de la répartition des données.

En ce qui concerne les performances du classificateur gaussien, il faut tout d'abord souligner que l'hypothèse gaussienne n'est pas vérifiée en pratique par les paramètres cepstraux issus du signal de parole (cf. histogrammes des coefficients cepstraux présentés dans [Rose 90] et [Rabiner 89, VI.E]). D'autre part, le classificateur gaussien est un cas particulier d'un autre modèle que nous verrons plus loin : le mélange de densités gaussiennes. Les résultats présentés dans [Rose 90] montrent que le classificateur gaussien fournit de moins bon résultats que le modèle de mélange de densités gaussiennes[2]. Ce résultat indique simplement que le modèle de mélange de densités gaussiennes est plus susceptible de représenter la distribution réelle des paramètres cepstraux.

Toutefois, les techniques inspirées du modèle gaussien présentent certains attraits : elles sont peu coûteuses en termes de temps de calcul et présentent des performances déjà satisfaisantes pour certaines applications [Bimbot 93], [Bimbot 94b]. De plus, elles nécessitent l'estimation d'un faible nombre de paramètres, et peuvent donc être mise en œuvre même lorsque les enregistrements disponibles pour l'apprentissage sont très courts (moins de 10 secondes de parole) [Bimbot 94b]. Il serait intéressant de comparer la robustesse de ce type de modèle, dans des conditions d'apprentissage très court, avec celle de modèles dont la structure est plus complexe.

2.2.2.d Représentation par quantification vectorielle

Technique non-paramétrique globale.
Références : [Soong 85], [Rosenberg 86], [Soong 87], [Soong 88], [Rosenberg 91], [Matsui 91], [Matsui 92], [Furui 94] ainsi que [Makhoul 85] et [Rabiner 93] pour ce qui concerne plus spécifiquement la quantification vectorielle.

La quantification vectorielle (*Vector Quantization* ou VQ) est une méthode non-paramétrique qui permet de décrire un ensemble de données par un faible nombre de vecteurs formant un *dictionnaire* (*codebook*) associé aux données. Le dictionnaire est en général calculé de telle façon que la distance, ou *distorsion*, moyenne entre un vecteur issu des données et son plus proche voisin dans le dictionnaire soit la plus faible possible [Makhoul 85]. La quantification vectorielle est une technique de groupage (*clustering*) qui est d'autant plus adaptée que les données présentent naturellement des "points d'accumulation" autour desquels la densité de vecteurs issus des données est importante [Duda 73, §6], [Makhoul 85]. Compte tenu de la nature du signal de parole, le

[2]D'une manière générale, tous les auteurs travaillant avec ce type de modèle s'accordent pour trouver que l'efficacité de la reconnaissance augmente avec le nombre de densités gaussiennes constitutives du mélange, du moment que les données disponibles pour l'apprentissage sont suffisantes. [Matsui 92], [Rose 90], [Tseng 92]

choix d'un tel modèle semble assez judicieux. En général, la quantification vectorielle est réalisée par une méthode dite de *binary (spliting) K-means* (optimisations successives de dictionnaires de taille croissante) qui permet de contourner le délicat problème de l'initialisation de l'algorithme de recherche itérative des vecteurs du dictionnaire [Makhoul 85, §V].

Pour la reconnaissance du locuteur, la mesure de similarité entre deux jeux de mesures consiste à évaluer la distorsion moyenne d'un des deux ensembles de mesures en utilisant le dictionnaire optimisé pour l'autre jeu de mesures par quantification vectorielle.

La caractérisation de la distribution des données obtenue par la quantification vectorielle est en fait voisine de celle fournie par un modèle de mélange de densités gaussiennes (cf. paragraphe suivant). Les performances des deux types de systèmes sont donc assez proches [Matsui 92]. Lorsque les données disponibles pour l'apprentissage sont suffisantes, il semble que le modèle de mélange de densités gaussiennes soit plus robuste (voir l'illustration graphique de la différence les deux modèles présentée dans [Matsui 92]). A l'opposé lorsque les enregistrements utilisés pour l'apprentissage sont de courte durée (moins de 20 secondes), il semble que la quantification vectorielle fournisse une description plus fiable que le modèle de mélange gaussien qui nécessite l'estimation d'un grand nombre de paramètres [Matsui 92].

Il est possible de donner une estimation du volume de données nécessaire pour l'apprentissage en notant que les performances de la reconnaissance ne s'améliorent que très peu lorsque le nombre de vecteurs du dictionnaire dépasse 64 à 128 [Matsui 92], [Soong 87]. En général, on considère que l'algorithme de quantification vectorielle ne fonctionne de manière satisfaisante que si on dispose d'au moins 20 à 50 fois plus de vecteurs de données que de vecteurs du dictionnaire [Makhoul 85, §V.E], [Rosenberg 86]. En supposant que les paramètres cepstraux sont mesurés toutes les 10 ms, on obtient une durée totale minimale des enregistrements utilisés pour l'apprentissage de l'ordre de 20 secondes (64 vecteurs du dictionnaire $\times$ 30 $\times$ 10 millisecondes). De plus, il faut se souvenir que cette durée de 20 secondes n'inclut pas les fenêtres de "silence" qui existent dans tout enregistrement de parole (cf. paragraphe 2.2.1). En pratique, on commence à observer une diminution très nette des performances lorsque la durée des enregistrements d'apprentissage devient inférieure à une dizaine de secondes [Matsui 91].

2.2.2.e Modélisation par mélange de densités gaussiennes

> Technique paramétrique globale.
> Références : [Reynolds 92], [Rose 90], [Rose 91], [Rose 94], [Reynolds 94b], [Tseng 92], [Matsui 92] ainsi que [Duda 73, §6] et [Dempster 77], [Redner 84] pour les aspects concernant le modèle de mélange de densités gaussiennes et l'algorithme EM.

Le modèle de mélange de densités gaussiennes (*gaussian mixture*)[3] consiste à supposer que la distribution des données peut être décrite comme une somme pondérée de densités gaussiennes (ici multidimensionnelles) [Duda 73, §6], [Reynolds 94b].

Ce modèle de mélange est classique dans le domaine de la reconnaissance de forme car il correspond à une situation où les données appartiennent à un ensemble de *classes* distinctes, avec une probabilité d'appartenance propre à chaque classe. Le cas particulier considéré ici est celui où dans chaque classe les données suivent une loi gaussienne. Ce choix tient essentiellement

[3]Dans le domaine de la reconnaissance du locuteur la terminologie utilisée pour décrire ce modèle est souvent variable. En particulier, dans [Matsui 92] et [Matsui 94b], c'est bien ce modèle qui est décrit comme étant un modèle de Markov continu à un seul état. Cette dénomination n'est pas à conseiller car le modèle de Markov est fondamentalement un modèle séquentiel (où l'ordre d'apparition des données est pris en compte) alors qu'il devient un modèle global dans sa version dégénérée à un seul état. De même, la dénomination introduite dans [Tseng 92] est moins explicite que le terme de mélange de densités gaussiennes utilisé par [Reynolds 94b].

au fait que la loi gaussienne appartient à une famille de distributions (dite exponentielles) pour lesquelles le problème de l'identification des composantes du mélange se trouve simplifié [Duda 73, §6], [Redner 84]. Pour le signal de parole, ce modèle ne paraît donc pas déraisonnable et il est d'autre part assez proche de la caractérisation fournie par la quantification vectorielle (cf. paragraphe précédent). La différence étant qu'avec la quantification vectorielle, on se contente de mettre en évidence un certain nombre de "points d'accumulation" des paramètres mesurés, alors qu'avec le modèle de mélange de densités gaussiennes, on cherche en plus à décrire la distribution des paramètres mesurés autours de ces points d'accumulation [Duda 73, §6.6].

Dans le cadre de la reconnaissance du locuteur, l'identification des paramètres du modèle est toujours réalisée grâce à l'*algorithme EM* [Rose 90], [Rose 94], [Tseng 92] qui recherche de manière itérative les paramètres permettant de maximiser localement la vraisemblance des données d'apprentissage [Dempster 77], [Redner 84]. La mesure de similarité est obtenue par calcul de la vraisemblance du jeu de mesures à tester (en pratique on utilise plutôt le logarithme de la vraisemblance) compte tenu du modèle déterminé avec les données d'apprentissage [Reynolds 94b].

Nous avons déjà eu l'occasion d'évoquer les performances de ce type de techniques au cours des paragraphes précédents. Rappelons ici simplement les conclusions : en mode indépendant du texte (avec un vocabulaire non contraint), et lorsque les données disponibles pour l'apprentissage sont suffisantes, le modèle de mélange gaussien permet d'obtenir de meilleures performances que celles des autres techniques décrites (classificateur gaussien et quantification vectorielle). Cependant, lorsque la durée des enregistrements utilisés pour la phase d'apprentissage est faible (inférieure à 20 secondes) la méthode utilisant le modèle de mélange de densités gaussiennes semble moins efficace compte tenu du nombre important de paramètres qu'il est nécessaire d'estimer.

Plusieurs points méritent d'être étudiés concernant ce modèle de mélange de densités gaussiennes. Le premier concerne la structure des densités gaussiennes composant le mélange. Une simplification souvent utilisée consiste à considérer que les densités gaussiennes composant le mélange possèdent toutes une matrice de covariance diagonale [Matsui 92], [Reynolds 94b], [Tseng 92]. Cette simplification est plus réaliste compte tenu de la difficulté posée par l'estimation complète des matrices de covariance [Rabiner 89, §VI.E]. De plus, cette hypothèse ne semble pas déraisonnable compte tenu du fait que les coefficients cepstraux sont pratiquement décorrélés (cf. paragraphe 2.2.1) [Tseng 92]. Les résultats obtenus dans [Tseng 92] semblent toutefois indiquer que cette restriction contribue à dégrader légèrement les performances de la reconnaissance. Pour vérifier ce résultat, il faudrait, à durée d'apprentissage égal, comparer les bénéfices respectifs de l'utilisation de matrices de covariance complètes et d'une augmentation équivalente du nombre de composantes du mélange. En général, c'est le second élément qui est considéré comme plus bénéfique [Rabiner 89, §VI.E].

Un autre point important concerne la méthode d'apprentissage. On sait en effet que le problème de l'identification des composantes du mélange est par nature très complexe [Duda 73, §6.4]. De plus, l'algorithme EM est susceptible de fournir de multiples solutions avec, qui plus est, une convergence très lente [Redner 84]. Le problème de l'initialisation de l'algorithme d'apprentissage est donc très important. Dans les applications de reconnaissance du locuteur, on trouve à la fois des méthodes d'initialisation très simples (partition arbitraire [Matsui 92]) et des méthodes plus élaborées, et certainement plus performantes, (détermination initiale des paramètres à l'aide d'une procédure de quantification vectorielle [Rose 91]). Il serait très intéressant de vérifier l'influence de la méthode d'initialisation utilisée[4].

[4] [Rose 90] présente une comparaison entre deux procédures d'initialisation dont les résultats mériteraient d'être vérifiés compte tenu de l'hypothèse très restrictive imposée au modèle dans cette étude (matrice de covariance commune à toutes les composantes du mélange).

2.2.2.f Modélisation par modèle de Markov caché

Technique paramétrique séquentielle.
Références : [Savic 90], [Rosenberg 90], [Tishby 91], [Rosenberg 91], [Matsui 92], [Webb 93] ainsi que [Rabiner 89] et [Rabiner 93] pour ce qui concerne plus spécifiquement les modèles de Markov cachés et leurs applications dans le domaine du traitement de la parole.

Un défaut commun à la plupart des techniques présentées précédemment (classificateur gaussien, caractérisation par quantification vectorielle et modélisation par mélange de densités gaussiennes) est le caractère global : ces techniques ne tiennent aucun compte de l'ordre dans lesquelles sont présentées les fenêtres de signal. Pour le modèle de mélange de densités gaussiennes ainsi que pour le classificateur gaussien, on suppose même que les paramètres mesurés dans des fenêtres distinctes sont statistiquement indépendants. En pratique, cette hypothèse n'est pas vérifiée car les mesures effectuées dans des fenêtres voisines ne sont pas indépendantes. Une méthode permettant de prendre en compte certains aspects séquentiels, qui s'est avérée très efficace dans le cadre de la reconnaissance de la parole, consiste à utiliser un modèle de Markov caché (*Hidden Markov Model* ou HMM)[5].

Le modèle de Markov caché est un modèle statistique séquentiel qui suppose que les caractéristiques observés forment une succession d'*états* distincts. Un tel modèle est entièrement caractérisé par la donnée de trois éléments :

1. Les *probabilités initiales* de se trouver dans chaque état.
2. Les *probabilités de transition* qui décrivent les passages possibles entre les différents états.
3. Les *probabilités de sortie* qui à proprement parler représentent les distributions conditionnelles des caractéristiques observées en fonction de l'état du modèle.

Il existe plusieurs types de modèles de Markov qui correspondent aux différents choix possibles en ce qui concerne le second, et surtout, le troisième élément du modèle. Pour la reconnaissance du locuteur, le choix le plus fréquent consiste à utiliser un modèle de Markov "continu" où la distribution conditionnelle dans chaque état est un mélange de densités gaussiennes [Matsui 92], [Rosenberg 90], [Webb 93]. Un autre choix usuel consiste à utiliser un modèle de Markov dit "autorégressif" [Savic 90] pour lequel les paramètres de départ sont directement les échantillons du signal qui sont supposés suivre un modèle AR dépendant de l'état du modèle ([Tishby 91] présente une extension de cette démarche à un mélange de modèles autorégressifs).

Un résultat très important concernant l'utilisation de modèles de Markov cachés est le fait qu'*en mode indépendant du texte, l'information supplémentaire apportée par les transitions entre états n'améliore pas les performances de la reconnaissance du locuteur* [Matsui 92] (repris par [Matsui 94a]). Notons que les conclusions présentées dans [Tishby 91] vont dans le même sens alors même que le vocabulaire utilisé est fortement contraint (limité à des suites de chiffres). Ceci indique donc que la complexité accrue d'un modèle de Markov par rapport à son équivalent non-séquentiel (c'est à dire le mélange de densités gaussiennes pour un modèle de Markov continu) ne se justifie que lorsque le texte est fixé, ou tout au moins lorsque le vocabulaire disponible est très restreint. Il serait d'ailleurs intéressant de comparer, dans ces dernières conditions, les performances respectives des techniques utilisant un modèle de Markov et de celles fondées sur un "double" modèle global (avec caractéristiques instantanées et variationnelles).

[5]En fait, il existe un autre moyen de prendre en compte l'aspect séquentiel, même avec un modèle de type global, qui consiste à adjoindre des paramètres variationnels (cf. paragraphe 2.2.1) aux paramètres "instantanés" usuels [Soong 88], [Tseng 92].

2.2.2.g Autres méthodes récentes

Parmi les méthodes étudiées récemment pour évaluer la similarité entre deux jeux de caractéristiques dans le cadre de la reconnaissance du locuteur, on peut citer l'utilisation de la quantification matricielle [Chen 93], [Rosenberg 91] ainsi que la modélisation autorégressive multidimensionnelle [Bimbot 93], [Furui 94]. Ces méthodes constituent toutes deux des généralisations séquentielles de méthodes globales (respectivement, la quantification vectorielle et le quantificateur gaussien). Par analogie avec ce que nous avons vu au paragraphe précédent concernant le modèle de Markov, on peut donc penser que ces techniques sont plutôt susceptibles d'apporter des améliorations dans le cadre des applications en mode dépendant du texte ou à vocabulaire restreint.

Depuis quelques années, on aussi trouve plusieurs exemples d'applications de réseaux dit "neuromimétiques" dans le cadre de la reconnaissance du locuteur [Farell 94], [Furui 94].

2.2.3 Modes de décision

Dans ce paragraphe, nous considérons la dernière partie de la figure 2.1, c'est à dire la prise de décision. La stratégie mise en jeu dans cette partie dépend essentiellement du type d'application : identification ou vérification.

2.2.3.a Pour l'identification

La stratégie est assez simple puisqu'il s'agit d'évaluer la similarité des caractéristiques mesurées avec toutes les références correspondant à chacun des locuteurs autorisés. Le locuteur identifié est celui pour lequel la similarité est la plus grande [Furui 94]. Notons que le coût de calcul de cette opération d'identification, ainsi que le volume des données qu'il est nécessaire de stocker, croissent linéairement avec la taille du groupe de locuteurs autorisés.

La situation est plus complexe lorsqu'on a affaire à un groupe ouvert car il est en plus nécessaire de rejeter les locuteurs n'appartenant pas au groupe de locuteurs autorisés. En général, la démarche adoptée consiste à effectuer d'abord l'identification, puis à utiliser une stratégie de vérification pour rejeter les éventuels imposteurs (en considérant que l'identité revendiquée est celle déterminée lors de la phase d'identification). Il existe toutefois d'autres stratégies raisonnables pour vérifier si un locuteur appartient bien au groupe de locuteurs autorisés [Webb 93].

On trouve peu de variations autour de cette stratégie de base si ce n'est dans [Noda 89] qui propose une modification (destinée à la criminalistique) visant non pas à désigner un seul locuteur mais un ensemble de locuteurs "probables". Cette modification a pour effet d'augmenter la fiabilité de la décision au détriment de sa précision, puisque l'on se retrouve avec un ensemble de locuteurs potentiels. Les résultats présentés dans [Webb 93] suggèrent d'ailleurs que cette méthode très efficace : avec la base de données utilisée dans [Webb 93], le taux d'identification erronée est de l'ordre de 2% et il diminue de moitié dès lors que l'on désigne non plus un locuteur mais deux locuteurs potentiels (les deux locuteurs pour lesquels la mesure de similarité est la plus grande). [Gish 94] présente plusieurs rechniques destinées à accroître la robustesse de l'identification notamment en segmentant les données de test et en combinant les scores obtenus sur les différents segments.

Une autre méthode pour augmenter la fiabilité de la reconnaissance (qui est applicable aussi bien pour la vérification) consiste à utiliser une stratégie séquentielle de décision, c'est à dire à suspendre la décision lorsque la similarité évaluée n'est pas suffisante [Furui 81a]. Ceci suppose bien sûr qu'il soit possible d'obtenir plusieurs enregistrements de test, par exemple en demandant au locuteur de réitérer. Pour certaines des applications où cette possibilité n'existe pas, il peut

aussi être raisonnable d'autoriser une décision "nulle" (pas de décision de reconnaissance dans les cas litigieux).

2.2.3.b Pour la vérification

L'attitude adoptée en général consiste à fixer un seuil sur la mesure de similarité : au dessus, le locuteur est rejeté, en dessous, le locuteur est accepté (comme étant celui dont l'identité est revendiquée) [Furui 94]. Toutefois, comme le montre de manière très claire [Noda 89], *l'utilisation d'un seuil fixé, identique pour tous les locuteurs, conduit à des taux d'erreur variables en fonction du locuteur.*

Une réponse très courante à ce problème consiste à fixer les seuils individuels de vérification *a posteriori*. Les seuils sont alors évalués grâce à des tests systématiques avec tous les locuteurs et tous les enregistrements disponibles pour l'apprentissage. On fixe en général le seuil de façon à obtenir des taux de rejets erronés et d'acceptations erronées de même valeur (cf. paragraphe 1.2.1). [Furui 81a] présente une stratégie empirique alternative visant à remplacer cette procédure très coûteuse, et surtout peu réaliste compte tenu du fait qu'elle suppose que tous les imposteurs potentiels soient connus.

Récemment, une autre solution a été proposée qui consiste à *normaliser les mesures de similarité*, le seuil de décision restant lui fixé indépendamment du locuteur [Furui 94], [Matsui 93], [Matsui 94b], [Reynolds 94b]. L'intérêt principal de cette normalisation est de réduire les différences individuelles intrinsèques qui rendent nécessaires l'usage de seuils individuels. De plus, cette méthode semble aussi limiter les conséquences de la variabilité "accidentelle" introduite par le canal de transmission [Furui 94]. Le principe de la normalisation de la mesure de similarité consiste à soustraire, à la mesure de similarité obtenue pour le locuteur dont l'identité est revendiquée, une mesure de similarité moyenne obtenue pour un groupe de locuteurs représentatifs (dit *cohort speakers*). Le débat reste ouvert quant à la manière dont il convient de composer ce groupe représentatif [Furui 94], [Reynolds 94b]. [Matsui 94b] présente une nouvelle approche, qui semble être assez efficace, consistant à effectuer la normalisation, non plus avec un groupe de locuteurs représentatifs, mais directement avec un modèle représentatif (dans le cadre du modèle de mélange de densités gaussiennes).

2.2.4 Prise en compte la variabilité

2.2.4.a Variabilité intra-locuteur

Un résultat très important est le fait que les caractéristiques mesurées pour un même locuteur varient de manière notable au cours du temps : plus l'intervalle de temps séparant les dates des enregistrements d'apprentissage et de test augmente, plus les performances de reconnaissance se dégradent [Furui 86], [Furui 81b], [OShaughnessy 86], [Rosenberg 91], [Soong 87]. Lorsque le système de reconnaissance est destiné à une utilisation répétée, cet effet peut être atténué par une mise à jour régulière des données de références de chaque locuteur [Doddington 85], [Rosenberg 76], [Rosenberg 91].

D'une manière générale, on obtient de meilleurs résultats en réalisant l'apprentissage sur plusieurs enregistrements distincts [Furui 81a], si possible enregistrés lors de séances suffisamment séparées dans le temps [Furui 81b], [Soong 87]. Cependant, pour les applications où l'adaptation régulière des références est impossible, il faut s'attendre à une diminution des performances de reconnaissance à long-terme (à l'échelle de plusieurs mois).

Enfin, même une adaptation régulière peut se révéler insuffisante face à des variations très circonstancielles des caractéristiques du locuteur dues, entre autres, à son état de santé [Rosenberg 76].

2.2.4.b Variations du canal de transmission

Nous avons déjà souligné que quel que soit le type d'application envisagée (identification ou vérification), il est nécessaire de prendre des mesures particulières dès lors que le canal de transmission est susceptible de varier (en particulier pour les enregistrements réalisés par téléphone). La méthode la plus efficace connue à ce jour est la *normalisation cepstrale* qui ne peut être mise en œuvre que si l'on utilise des paramètres de type cepstraux. Cette normalisation consiste simplement à soustraire aux paramètres cepstraux la valeur moyenne calculée sur tout l'enregistrement disponible (cf. paragraphe 2.2.1) [Furui 81a], [Furui 94], [Gish 86] et [Gish 90], [Reynolds 94b], [Rose 90]. Il semble que cette méthode de normalisation cepstrale reste plus efficace que d'autres méthodes proposées récemment [Reynolds 94a][6].

Toutefois, si cette méthode est satisfaisante lorsque le texte est fixé [Furui 81a], ainsi que dans les cas où la durée de l'enregistrement de test est suffisamment longue [Reynolds 94b] (plus de 20 secondes), elle devient contestable dans les cas inverses (texte libre et de courte durée) [Gish 86], [Rosenberg 91]. En effet, dans ces dernières conditions, la quantité déterminée par moyenne des paramètres cepstraux présente elle même une très forte variabilité, liée entre autres au texte prononcé. Dans le cas de textes courts (en mode indépendant du texte), la normalisation cepstrale contribue donc à augmenter la variabilité ce qui conduit à une dégradation des performances. Malgré ces défaut, il faut cependant souligner qu'il n'existe pas à l'heure actuelle d'alternative, applicable en pratique, à la méthode de normalisation cepstrale pour les enregistrements téléphoniques (voir en particulier [Krasner 84], [Gish 85], [Gish 86] et [Gish 90]).

Il faut nuancer ce jugement en notant que toutes les techniques visant à accroître la robustesse des mesures de similarité (voir, par exemple, [Gish 90] et [Matsui 91]), et de manière plus générale, la recherche de modèles plus robustes permettent de limiter l'effet des variations du canal de transmission. D'autre part, ces variations sont aussi prises en compte par la méthode de normalisation de la mesure de similarité mentionnée plus haut.

Une dernière remarque est que tout ce qui vient d'être évoqué ne concerne que le cas de variations supposées linéaires du canal de transmission. Un autre cas très important de variation du canal de transmission est lié à la présence de bruit de fond[7]. Sur ce point on peut consulter [Reynolds 92], [Rose 91] et [Rose 94] qui présentent une démarche très complète où le bruit de fond est intégré explicitement au le modèle utilisé pour la reconnaissance. Une des difficultés rencontrées dans ces publications est le fait que l'influence du bruit de fond sur les paramètres utilisés habituellement (coefficients cepstraux, ou log-énergie de banc de filtre) est "complexe" (elle se traduit par des modifications non-linéaires).

2.3 Performances et problèmes actuels

Nous avons eu l'occasion tout au long de cette partie consacrée aux techniques de reconnaissance automatique du locuteur de souligner les différents problèmes auxquels sont confrontés les chercheurs travaillant dans ce domaine. D'une manière générale, l'amélioration des techniques utilisées permet maintenant d'envisager des conditions de fonctionnement de plus en plus proches

[6] La référence [Sankar 94] (consacrée à la reconnaissance de la parole) présente une technique de compensation plus élaborée que la simple normalisation cepstrale puisqu'elle prend en compte les caractéristiques du modèle utilisée. D'un point de vue méthodologique, cette technique repose sur le même principe que celle étudiée dans [Rose 94] pour le bruit de fond.

[7] D'une manière générale, la prise en compte du bruit de fond à fait l'objet de nettement moins d'études dans le cadre de la reconnaissance du locuteur que dans celui de la reconnaissance de la parole. L'explication en est qu'une des applications phare de la reconnaissance du locuteur est le paiement téléphonique pour lequel le problème rencontré est principalement celui des variations linéaires.

des situations rencontrées en pratique (enregistrements de courte durée, variations du canal de transmission, présence de bruit de fond, nombre important de locuteurs autorisés, etc.). Toutefois, nous avons vu que les solutions à ces différents problèmes, telles qu'elles sont utilisées à l'heure actuelle, présentent encore de nombreux défauts. Ceci explique que la reconnaissance du locuteur reste un sujet de recherche et ne saurait être considérée comme une technique acquise.

Plutôt que de mentionner de nouveau l'ensemble des problèmes posés par la reconnaissance du locuteur, nous avons choisi d'insister sur un défaut structurel qui nous semble être actuellement un des points faibles de la recherche dans le domaine : celui de l'*évaluation*. Il faut noter que nous n'avons pratiquement pas fourni de données chiffrées concernant l'efficacité des différentes méthodes de reconnaissance du locuteur. Il nous semble en effet que l'évaluation des méthodes soulève de très sérieuses interrogations.

Nous avons vu au paragraphe 1.3 que l'évaluation des méthodes de reconnaissance du locuteur s'effectue actuellement systématiquement par une démarche qui peut être qualifiée d'empirique (constitution d'une base de données, tests systématiques de reconnaissance, présentation des taux d'erreur). L'interprétation des données chiffrées fournies par cette évaluation empirique est très délicate, et les résultats obtenus ne permettent pas forcément de répondre à certaines questions fort importante en pratique.

2.3.1 Comparaison des méthodes entre elles

L'évaluation empirique permet de comparer des méthodes entre elles uniquement si on utilise systématiquement la même base de données et dans les mêmes conditions. C'est une des raisons pour lesquelles la constitution de bases de données spécifiques à la reconnaissance du locuteur, ainsi que leur diffusion auprès du plus grand nombre de chercheurs sont devenus des aspects très important du domaine [Bimbot 94a], [Godfrey 94]. L'acquisition des bases de données spécialisées décrites dans [Godfrey 94] est fortement conseillée pour toute équipe de recherche travaillant dans le domaine. On ne peut d'ailleurs que se féliciter de voir un nombre toujours croissant d'articles qui présentent des résultats obtenus sur ces bases de données "classiques" [Bimbot 94b], [Farell 94], [Kao 93], [Reynolds 94b], [Wagner 94], [Yu 93]. Il ne faut toutefois par perdre de vue le fait que certaines de ces bases de données n'ont pas été constitué spécifiquement pour la reconnaissance du locuteur. Par conséquent, les évaluations obtenues avec celles-ci ne doivent pas être considérée comme réalistes (c'est notamment le cas pour la base de données TIMIT [Reynolds 94b]).

Par ailleurs, il faudrait définir une manière standard d'utiliser ces bases de données. Ainsi, le plus souvent, tous les locuteurs autorisés sont utilisés tour à tour comme imposteur pour chacun des autres locuteurs autorisés (dans le cas d'une application de vérification). Cette manière de faire n'est pas vraiment réaliste et il serait sûrement plus prudent de prévoir deux groupes (locuteurs autorisés et imposteurs) séparés [Bimbot 94a].

Cependant, l'utilisation de bases de données communes ne résout pas complètement le problème car la nature empirique de la démarche d'évaluation ne fournit pas les moyens d'analyser les différences de performances constatées. Pour cette raison, les comparaisons entre méthodes de reconnaissance de locuteur présentées dans la littérature donnent souvent des résultats assez décevants et de portée limitée. Pour faire progresser les connaissances dans ce domaine, il serait très bénéfique d'effectuer des expériences visant à déterminer le degré d'adéquation réel des modèles utilisées avec les données de départ (les paramètres cepstraux mesurés sur les signaux de parole). Il est d'autre part important de mette en évidence, lorsque c'est possible, les liens théoriques existants entre les différentes techniques ou modèles utilisés.

2.3.2 Nature des erreurs

Un autre problème lié à l'évaluation est le fait que le taux d'erreur semble masquer une réalité plus complexe. Nous avons déjà signalé que de nombreuses références indiquent que les erreurs de reconnaissance ne se répartissent pas uniformément selon les locuteurs : la majorité des locuteurs présentant un faible taux d'erreur tandis qu'un faible groupe de locuteur présente un taux d'erreur important [Doddington 85], [Furui 81a], [Oglesby 94], [Soong 87]. Cette situation est assez préoccupante, et elle met bien en évidence le fait qu'à l'heure actuelle, il existe très peu de résultats permettant d'analyser la nature des erreurs commises par les systèmes de reconnaissance du locuteur[8].

Cette question est très importante car elle se traduit en pratique par le fait que le système fonctionne moins efficacement pour certains locuteurs [Noda 89]. Il serait fortement souhaitable de pouvoir prévoir ce type de comportement. En particulier, il serait très utile de faire la part entre ce qui est lié à *des facteurs passagers* (à savoir si le locuteur est plus ou moins coopérant, plus ou moins attentif, s'il est fréquemment sujet à des affections de la voix, s'il est habitué à l'utilisation du système de reconnaissance) et ce qui provient des *caractéristiques intrinsèques* du locuteur (fréquence fondamentale grave ou aiguë, particularités d'élocution, etc.). Les deux types de facteurs semblent entrer en jeu [Doddington 85], [Noda 89], [Wagner 94].

2.3.3 Validité de l'évaluation

A l'heure actuelle, la plupart des publications concernant la reconnaissance du locuteur rapportent des performances d'un très bon niveau : taux d'erreur inférieurs à 15%, et même, très souvent, inférieurs à 5%. La comparaison entre plusieurs techniques ainsi que l'influence des divers paramètres se traduisent donc en général par des variations de l'ordre de quelques pourcent. Malheureusement, ces variations sont du même ordre (et même souvent bien inférieures) que celles obtenues en changeant la base de données utilisées par l'évaluation.

[Reynolds 94b] présente une illustration très intéressante de ce phénomène pour une application d'identification du locuteur sur un groupe de 100 personnes (en utilisant le modèle de mélange de densité gaussienne). Les résultats obtenus pour trois bases de données classiques différentes sont les suivants :

Base de données	Canal de transmission	Conditions d'enregistrement	Taux d'erreur
TIMIT	enregistrement direct	au cours d'une seule session	$< 1\%$
NTIMIT	enregistrement téléphonique (poste fixé, liaison de mauvaise qualité)	au cours d'une seule session	20%
SWITCHBOARD	enregistrement téléphonique (différents postes téléphoniques)	pendant plusieurs semaines	16%

Ces résultats montrent tout d'abord que la baisse de qualité des enregistrements due à l'utilisation du téléphone se traduit par une énorme dégradation des performances (comparer les performances obtenues avec la base de données TIMIT et avec les deux autres). De plus, le décalage entre les performances obtenues avec les bases de données NTIMIT et SWITCHBOARD souligne l'importance de facteurs qui peuvent apparaître à première vue comme secondaires (qualité de la transmission, variabilité due à l'usage de différents types de postes téléphoniques, enregistrement en différentes occasions).

[8] On trouvera quelques résultats très intéressants sur ce sujet notamment dans [Webb 93], et [Wagner 94]. Les résultats obtenus dans cette dernière référence étant d'ailleurs assez troublants ...

Un autre exemple analogue de ce type de variations est rapporté par [Webb 93], il s'agit du fait que les performances d'un système de reconnaissance sont toujours moins bonnes lorsque l'on utilise que des locuteurs du même sexe. Il est donc possible "d'augmenter artificiellement" les performances d'un système donné en l'évaluant sur une base de données composée de locuteur des deux sexes (ce qui est très souvent fait en pratique). Il est clair que pour une application de type vérification, une telle base de données est peu réaliste car un éventuel imposteur masculin essaiera, en général, plutôt de se faire passer pour un des locuteurs masculins autorisés. En allant plus loin, il faudrait aussi prendre en compte le fait qu'un éventuel imposteur va chercher à se faire passer non pas pour un des locuteurs autorisés choisi au hasard, mais plutôt pour celui dont la voix est la plus susceptible d'être confondu avec la sienne [Bimbot 94a].

Un dernier exemple, de variation des performances de la reconnaissance en fonction de la difficulté de la tâche nous est fournie par [Webb 93] et [Reynolds 94b]. Il s'agit du fait que les performances de la reconnaissance *dans une application d'identification* se dégradent lorsque le nombre de locuteurs augmente. Ceci constitue déjà un problème en soi, qui est aggravé par le fait que cette décroissance des performances est très différente selon les conditions dans lesquelles sont réalisés les enregistrements. Ainsi dans [Webb 93], la variation du taux d'erreur quand le nombre de locuteur augmente de 50 à 300 est quasiment négligeable lorsque le même type de microphone téléphonique est utilisée lors de la phase d'apprentissage et lors du test. Au contraire, lorsque les microphones sont différents lors de l'apprentissage et lors du test, le taux d'erreur va jusqu'à augmenter de plus de 15% (passant de 4% pour 50 locuteurs à 20% pour 300 locuteurs). De même, dans [Reynolds 94b], l'augmentation du taux d'erreur lorsque l'on passe de 20 à 160 locuteurs est inférieure au pour-cent pour la base de données TIMIT alors qu'elle est de 17% (on passe de 6% à 23%) pour NTIMIT (cette différence étant entièrement attribuable à l'utilisation du téléphone, cf. tableau 2.3.3).

Il faut donc être très prudent avec les données chiffrées obtenues lors d'une évaluation empirique. Les conclusions présentées en 1985 par G. Doddington dans [Doddington 85] restent tout à fait d'actualité : "[...] Il n'est pas raisonnable de s'attendre à ce qu'il soit possible d'atteindre des performances arbitrairement bonnes à la seule condition d'améliorer la mesure des caractéristiques du locuteur et de développer les algorithmes de traitement. [...] les performances d'un système de reconnaissance du locuteur dépendent du degré de contrôle qu'il est possible d'exercer sur les conditions dans lequel il fonctionne" (traduit de l'anglais). On pourrait même ajouter que dans de nombreux cas, les performances dépendent beaucoup plus des conditions expérimentales que de la technique particulière utilisée (en se limitant aux techniques "raisonnables").

En reconnaissance du locuteur, il est donc extrêmement important de préciser la nature de l'application envisagée. Avant même d'avancer des chiffres concernant les performances d'une méthode, il serait important d'être capable de quantifier la difficulté intrinsèque des situations dans lesquelles elle est susceptible de fonctionner. Une caractérisation purement théorique de cette difficulté semble vouée, eu égard aux connaissances actuelles, à faire appel à des modèles plus ou moins arbitraires [Oglesby 94]. Une solution intéressante consisterait à calibrer les bases de données par rapport aux performances de reconnaissance du locuteur obtenues par les auditeurs humains [Doddington 85], [Bimbot 94a]. Malheureusement, une telle démarche est très lourde à mettre en œuvre car elle nécessite des tests auditifs extensifs. De plus, étant donné l'état actuel des connaissances concernant les processus complexes mis en jeu lors de la reconnaissance par des auditeurs humains, l'interprétation de tels tests constitue en elle même une difficulté non négligeable.

Chapitre 3

Reconnaissance du locuteur en criminalistique

3.1 Spécificités de l'application en criminalistique

Le paragraphe suivant présente ce qui constitue en fait le point le plus important de cette seconde partie, à savoir, les raisons pour lesquelles le domaine criminalistique présente plusieurs caractéristiques très différentes, voire incompatibles, avec les applications "usuelles" de reconnaissance du locuteur telles que nous les avons évoquées jusqu'à présent [Hollien 90], [Kunzel 94]. Cette partie reprend plusieurs des arguments et exemples présentés dans [Kunzel 94].

3.1.1 Situation-type d'expertise judiciaire

En France, c'est le juge d'instruction chargé d'une procédure judiciaire qui désigne, si besoin est, un expert (éventuellement à la demande d'une des parties). Une fois le rapport de l'expert remis, une des parties peut, si elle l'estime nécessaire, demander une contre-expertise à un autre expert (pour plus d'information concernant la désignation et le statut des experts, on peut consulter [Buquet 91, Chap. IX]). Le cadre juridique est très différent dans des pays comme les Etats-Unis et la Grande Bretagne où les deux parties peuvent faire appel à leur propre expert [Kunzel 94]. La situation la plus fréquente dans le cadre de l'expertise judiciaire est la suivante :

1. La pièce à expertiser proprement dite est constituée par un **enregistrement de question**, enregistré au moment des faits, en général dans de mauvaises conditions techniques (le plus souvent, sur un répondeur téléphonique) et dans des situations psychologiques très particulières (menaces téléphoniques, demande de rançon, appel d'une personne kidnappée, etc.). Il s'agit la plupart du temps d'un enregistrement assez court (une vingtaine de secondes ou moins). Notons qu'il existe parfois le cas inverse (c'est notamment le cas pour les enregistrements provenant d'écoutes téléphoniques) où les documents à expertiser représentent un volume problématique (plusieurs heures de conversations téléphoniques). Assez souvent, la voix enregistrée est volontairement altérée, éventuellement par un dispositif particulier.

Les statistiques du laboratoire du BKA[1] citées dans [Kunzel 94] permettent de se faire une idée de la mauvaise qualité des enregistrements généralement soumis à expertise : dans 95% des cas il s'agit d'enregistrements téléphoniques, dans 20% des cas l'enregistrement dure moins de 20 s et dans 15% des cas, le locuteur cherche manifestement à déguiser sa voix.

2. On dispose par ailleurs d'un **enregistrement de comparaison** de l'auteur présumé de l'enregistrement de question. En général, l'enregistrement de comparaison est toujours nettement postérieur à celui de question (au minimum de plusieurs mois, et parfois de plusieurs années) et souvent enregistré dans des circonstances psychologiques difficiles (l'enregistrement est en général effectué à la demande de l'expert alors que la personne est détenue). Dans la plupart des cas, il semble illusoire de faire répéter le texte exact de l'enregistrement de question, l'enregistrement de comparaison contient donc une conversation dont le sujet n'a pas de rapport avec l'enregistrement de question.

3. Le rôle de l'expert consiste à valider, ou bien à exclure, l'hypothèse selon laquelle les deux enregistrements correspondent bien au même locuteur. Il lui est de plus demandé de quantifier la fiabilité de sa conclusion en fournissant un pourcentage de confiance.

La tâche confiée à l'expert judiciaire est donc proche de la *vérification du locuteur en mode indépendant du texte*. Toutefois, le fait que le locuteur ne soit pas coopérant doit être pris en compte [Kunzel 94] : contrairement aux applications classiques de vérification du locuteur, l'auteur de l'enregistrement de question ne déclare pas son identité, il n'a même en général aucun intérêt à être reconnu (en particulier, la notion d'imposteur n'est pas vraiment pertinente dans ce contexte). Ainsi dans le cadre de l'expertise judiciaire, le problème de la *variabilité intra-locuteur* des caractéristiques de la voix devient crucial puisque les locuteurs n'ont pas de raison de s'astreindre à une élocution relativement homogène (et même souvent intérêt à faire le contraire). Une autre particularité de la criminalistique est le fait que l'on travaille souvent en groupe totalement ouvert : il n'y a en général pas de groupe de locuteurs autorisés, connus a priori, comme dans le cas de certaines applications commerciales [Kunzel 94].

3.1.2 Contraintes propres au domaine

Par rapport aux applications commerciales "usuelles", l'application en criminalistique présente la particularité de cumuler la plupart des facteurs négatifs susceptibles de limiter l'efficacité des techniques de reconnaissance. Ces facteurs négatifs, qui sont liés aux conditions dans lesquelles sont réalisées les enregistrements dans le domaine judiciaire, sont principalement :

Mauvaises conditions d'enregistrements Le fait que les conditions d'enregistrement de la pièce de question ne soient pas maîtrisées pose de sérieux problèmes car la variabilité due au canal de transmission peut devenir très importante par rapport à la variabilité inter-locuteur. En particulier, nous avons vu au paragraphe 2 que certaines techniques qui fonctionnent pour des enregistrement effectués "en direct" deviennent peu efficaces lorsqu'elles sont utilisées pour des enregistrements téléphoniques [Doddington 85], [Furui 81a], [Gish 85], [Gish 86], [Kunzel 94]. Ceci est lié au fait qu'en plus de contribuer à filtrer le signal de parole, l'utilisation du téléphone constitue une source importante de variabilité. Les variations constatées sont surtout dues à

[1]Le laboratoire du BKA (pour *BundesKriminAlamt*, ce qui correspond à la Direction Centrale de la Police Judiciaire en France) est le laboratoire central de police scientifique allemand qui est situé à Wiesbaden. Ce laboratoire comprend un département spécialisé dans la reconnaissance du locuteur. Ce département, le plus grand du genre en Europe, a traité environ 2000 cas depuis le début des années 80 (d'aprèes [Kunzel 94]).

la très médiocre qualité des microphones téléphoniques mais aussi à la diversité du matériel téléphonique, y compris de la liaison elle même (voir l'étude très complète réalisée aux Etats-Unis et décrite dans [Carey 84]). Les performances de la reconnaissance sont aussi affectées par la qualité du matériel d'enregistrement (c'est surtout vrai pour le microphone) ainsi que par la présence éventuelles de sources de bruit au moment de l'enregistrement. Dans le domaine de la criminalistique, la qualité des appareils d'enregistrement est souvent très mauvaise, notamment lorsqu'il s'agit d'écoutes téléphoniques en général réalisées avec des dispositifs à sauvegarde de bande (faible vitesse de défilement de la bande magnétique).

Ancienneté de l'enregistrement de question Le fait que les enregistrements à comparer ne soient pas contemporains constitue une difficulté très importante car les caractéristiques d'un locuteur varient de manière non négligeable entre deux séances d'enregistrements distantes dans le temps [Furui 86], [Furui 81b], [OShaughnessy 86], [Rosenberg 91], [Soong 87]. Cet effet peut aussi être accentué par des éléments circonstanciels liés, entre autres, à l'état de santé du locuteur [Rosenberg 76], [Kunzel 94] ou à des facteurs externes [Braun 94]. Nous avons signalé (au paragraphe 2.2.4.a) que dans la plupart des systèmes commerciaux de reconnaissance du locuteur, cette difficulté est surmontée par une mise à jour régulière des données de références de chaque locuteur [Doddington 85], [Rosenberg 76], [Rosenberg 91]. Une telle mise à jour n'est possible que pour un locuteur qui utilise relativement fréquemment le système de reconnaissance. En criminalistique, cette possibilité d'adaptation est totalement exclue.

Courte durée des enregistrements La courte durée de l'enregistrement de question concourt aussi à limiter les performances de la reconnaissance en rendant peu fiable la mesure des caractéristiques du locuteur [Gish 86], [Kunzel 94], [Matsui 91]. Notons de plus que la courte durée de l'enregistrement se traduit souvent par un message enregistré possédant un très faible contenu phonétique ce qui ne simplifie pas la mesure de caractéristiques significatives du signal de parole [Soong 87].

Etat psychologique du locuteur La situation psychologique du locuteur au moment de l'enregistrement influe notablement sur les caractéristiques de la parole. En particulier, plusieurs études portant sur la reconnaissance de locuteurs en état de stress ont montré une diminution significative des possibilités de reconnaissance [Hollien 82], [Hollien 90]. D'une manière plus générale, même si le locuteur n'est pas toujours dans un état psychologique très "délicat", le simple fait qu'il ne soit pas coopérant (c'est à dire qu'il ne cherche pas consciemment à parler de manière claire et distincte) est déjà une source notable de variabilité.

Déguisement de la voix La présence éventuelle d'un travestissement volontaire de la voix, éventuellement à l'aide d'un dispositif particulier, peut rendre la reconnaissance très difficile, voire quasiment impossible [Hollien 90] [2]. Notons à cet égard que la commercialisation, auprès du grand public, de matériel audionumérique (synthétiseurs-échantillonneurs, cartes audionumérique pour micro-ordinateurs) risque de rendre les modifications "artificielles" de la voix plus fréquentes dans les années à venir.

Faible représentativité D'une manière générale, les enregistrements dont on dispose dans le domaine criminalistique se caractérisent par une très faible représentativité. Par définition, il n'y a pas de groupe de locuteurs autorisés, il est donc difficile de cerner précisément la répartition

[2] Il est important de se souvenir que dans la littérature concernant la reconnaissance du locuteur, on considère en général le cas de locuteurs coopérants. Le problème du déguisement de la voix est donc vu sous l'angle de l'*imitation* de la part d'un imposteur [Atal 76], [Das 71], [Doddington 85], [Rosenberg 76]. Par contre, dans le cadre de la criminalistique, on a généralement affaire à des locuteurs non-coopérants qui cherchent à travestir leur voix sans pour autant imiter une autre personne. Ce dernier problème a surtout été étudié dans le cas de la reconnaissance perceptive (écoute par un auditeur humain) [Hirson 93], [Hollien 82], [Reich 79], [Reich 81].

des caractéristiques des locuteurs potentiels. De plus, on ne dispose, en général, que d'un seul enregistrement de comparaison, ce qui rend impossible l'estimation du degré de variabilité habituel des caractéristiques du locuteur présumé. Ce manque de connaissances pose un problème pour fixer les seuils de décision dans une application de type *vérification* (voir la pragraphe 2.2.4). Cette question revêt une importance particulière dans le cadre de la criminalistique puisqu'il est nécessaire (idéalement) non seulement de fournir une décision mais aussi d'estimer la fiabilité de cette décision [Noda 89].

3.1.3 Place de la reconnaissance

Il faut retenir que dans le domaine de la criminalistique, le problème rencontré est en général celui de la vérification du locuteur en mode indépendant du texte. La particularité du domaine réside dans le fait que les locuteurs sont le plus souvent *non-coopérants* ce qui entraîne une variabilité très importante des caractéristiques d'un même locuteur. Cette variabilité se trouve, le plus souvent, aggravée par des facteurs externes (enregistrement téléphonique, de courte durée, etc.). Par conséquent, il ne faut pas s'attendre à obtenir dans ce domaine des performances comparables à celles des systèmes de reconnaissance fonctionnant avec des locuteurs coopérants.

Cependant, un point très important souligné dans [Hollien 90] est le fait que le travail confié à l'expert dépasse en général largement le cadre de la reconnaissance du locuteur. Quasiment systématiquement, il est en effet demandé à l'expert de retranscrire les propos enregistrés sur la pièce de question. Cette tâche peut déjà s'avérer très délicate notamment lorsque plusieurs personnes s'expriment simultanément. Pour faciliter ce travail, il peut être utile de disposer d'un bon outil d'édition de signal et d'un moyen de réaliser simplement quelques traitements (filtrage, modification de vitesse de lecture, etc.) [Hollien 90]. Très souvent, il est demandé à l'expert de préciser dans quels conditions techniques l'enregistrement de question a pu être effectué. Il est donc nécessaire de pouvoir mettre en œuvre quelques mesures visant à déterminer, par exemple, si l'enregistrement a été réalisé par téléphone ou bien s'il présente des coupures. Cette dernière tâche peut d'ailleurs devenir quasiment impossible si les coupures ont été réalisées à l'aide d'un matériel approprié de montage. Sur ce dernier point, compte tenu des possibilités actuellement offertes au grand public dans le domaine audionumérique (éditeurs de signal sur micro-ordinateur par exemple) il est à craindre qu'à l'avenir la tâche de vérification de l'authenticité de la pièce de question ne deviennent beaucoup plus problématique (les méthodes présentées dans [Hollien 90] concernent essentiellement le cas des enregistrements sur bande magnétique analogique qui tendent à disparaître à l'heure actuelle).

D'une manière plus générale, *l'expertise "vocale" dans le domaine judiciaire repose très souvent sur des aspects totalement étranger à la reconnaissance du locuteur.* Les divers exemples cités dans [Hollien 90] montrent que dans une proportion non négligeable des cas, une expertise qui se présente à l'origine sous la forme d'une tâche de reconnaissance du locuteur est en fait "résolue" en ayant recours à des arguments totalement différents (présence de montage, incompatibilité entre les techniques d'enregistrements, etc.). En particulier, dans ce domaine, les propos prononcé par les locuteurs sur la pièce à expertiser peuvent avoir en eux-mêmes une valeur d'identification [Hollien 90].

Il faut donc se garder de limiter l'application en criminalistique uniquement à un problème de reconnaissance du locuteur. Dans la suite, nous considérons uniquement les aspects liés à la reconnaissance du locuteur (telle qu'elle a été définie au paragraphe 1) tout en sachant que ceux-ci ne représente en général qu'un élément parmi d'autres dans les cas concrets.

3.2 Recherches sur les méthodes de reconnaissance

Une autre particularité du domaine est le fait qu'il existe une pression très importante de la part des divers acteurs (juges, policiers, opinion publique, etc.) pour utiliser la reconnaissance du locuteur dans le cadre d'expertises judiciaires. Cet état de choses a différentes explications dont la principale est très certainement la *sous-estimation de la difficulté de la tâche de reconnaissance du locuteur par les personnes naïves dans le domaine* : la plupart des gens sont persuadés (à tort comme nous le verrons au paragraphe 3.4) qu'il leur est possible, sans se tromper, d'identifier une personne à l'écoute de sa voix [Kunzel 94]. De plus, le cinéma, la télévision ainsi que la littérature policière (ou d'espionnage) semblent avoir popularisé l'idée qu'il existe effectivement des techniques scientifiques permettant d'analyser et de reconnaître les voix. Cette méconnaissance du problème rend possibles les déclarations les plus farfelues [NS 93], [Vincent 93] malgré les protestations des spécialistes du traitement de la parole [GCP 90].

Cette pression se traduit en pratique par le fait que certaines techniques ont été utilisées, pas forcément en France d'ailleurs, dans le cadre d'expertises judiciaires sans avoir été au préalable suffisamment évaluées. Il en va de même pour certains experts judiciaires dont les compétences ne sont pas toujours avérées de manière indiscutable [Hollien 90], [Kunzel 94]. Afin d'attirer l'attention sur cette situation déplorable, nous avons choisi de débuter cette revue des méthodes de reconnaissance du locuteur utilisées en criminalistique en relatant l'épisode édifiant des *Voiceprints*.

3.2.1 La mésaventure des "voiceprints"

En 1962, aux Etats-Unis, un chercheur prétend avoir trouvé un moyen quasiment infaillible permettant de reconnaître les locuteurs. Cette méthode consiste simplement à comparer visuellement les spectrogrammes d'un même mot (ou d'une même phrase) prononcé par les locuteurs à comparer [Doddington 85], [Kunzel 94], [Hollien 90]. Aussitôt, cette méthode est baptisée par ceux qui la défendent "empreinte vocale"[3] (*voiceprints*).

Malheureusement, très vite des résultats scientifiques viennent démontrer que la comparaison visuelle de spectrogrammes ne constitue pas une méthode fiable de reconnaissance. Les experts utilisant les *voiceprints* rétorquent en expliquant que la comparaison doit être effectuée avec des critères mystérieux, connus des seuls initiés, et qui relèvent du savoir faire de l'expert [Doddington 85], [Kunzel 94], [Hollien 90] ...

A l'heure actuelle, la plupart des chercheurs qui s'intéressent au domaine de la ciminalistique s'accordent pour dénoncer ce procédé. Une évaluation indépendante a même confirmé que cette méthode est encore moins fiable qu'une reconnaissance auditive (qui ne constitue déjà pas une méthode très fiable) [Doddington 85], [Kunzel 94], [Hollien 90]. Compte tenu de ce qui a été dit au paragraphe 2.2.2.a concernant le spectre moyen à long-terme, on peut ajouter que cette méthode n'est pas robuste vis à vis des variations du canal de transmission. On trouvera d'ailleurs plusieurs contre-exemples démontrant l'inefficacité de la comparaison de spectrogrammes comme outil de reconnaissance dans [Doddington 85], [Endres 71] et [Hollien 90]. Il n'empêche que cette méthode a été utilisée à l'occasion de plusieurs centaines de procès aux Etats-Unis et qu'elle demeure encore employée de nos jours dans certains pays [Hollien 90], [Kunzel 94].

[3] Comme le soulignent très justement [Doddington 85] et [Kunzel 94], cette assimilation fallacieuse avec les empreintes digitales a amplement contribué à promouvoir la méthode. Le terme d'empreinte vocale doit être systématiquement banni car il ne correspond pas à la réalité : si il est légitime de parler d'empreinte dans le cas des doigts qui correspondent bien à une caractéristique *physiologique* de l'individu, ce terme n'a aucun sens pour une manifestation largement *comportementale* comme la parole ou l'écriture. D'une manière générale, on peut dire que l'emploi du terme d'empreinte vocale cache souvent des motivations douteuses (voir par exemple [Vincent 93]).

La principale leçon à tirer de cet épisode malheureux est qu'il est indispensable d'évaluer sérieusement toute méthode destinée à être utilisée dans le cadre d'expertises judiciaires. A ce titre, il n'est pas raisonnable que les experts dans ce domaine cherchent à entourer les méthodes qu'ils utilisent d'un certain mystère. Bien au contraire, les seules méthodes qui peuvent être utilisées sont celles qui ont fait l'objet de travaux scientifiques poussés et sur lesquelles se dégagent un large consensus. La mésaventure des *voiceprints* souligne aussi les dangers de démarches pseudo-scientifiques qui exploitent des outils techniques (ici les spectrogrammes) à l'aide d'arguments subjectifs qui eux n'ont pas de fondement scientifique.

3.2.2 Références bibliographiques

Pour rendre compte des différents efforts de recherche dont nous avons trouvé la trace, nous avons choisit de les présenter, par ordre (à peu près) chronologique, sous la forme de brefs résumés.

Philips GmbH (Allemagne)
Projet : AUROS *(AUtomatic Recognition Of Speakers by computer)*
Référence : [Bunge 77]
Description : Système de reconnaissance automatique utilisant des mesures statistiques simples (moyennes). Pas spécialement spécifique pour la criminalistique. Recherche de caractéristiques non modifiées par les transmissions téléphoniques.
Statut : Pas de suites connues.

Rockwell International (sous contrat du *Law Enforcement Assistance Administration*)
Projet : SASIS *(Semi-Automatic Speaker Identification System)*
Référence : [Paul Jr 75]
Description : Système de reconnaissance semi-automatique où l'opérateur identifie et étiquette des événements phonétiques (grâce à un éditeur de signal avec visualisation de spectrogramme). Sur chaque événement localisé, le système réalise des mesures automatiques. Tests de grande échelle effectués sur une base de données comprenant 250 locuteurs couvrant différents styles linguistiques américains.
Statut : Le projet a été abandonné. D'après [Kunzel 94], une des raisons principal de son échec est le fait que seuls des spécialistes en phonétique étaient en mesure de l'utiliser.

Los Angeles County Sheriff's Department
Projet : CAVIS *(Computer Assited Voice Identification System)*
Référence : [Nakasone 88]
Description : Système de reconnaissance automatique utilisant des mesures statistiques simples (moyennes). Recherche de caractéristiques non modifiées par les transmissions téléphoniques.
Statut : Projet abandonné récemment (d'après [Kunzel 94]).

University of Florida, Institute for Advanced Study of Communication Processes
Projet : SAUSI *(Semi-AUtomatic Speaker Identification system)*
Auteurs : H. Hollien, ...
Références : [Hollien 90] et [Doherty 76], [Doherty 78], [Johnson 84]

Description : Système semi-automatique de mesure de caractéristiques du signal de parole (fréquence fondamentale de voisement, spectre moyen à long-terme, mesures de durées, de la vitesse d'élocution ...). Ce en plus d'une pratique importante de l'expertise judiciaire.

Statut : Le système SAUSI est toujours à l'état de recherche, et n'a pas été appliqué à des situations réelles de criminalistique (d'après [Kunzel 94]).

Fondazione Ugo Bordoni (Italie)

Projet : IDEM (*IDEntification Method*)

Auteurs : A. Paoloni, A. Federico, ...

Références : [Falcone 94], [Federico 93] et [Federico 87], [Federico 89]

Description : Système semi-automatique sur PC permettant l'édition de signal, l'étiquetage (visualisation de la forme d'onde, du spectrogramme ...). Inclut un module de décision statistique sur les paramètres mesurés (fréquence fondamentale de voisement, fréquences des formants).

Statut : Système toujours en développement. Pour l'instant, il s'agit d'un projet interne.

International Association for Forensic Phonetics (Grande-Bretagne[4])

Auteurs : P. French, A. Hirson, ...

Références : [Hirson 93], [Hirson 94], [Howard 93]

Description : Recherche sur la reconnaissance du locuteur en criminalistique d'une manière générale. Approche de type "phonétique assistée par ordinateur". Les auteurs travaillent visiblement sur des cas concrets. Cette association organise un congrès par an sur le sujet.

Laboratoire du *BundesKriminAlamt* (Allemagne)

Référence : [Kunzel 94]

Description : Pratique importante de la reconnaissance de locuteur dans le cadre d'expertises judiciaires. Recherches sur l'utilisation de mesures automatiques (fréquence fondamentale, spectre à long-terme, etc.)

A cette liste, il conviendrait d'ajouter [Noda 89] (recherche effectuée par le *National Research Institute of Police Science* au Japon) que nous avons déjà eu l'occasion de citer dans la partie 2 concernant les techniques de reconnaissance automatique. Cette référence présente plusieurs résultats très intéressants concernant les problèmes posés par l'application d'une technique automatique dans le cadre de la criminalistique.

3.2.3 Remarques concernant les références citées

Les divers projets de recherche présentés dans le paragraphe précédent illustrent clairement plusieurs tendances. La première est le fait que toutes les recherches visant à mettre au point un système de reconnaissance complètement automatique dans le cadre de la criminalistique ont échouées. On peut même ajouter que les praticiens de l'expertise dans le domaine expriment plus ou moins clairement leur défiance face aux approches automatiques [Hollien 90], [Kunzel 94].

[4]Cette association regroupe en fait des scientifiques de plusieurs pays travaillant dans ce domaine, mais il semble que l'essentiel de ses activités se déroule en Grande-Bretagne

Cette situation est liée aux difficultés propres au domaine criminalistique (voir le paragraphe 3.1) qui font que les performances des techniques automatiques sont, dans ce domaine, bien inférieures à celles obtenues dans les applications usuelles [Doddington 85].

Il faut tout de même noter que les techniques automatiques qui ont été étudiées dans le cadre de la criminalistique ([Bunge 77], [Nakasone 88]) ne constituent pas les techniques ni les plus efficaces, ni les plus robustes parmi celles connues à ce jour. En effet, à part pour le choix des caractéristiques utilisées (qui ont été mises au point de manière empirique afin de limiter l'influence de variations linéaires du canal de transmission), ces techniques appartiennent clairement à la catégorie décrite au paragraphe 2.2.2.a (classification par la valeur moyenne).

L'un des projets de recherche mentionnés présente des caractéristiques originales, il s'agit du projet IDEM. En effet, d'après ses auteurs [Falcone 94], il s'agit d'un système semi-automatique *destiné à être utilisé par des opérateurs non-spécialistes du domaine*. L'intervention de l'opérateur humain est effectivement très limitée puisqu'il se contente de confirmer et d'étiqueter les segments vocaliques déterminés par une procédure de pré-détection automatique [Federico 87]. Il faut d'ailleurs souligner que le principe de fonctionnement retenu est quasiment identique à celui exposé dans [Fatokakis 93], à la différence près que l'étape la plus délicate, et la plus susceptible de produire des erreurs, (étiquetage automatique des segments vocaliques) se fait avec l'intervention d'un opérateur humain. On peut penser que ce système constitue un intermédiaire relativement robuste, du fait de la nature des caractéristiques choisies et de l'intervention d'un opérateur, et raisonnablement fiable (bien que les chiffres fournis dans [Fatokakis 93] placent cette technique nettement en dessous de la plupart de celles mentionnées au paragraphe 2 en ce qui concerne les performances).

Pour les chercheurs impliqués dans les expertises judiciaires, le point le plus important semble être l'affirmation de l'importance du rôle de l'expert phonéticien. Le recours à des mesures objectives sur le signal de parole, réalisées à l'aide d'un ordinateur, est vu comme un élément secondaire visant essentiellement à illustrer les constatations de l'expert [Hollien 90], [Kunzel 94], [Hirson 93]. De plus, ces auteurs insistent sur le fait que l'expert ne doit pas se fonder uniquement sur des mesures acoustiques réalisées sur le signal de parole mais aussi sur des aspects phonétiques et linguistiques tels que la présence éventuelle d'une pathologie vocale, l'analyse du dialecte, la détection de particularités d'élocution, etc.

L'argumentation de ces auteurs repose sur le fait qu'un expert humain est capable d'émettre un jugement beaucoup plus robuste (qu'une technique automatique) notamment vis à vis des variations du canal de transmission, des dégradations présentes sur l'enregistrement ou d'un éventuel déguisement de la voix [Hollien 90], [Kunzel 94]. A l'heure actuelle, cette manière de voir est en général partagée par les auteurs travaillant sur les techniques de reconnaissance automatique [Doddington 85]. Toutefois, *la fiabilité des jugements de reconnaissance émis par des experts phonéticiens n'a pas été évalué de manière exhaustive comme c'est le cas pour les techniques entièrement automatiques.* Une des raisons principales en est qu'étant donnée la durée importante et incompressible d'une analyse par un expert humain, il est très difficile d'envisager une évaluation systématique portant sur plusieurs milliers de tests de reconnaissance (cf. paragraphe 1.3). En l'état actuel des choses, il est donc extrêmement difficile de quantifier la fiabilité des analyses effectuées par des experts.

Un autre point gênant est le fait qu'à la lecture des publications citées, il est souvent difficile de savoir jusqu'à quel degré le résultat de l'expertise est guidé par le jugement auditif de l'expert. En effet, bien qu'un des avantages de l'expert humain soit justement la possibilité d'effectuer une analyse auditive critique et analytique, les chiffres cité au paragraphe 3.4 donnent à penser que la fiabilité d'une expertise purement auditive reste peu satisfaisante. La valeur à accorder aux jugements auditifs émis par les experts dans le domaine de la criminalistique reste d'ailleurs un

sujet controversé [Kunzel 94]. Il n'en demeure pas moins qu'une grande partie de l'expertise semble reposer sur cette analyse auditive, et fait donc intervenir le jugement de l'expert [Hirson 93], [Kunzel 94], [Hollien 90]. Dans un domaine tel que celui des expertises judiciaires, il semble indispensable de chercher à formuler ces critères auditifs sous la forme de constatations objectives pour éviter tout ce qui peut être considéré comme arbitraire ou subjectif.

3.3 Méthodes utilisées en criminalistique

Dans cette partie, nous nous proposons de faire le point sur les éléments et les mesures qui semblent être actuellement employés par les experts pratiquant la reconnaissance du locuteur dans le domaine criminalistique.

3.3.1 Eléments phonétiques et linguistiques

Les auteurs de [Hollien 90], [Kunzel 94] indiquent qu'il est très important de caractériser la façon dont le locuteur s'exprime, aussi bien sur le plan linguistique que phonétique.

Il est tout d'abord utile de déterminer la (ou les) "famille(s)" à laquelle (auxquelles) le locuteur appartient. En général, on cherche essentiellement à préciser les origines régionales (accent, expressions) et sociales du locuteur (type d'élocution, vocabulaire) [Hollien 90], [Kunzel 94]. Il faut cependant souligner que selon les pays et les langues, ce type de constatations peut être plus ou moins discriminante. Ainsi, s'il est logique que cette partie occupe une place importante dans [Hollien 90] compte tenu du grand nombre de variantes présentées par l'anglais américain, on peut se demander ce qu'il en est pour le français. Enfin, les publications citées ne fournissent pas de résultats permettant de préciser la fiabilité de ce type de constatations, notamment en ce qui concerne la manière dont ces habitudes sont susceptibles de se modifier dans le temps.

Un autre point intéressant consiste à détecter la présence éventuelle d'une pathologie vocale, ou tout du moins d'une forte particularité. Même si cet élément n'est évidemment pas toujours présent, il est très important de l'exploiter lorsqu'il existe [Hirson 93], [Hollien 90], [Kunzel 94]. De plus, la caractérisation d'une éventuelle particularité vocale présente aussi l'intérêt d'être un élément objectif qui peut, en général, être illustré par des mesures effectuées à partir du signal de parole [Braun 94], [Hirson 93].

Il est clair qu'un type de "pathologie" qu'il est particulièrement important de détecter est la présence d'une modification volontaire de la voix. Il est en effet indispensable de détecter et de préciser la nature d'un éventuel déguisement de la voix pour éviter d'utiliser des techniques totalement inappropriées. La plupart des publications sur le sujet s'accordent pour trouver que c'est plus particulièrement sur ce point que la présence d'un expert phonéticien est indispensable [Hirson 93], [Hollien 90], [Kunzel 94]. A ce niveau, l'expert a pour rôle, à la fois, de mettre en évidence une éventuelle particularité vocale qui peut être très discriminante et difficile à modifier, ainsi que de détecter la présence d'un déguisement de la voix qui pourrait compromettre la validité des méthodes de reconnaissance utilisées. Là encore, les auteurs des publications citées fournissent peu de données permettant d'évaluer l'efficacité de ce type de constatations ou leur pertinence statistique (c'est à dire le pourcentage d'individus présentant effectivement des particularités phonétiques exploitables).

3.3.2 Mesures sur le signal de parole

3.3.2.a Caractéristiques "fréquentielles"

Parmi les mesures utilisées en criminalistique, les plus courantes et les plus largement acceptées sont les mesures de caractéristiques fréquentielles du signal de parole : *fréquence fondamentale de voisement et fréquences des formants dans les sections vocaliques.* Ces caractéristiques sont réputées efficaces car elles correspondent à des grandeurs relativement stables et très robustes notamment vis à vis de la présence de bruit de fond et des modifications linéaires dues au canal de transmission [Falcone 94], [Hollien 90], [Kunzel 94]. Les résultats obtenus dans [Endres 71] semblent toutefois indiquer que ces caractéristiques présentent des variations à long-terme (plusieurs années) non négligeables et, relativement, systématiques.

Pour la fréquence fondamentale, seule la valeur moyenne est en général utilisée [Doherty 76], [Hirson 94], [Kunzel 94]. [Doherty 78] présente une mesure simple destinée à caractériser la dispersion de la fréquence fondamentale. Toutefois, il ne semble pas qu'il soit possible d'exploiter de manière significative l'ensemble des variations temporelles de la fréquence fondamentale (comme c'est le cas par exemple dans [Sambur 75]). Ceci est certainement lié au fait que le texte prononcé n'est en général pas fixé.

Il semble utile de préciser que la mesure de la fréquence fondamentale moyenne n'est en aucun cas très discriminante (point qui n'apparaît pas de manière très claire dans les articles citées). En effet, compte tenu de la variabilité de la fréquence de voisement, notamment en fonction de la situation psychologique du locuteur [Hollien 90], [Hirson 94], il est tout à fait illusoire de mesurer sa valeur à plus de quelques pour-cent près. Cette incertitude fait qu'une proportion importante de locuteurs ne pourront pas être distingués grâce à cette mesure. La fonction de répartition représentée dans [Kunzel 94] montre par exemple que près de 35% des locuteurs masculins allemands ont une fréquence fondamentale moyenne située entre 110 Hz et 120 Hz. La mesure de fréquence fondamentale fournit donc surtout des résultats exploitables lorsqu'un des locuteurs à comparer s'éloigne fortement des caractéristiques moyennes.

Pour la mesure des fréquences de formants, la situation est assez semblable à la différence près que la variabilité est encore bien plus importante compte tenu de l'influence du contexte phonétique sur la réalisation des voyelles ainsi que de la difficulté intrinsèque de la mesure, notamment pour les voix possédant un fondamental élevé (femmes, enfants). Pour limiter au maximum cette variabilité, il convient donc de considérer chaque voyelle dans des contextes comparables, de ne travailler que sur les voyelles suffisamment distinctes (en ce qui concerne les fréquences de formants) et de disposer d'un nombre suffisant de réalisations de chaque voyelle [Hollien 90]. Ici encore, on peut penser que le résultat est d'autant plus exploitable que les fréquences mesurées s'éloignent significativement de la configuration moyenne. Toutefois, le système décrit dans [Falcone 94] repose uniquement sur une classification effectuée dans l'espace vectoriel défini par les deux (ou trois) premières fréquences de formants [Federico 87].

3.3.2.b Caractéristiques "spectrales"

Le terme "spectral" est utilisé ici pour désigner les cas où l'ensemble du spectre du signal de parole est pris en compte (le terme "fréquentiel" étant réservé aux cas où l'on ne considère que la valeur de fréquences particulières). On distingue deux types de caractéristiques spectrales : des mesures globales de la répartition spectrale du signal, et des mesures spectrales ponctuelles concernant des événements phonétiques particuliers.

Dans le domaine de la criminalistique, c'est souvent la première démarche qui est utilisée. Malheureusement, la seule mesure globale qui a été étudié dans ce cadre est le spectre moyen

à long terme [Doherty 76], [Doherty 78], [Hollien 90], [Kunzel 94]. Nous avons vu au paragraphe 2.2.2.a que cette technique ne peut donner des résultats significatifs que lorsque les conditions d'enregistrements sont maîtrisées. Il est à ce titre assez surprenant de constater que cette technique semble néanmoins être utilisée dans le cadre d'expertises judiciaires [Kunzel 94].

Nous avons vu (paragraphe 2) que dans le cadre d'une technique automatique la seconde démarche n'était pas envisageable compte tenu de la difficulté représentée par la localisation automatique d'événements phonétiques dans le signal de parole. En criminalistique, cette approche peut être praticable puisque l'étiquetage du signal est réalisé par un expert humain. Parmi, les phonèmes dont les spectres sont plus ou moins caractéristiques du locuteur, on trouve mentionnées surtout les consonnes nasales [m] et [n], qui sont réputées difficilement modifiables par le locuteur du fait de leur mécanisme de production [Sambur 75], [Hollien 90]. L'utilisation de la consonne [s] a aussi été étudiée [Sambur 75] : son spectre semble être peu modifié par certains types de déguisements vocaux [Hirson 93]. Dans tous les cas, on manque de données permettant de savoir quelles sont exactement les caractéristiques spectrales significatives (forme complète du spectre ou, par exemple, formants dans le cas des nasales) et quelle fiabilité de reconnaissance il faut attendre de type de mesure.

3.3.2.c Caractéristiques temporelles

Un dernier type de mesures concerne ce qu'ont peut appeler des caractéristiques temporelles. En général, il s'agit de durées caractéristiques mesurées à partir de la courbe représentant l'évolution temporelle du niveau sonore du signal. Selon les cas, on s'intéresse directement aux variations temporelles de l'intensité ou bien à des données plus synthétiques visant à caractériser le rythme d'élocution [Hollien 90], [Johnson 84] [Kunzel 94]. Dans tous les cas, bien que les auteurs s'accordent sur l'importance de ce type de caractéristiques [Hollien 90], [Kunzel 94], la nature exacte des mesures à effectuer n'est pas décrite. De plus, il est difficile de savoir dans quel mesure ces considérations peuvent être appliquées aux cas où le texte n'est pas fixé.

3.4 Reconnaissance par les auditeurs humains

Ce paragraphe traite d'un aspect lié à la reconnaissance du locuteur qui joue un rôle très important en criminalistique : l'étude des mécanismes qui permettent à un auditeur humain de reconnaître un locuteur. Ce sujet, qui est loin d'être complètement maîtrisé, fait l'objet de nombreuses études de type psychoacoustique (voir, par exemple, [Hollien 90], [Itoh 88], [LaRiviere 75]). Ce type d'études est en général mené par des équipes distinctes de celles que nous avons mentionnées au paragraphe 2 car les recherches actuelles sur les techniques de reconnaissance du locuteur ne font que très peu référence au fonctionnement de la perception humaine.

Par contre, la reconnaissance du locuteur par les auditeurs humains constitue un sujet très important dans le cadre de l'application en criminalistique. En effet, dans le cadre d'une expertise judiciaire, il est fréquent d'avoir affaire à des "témoignages auditifs" [Kunzel 94], c'est à dire à des personnes qui soutiennent avoir identifié, par exemple, l'auteur d'un coup de fil. Il est très utile de pouvoir préciser quelle doit être la confiance à accorder à ce type de témoignage. De plus, il faut être en mesure, lorsque c'est possible, de mettre en œuvre des procédures visant à vérifier la fiabilité de tels témoignages auditifs.

3.4.1 Performances de la reconnaissance auditive

3.4.1.a Auditeur "naïf"

D'une manière générale, les performances de reconnaissance obtenues par des auditeurs humains peuvent être qualifiées de moyennes : en situation idéale (chaque locuteur prononce au moins une phrase avec une élocution "naturelle", bonnes conditions d'enregistrement, etc.), le *taux d'erreur pour une tâche de vérification auditive* est de l'ordre de 8% [Reich 79] à 12% [Homayounpour 93]. En *identification*, compte tenu de la difficulté intrinsèque de la tâche (voir paragraphe 1.1.1), les performances se dégradent très rapidement : [Hollien 82] rapporte des taux d'erreurs de l'ordre de 60% pour une expérience d'identification auditive avec un groupe de 10 personnes. Ce chiffre n'est pas surprenant compte tenu du fait que la tâche d'identification fait fortement appel à la mémoire auditive. Ainsi pour identifier un locuteur parmi 10, il faut au préalable mémoriser les caractéristiques vocales de 10 locuteurs différents. Habituellement, cette tâche "d'apprentissage" des caractéristiques d'un locuteur n'est effectuée (inconsciemment) que pour les personnes familières que l'on a fréquemment l'occasion d'entendre (voir paragraphe suivant).

Il est fort probable que les résultats chiffrés mentionnés ci-dessus constituent en fait une évaluation plutôt optimiste des performances de la reconnaissance auditive car les enregistrements utilisés dans [Hollien 82] et [Reich 79] sont tous contemporains (c'est à dire qu'on ne prend pas en compte les conséquences de la variabilité intra-locuteur à long-terme).

La reconnaissance auditive est assez robuste vis à vis des conditions d'enregistrement (notamment en ce qui concerne la présence de bruit de fond, ou le filtrage) [Doddington 85]. Par contre, les performances de la reconnaissance auditive sont fortement affectées par les variations propres au locuteur : intervalle long entre les enregistrements à comparer [Hollien 90], état de stress [Hollien 82] ou de manière plus générale d'émotivité [Homayounpour 93], ou élocution particulière (voix chuchotée [Hollien 90], élocution très lente [Reich 79]).

La dégradation la plus importante des performances est constatée en présence d'un déguisement volontaire de la voix [Hollien 90], [Hollien 82], [Reich 79]. En présence d'un déguisement de la voix, la reconnaissance auditive fournit des performances quasiment inexploitables (proches d'un choix au hasard) [Hollien 82], [Reich 79]. Par contre, il semble que la présence d'un déguisement de la voix puisse être détectée assez facilement à l'écoute[5] [Reich 81]. Ceci se traduit d'ailleurs par le fait que face à un déguisement de la voix, la proportion des rejets erronés (par rapport au total des erreurs commises) augmente significativement : les auditeurs hésitent à déclarer que deux voix proviennent du même locuteur lorsqu'ils détectent une "anomalie" [Reich 79].

Les performances de la reconnaissance auditive diminuent significativement lorsque le texte prononcé est très court [Bricker 66], [Hollien 90]. En fait, il semble surtout que ce soit le nombre de phonèmes qui soit déterminant : en dessous d'une dizaine de phonèmes (mots isolés, courtes expression du type "Allo, bonjour", etc.), les performances se dégradent très vite [Bricker 66]. En pratique, la dégradation est certainement plus importante que ce que laissent supposer les résultats de [Bricker 66] compte tenu du fait que cette étude utilise uniquement des auditeurs familiers de locuteurs.

Enfin, la "mémoire" auditive ne semble pas être très importante : les jugements de reconnaissance deviennent quasiment inexploitables dès que l'intervalle entre la première présentation de la voix inconnue et le test de reconnaissance excède plusieurs mois [Doddington 85], [Hollien 90]. En pratique, compte tenu des délais caractéristiques de l'application en criminalistique, les témoignages auditifs sont donc en général peu exploitables (saufs lorsqu'ils émanent de personnes

[5] Il faut toutefois considérer avec précaution le résultat de cette étude car les instructions données aux locuteurs étaient de déguiser leur voix de façon à la rendre la plus méconnaissable possible. Dans une situation réelle, il est fort probable qu'une personne qui déguise sa voix veille également à préserver un caractère relativement naturel.

familières avec locuteur présumé).

3.4.1.b Personne familière

Le facteur qui influe le plus sur la fiabilité de la reconnaissance auditive est la familiarité de l'auditeur avec les locuteurs à reconnaître [Hollien 90]. Il semble en effet qu'il nous soit possible de reconnaître beaucoup plus sûrement les locuteurs lorsque nous avons fréquemment eu l'occasion d'entendre leur voix [Doddington 85]. Ainsi, dans l'expérience décrite dans [Hollien 82] (identification auditive sur un groupe de dix locuteurs), alors que le taux d'erreur est de 60% pour des auditeurs ne connaissant pas les locuteurs, il tombe à 2% pour des auditeurs familiers des locuteurs. Cette même expérience montre aussi que les personnes familières fournissent des décision d'identification beaucoup plus robustes vis à vis de variations inconscientes (état de stress) ou intentionnelles (déguisement de la voix) des caractéristiques des locuteurs.

3.4.1.c Spécialiste

Une autre catégorie d'auditeur un peu particulière est le spécialiste (sous entendu dans le domaine de l'expertise vocale). Il est important de noter que dans cette partie nous ne considérons que la reconnaissance auditive (par l'écoute), à l'exclusion de toute autre opération. Cette distinction est importante car nous avons vu qu'un expert ne se contente en général pas d'une simple écoute.

Une première remarque est que la vigueur même de la controverse sur ce sujet [Hollien 90], [Kunzel 94] indique en elle même que la supériorité du spécialiste par rapport à l'auditeur "naïf" n'est certainement pas très significative. L'expérience décrite dans [Reich 79] tend à montrer que les performances de vérification auditives obtenues par les spécialistes (en l'occurrence, des universitaires spécialisés en phonétique) sont comparables à celles d'auditeurs naïfs. Toutefois, la supériorité des spécialistes se manifeste dans cette étude par le fait que leurs décisions sont beaucoup plus robustes vis à vis de tentatives de déguisement, ou de modification des caractéristiques de la voix. Cependant, les taux d'erreurs en vérification auditive obtenus avec des voix déguisées restent élevés, même lorsque l'auditeur est un spécialiste de phonétique (20% dans [Reich 79], 35% dans [Hirson 93]).

3.4.2 Procédures de vérification de la reconnaissance auditive

La discussion précédente a permis de mettre en évidence le fait que la reconnaissance auditive n'est pas extrêmement fiable. Cet état de chose a deux conséquences importantes dans le cadre de la criminalistique : premièrement, même un expert ne saurait se contenter d'une simple écoute, en second lieu, les témoignages reposant sur la reconnaissance auditive ne sont pas infaillibles, même s'ils émanent de familiers. Il est donc indispensable de mettre en œuvre des procédures visant à évaluer la solidité de tout témoignage auditif.

[Kunzel 94] mentionne plusieurs procédures de choix parmi un groupe de voix (incluant la voix du locuteur présumé) pour vérifier la fiabilité d'un témoignage auditif. Ce type de procédure est analogue à ce qui se fait en criminalistique pour la reconnaissance visuelle (procédure dite de *line-up*). Un point extrêmement délicat est le fait qu'il est nécessaire de constituer un ensemble d'échantillons de voix "comparables" avec la voix du locuteur présumé. A ce niveau, [Kunzel 94] précise qu'il est indispensable de faire appel à un expert phonéticien pour sélectionner des locuteurs possédant des caractéristiques compatibles avec celles du locuteur présumé. Notons de plus, que si la voix du locuteur présumé à été enregistrée dans des conditions particulières (par exemple par téléphone), il est aussi nécessaire de simuler ces conditions sur les enregistrements des autres locuteurs.

Bibliographie

[Assaleh 94] K. T. Assaleh and R. J. Mammone. New LP-derived features for speaker identification. *IEEE Trans. Speech and Audio Processing*, vol. 2, no. 4, pp. 630–638, 1994.

[Atal 76] B. S. Atal. Automatic recognition of speakers from their voices. *Proc. IEEE*, vol. 64, no. 4, pp. 460–475, 1976.

[Bimbot 93] F. Bimbot and L. Mathan. Text-free speaker recognition using an arithmetic-harmonic sphericity measure. In *Proc. EUROSPEECH*, 1993.

[Bimbot 94a] F. Bimbot, G. Chollet, and A. Paoloni. Assessment methodology for speaker identification and verification systems: An overview of SAM-A Esprit project 6819 - Task 2500. In *Proc. ESCA Workshop on Automatic Speaker Recognition Identification and Verification ([ESCA 94])*, pp. 75–82, 1994.

[Bimbot 94b] F. Bimbot and L. Mathan. Second-order statistical measures for text-independent speaker identification. In *Proc. ESCA Workshop on Automatic Speaker Recognition Identification and Verification ([ESCA 94])*, pp. 51–54, 1994.

[Botte 88] M. C. Botte, G. Canevet, and L. Demany. *Psychoacoustique et perception auditive*. INSERM, Paris, 1988.

[Boves 94] L. Boves, T. Bogaart, and L. Bos. Design and recording of large data bases for use in speaker verification and identification. In *Proc. ESCA Workshop on Automatic Speaker Recognition Identification and Verification ([ESCA 94])*, pp. 43–46, 1994.

[Braun 94] A. Braun. The effect of cigarette smoking on vocal parameters. In *Proc. ESCA Workshop on Automatic Speaker Recognition Identification and Verification ([ESCA 94])*, pp. 161–164, 1994.

[Bricker 66] P. D. Bricker and S. Pruzansky. Effects of stimulus content and duration on talker identification. *J. Acoust. Soc. Am.*, vol. 40, no. 6, pp. 1441–1449, 1966.

[Bricker 71] P. D. Bricker, R. Gnanadesikan, M. V. Mathews, S. Pruzansky, P. A. Tukey, K. W. Wachter, and J. L. Warner. Statistical techniques for talker identification. *Bell System Technical Journal*, vol. 50, no. 4, pp. 1427–1454, 1971.

[Bunge 77] E. Bunge. Automatic speaker recognition system AUROS for security systems and forensic voice identification. In *Proc. 1977 Internat. Conf. on Crime Countermeasures-Sci. & Eng.*, pp. 1–7, 1977.

[Buquet 91] A. Buquet. *L'expertise des écritures.* Presses du CNRS, 1991.

[Calliope 89] Calliope. *La parole et son traitement automatique.* Collection technique et scientifique des télécommunications. Masson, 1989.

[Carey 84] M. B. Carey, H. T. Chen, A. Descloux, J. F. Ingle, and K. I. Park. 1982/1983 End office connection study: analog voice and voiceband data transmission performance characterization of the public switched network. *AT&T Bell Laboratories Technical Journal*, vol. 69, no. 9, pp. 2059–2119, 1984.

[Chen 93] M-S. Chen, P-H. Lin, and H-C. Wang. Speaker identification based on a matrix quantization method. *IEEE Trans. Signal Processing*, vol. 41, no. 1, pp. 398–403, 1993.

[Cheung 78] R. S. Cheung and B. A. Eisenstein. Feature selection via dynamic programming for text-independent speaker identification. *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 26, no. 5, pp. 397–403, 1978.

[Das 71] S. K. Das and W. S. Mohn. A scheme for speech processing in automatic speaker verification. *IEEE Trans. Audio Electroacoust.*, vol. 19, pp. 32–43, March 1971.

[Davis 80] S. B. Davis and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 28, pp. 357–366, August 1980.

[Dempster 77] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. Royal Statist. Soc. Ser. B (methodological)*, vol. 39, no. 1, pp. 1–22 et 22–38 (discussion), 1977.

[Doddington 85] G. Doddington. Speaker recognition - Identifying people by their voices. *Proc. IEEE*, vol. 73, no. 11, pp. 1651–1664, 1985.

[Doherty 76] E. T. Doherty. An evaluation of selected acoustic parameters for use in speaker identification. *J. Phonetics*, vol. 4, pp. 321–326, 1976.

[Doherty 78] E. T. Doherty and H. Hollien. Multiple-factor speaker identification of normal and distorted speech. *J. Phonetics*, vol. 6, pp. 1–8, 1978.

[Duda 73] R. O. Duda and P. E. Hart. *Pattern classification and scene analysis.* John Wiley & Sons, Inc., New York, 1973.

[Endres 71] W. Endres, W. Bambach, and G. Flösser. Voice spectrograms as a function of age, voice disguise, and voice imitation. *J. Acoust. Soc. Am.*, vol. 49, no. 6 (Part 2), pp. 1842–1848, 1971.

[ESCA 94] European Speech Communication Association (ESCA). *ESCA Workshop on Automatic Speaker Recognition Identification and Verification. Martigny, Switzerland, April 5-7*, 1994.

[Falcone 94] M. Falcone and N. De Sario. A PC speaker identification system for forensic use: IDEM. In *Proc. ESCA Workshop on Automatic Speaker Recognition Identification and Verification ([ESCA 94])*, pp. 169–172, 1994.

[Farell 94] K. R. Farell, R. J. Mammone, and K. T. Assaleh. Speaker recognition using neural networks and conventional classifiers. *IEEE Trans. Speech and Audio Processing (special issue on neural networks for speech)*, vol. 2, no. 1 (part II), pp. 194–205, January 1994.

[Fatokakis 93] N. Fatokakis, A. Tsopanoglou, and G. Kokkinakis. A text-independent speaker recognition system based on vowel spotting. *Speech Commun.*, vol. 12, no. 1, pp. 57–68, 1993.

[Federico 87] A. Federico, G. Ibba, and A. Paoloni. A new automated method for reliable speaker identification and verification over telephone channels. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 1457–1460, 1987.

[Federico 89] A. Federico, G. Ibba, A. Paoloni, N. De Sario, and B. Saverione. Comparison between automatic methods and human listeners in speaker recognition tasks. In *Proc. EUROSPEECH*, pp. 279–282, 1989.

[Federico 93] A. Federico and A. Paoloni. Bayesian decision in the speaker recognition by acoustic parametrization of voice samples over telephone lines. In *Proc. EUROSPEECH*, pp. 2307–2310, 1993.

[Furui 81a] S. Furui. Cepstral analysis technique for automatic speaker verification. *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 29, no. 2, pp. 254–272, 1981.

[Furui 81b] S. Furui. Comparison of speaker recognition methods using statistical features and dynamic features. *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 29, no. 3, pp. 342–350, 1981.

[Furui 86] S. Furui. Research on indivituality features in speech waves and automatic speaker recognition techniques. *Speech Commun.*, vol. 5, no. 2, pp. 183–197, 1986.

[Furui 94] S. Furui. An overview of speaker recognition technology. In *Proc. ESCA Workshop on Automatic Speaker Recognition Identification and Verification ([ESCA 94])*, pp. 1–9, 1994.

[GCP 90] GCP. Motion adoptée par le bureau du Groupe Communication Parlée (GCP) de la Société Française d'Acoustique (SFA), 1990.

[Gish 85] H. Gish, K. Karnofsky, M. Krasner, S. Roucos, R. Schwartz, and J. Wolf. Investigation of text-independent speaker identification over telephone channels. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 379–382, 1985.

[Gish 86] H. Gish, M. Krasner, W. Russel, and J. Wolf. Methods and experiments for text-independent speaker recognition over telephone channels. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 865–868, 1986.

[Gish 90] H. Gish. Robust discrimination in automatic speaker identification. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 289–292, 1990.

[Gish 91] H. Gish, M-H. Siu, and R. Rohlicek. Segregation of speakers for speech recognition and speaker identification. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 873–876, 1991.

[Gish 94] H. Gish and M. Schmidt. Text-idependent speaker identification speaker. *IEEE Signal Processing Magazine*, vol. 11, no. 4, pp. 18–32, October 1994.

[Godfrey 94] J. Godfrey, D. Graff, and A. Martin. Public databases for speaker recognition and verification. In *Proc. ESCA Workshop on Automatic Speaker Recognition Identification and Verification ([ESCA 94])*, pp. 39–42, 1994.

[Goldstein 75] U. G. Goldstein. Speaker-identifying features based on formant tracks. *J. Acoust. Soc. Am.*, vol. 59, no. 1, pp. 176–182, 1975.

[Hirson 93] A. Hirson and M. Duckworth. Glottal fry and voice disguise: a case study in forensic phonetics. *J. Biomed. Eng.*, vol. 15, no. 3, pp. 193–200, 1993.

[Hirson 94] A. Hirson, P. French, and D. Howard. Speech fundamental frequency over the telephone and face-to-face: some implications for forensic phonetics. In *Studies in General and English Phonetics: Essays in Honour of Professor J D O'Connor*. Routledge, London, 1994.

[Hollien 82] H. Hollien, W. Majewski, and E. T. Doherty. Perceptual identification of voices under normal, stress and disguise speaking conditions. *J. Phonetics*, vol. 10, pp. 139–148, 1982.

[Hollien 90] H. Hollien. *The Acoustics of Crime - The New Science of Forensic Phonetics*. Plenum Press, 1990.

[Homayounpour 93] M. M. Homayounpour, J. P. Goldman, G. Chollet, and J. Vaissière. Performance comparison of machine and human speaker verification. In *Proc. EUROSPEECH*, pp. 2295–2298, 1993.

[Howard 93] D. M. Howard, A. Hirson, J. P. French, and J. E. Szymanski. A survey of fundamental frequency estimation techniques used in forensic phonetics. *Proceedings of the Institute of Acoustics*, vol. 15 (Part 7), pp. 207–215, 1993.

[Itoh 88] K. Itoh and S. Saito. Effects of acoustical feature parameters on perceptual speaker identity. *Review of the Electrical Communications Laboratories*, vol. 36, no. 1, pp. 135–141, 1988.

[Johnson 84] C. C. Johnson, H. Hollien, and J. W. Hicks. Speaker identification utilizing selected temporal speech features. *J. Phonetics*, vol. 12, pp. 319–326, 1984.

[Juang 87] B-H. Juang, L. R. Rabiner, and J. G. Wilpon. On the use of bandpass liftering in speech recognition. *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 35, no. 7, pp. 947–954, 1987.

[Kao 93] Y-H. Kao, J. S. Barras, and P. K. Rajasekaran. Robustness study of free-text speaker identification and verification. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. II–379–II–382, 1993.

[Krasner 84] M. Krasner, J. Wolf, K. Karnofsky, R. Schwartz, S. Roucos, and H. Gish. Investigation of text-independent speaker identification techniques under conditions of variable data. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 18B.5.1–18B.5.4, 1984.

[Kunzel 94] H. J. Künzel. Current approaches to forensic speaker recognition. In *Proc. ESCA Workshop on Automatic Speaker Recognition Identification and Verification ([ESCA 94])*, pp. 135–141, 1994.

[LaRiviere 75] C. LaRiviere. Contributions of fundamental frequency and formant frequencies to speaker identification. *Phonetica*, vol. 31, pp. 185–197, 1975.

[Makhoul 85] J. Makhoul, S. Roucos, and H. Gish. Vector quantization in speech coding. *Proc. IEEE*, vol. 73, no. 11, pp. 1551–1588, November 1985.

[Markel 77] J. D. Markel, B. T. Oshika, and H. Gray, Jr. Long-term feature averaging for speaker recognition. *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 25, no. 4, pp. 330–337, 1977.

[Markel 79] J. D. Markel and S. B. Davis. Long-term feature averaging for speaker recognition. *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 27, no. 1, pp. 74–82, 1979.

[Matsui 91] T. Matsui and S. Furui. A text-independent speaker recognition method robust against utterance variations. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 377–380, 1991.

[Matsui 92] T. Matsui and S. Furui. Comparison of text-independent speaker recognition methods using VQ-distortion and discrete/continuous HMMs. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. II–157–II–160, 1992.

[Matsui 93] T. Matsui and S. Furui. Concatenated phoneme models for text-variable speaker recognition. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. II–391–II–394, 1993.

[Matsui 94a] T. Matsui and S. Furui. Comparison of text-independent speaker recognition methods using VQ-distortion and discrete/continuous HMMs. *IEEE Trans. Speech and Audio Processing*, vol. 2, no. 3, pp. 456–459, July 1994.

[Matsui 94b] T. Matsui and S. Furui. Similarity normalization method for speaker verification based on a posteriori probability. In *Proc. ESCA Workshop on Automatic Speaker Recognition Identification and Verification ([ESCA 94])*, pp. 59–62, 1994.

[Mohn 71] W. S. Mohn. Two statistical feature evaluation techniques applied to speaker identification. *IEEE Trans. Computers*, vol. 20, no. 9, pp. 979–987, 1971.

[Naik 89] J. M. Naik, L. P. Netsch, and G. R. Doddington. Speaker verification over long distance telephone lines. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 524–527, 1989.

[Naik 94] J. Naik. Speaker verification over the telephone network: Databases, algorithms and performance assessment. In *Proc. ESCA Workshop on Automatic Speaker Recognition Identification and Verification ([ESCA 94])*, pp. 31–38, 1994.

[Nakasone 88] H. Nakasone and C. Melvin. Computer assisted voice identification system (C.A.V.I.S.). In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 587–590, 1988.

[Noda 89] H. Noda. On the use of the information on individual speaker's position in the parameter space for speaker recognition. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 516–519, 1989.

[NS 93] NS. Ecoutes chinoises. *le Quotidient*, mercredi 12 décembre 1993.

[Oglesby 94] J. Oglesby. What's in a number ?: Moving beyond the equal error rate. In *Proc. ESCA Workshop on Automatic Speaker Recognition Identification and Verification ([ESCA 94])*, pp. 87–90, 1994.

[Openshaw 93] J. P. Openshaw, Z. P. Sun, and J. S. Mason. A comparison of composite features under degraded speech in speaker recognition. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. II–371–II–374, 1993.

[OShaughnessy 86] D. O'Shaughnessy. Speaker recognition. *IEEE ASSP Magazine*, pp. 4–17, October 1986.

[Paul Jr 75] J. E. Paul, Jr., A. S. Rabinowitz, J. P. Riganati, and J. M. Richardson. Development of analytical methods for a semi-automatic speaker identification system. In *Proc. 1975 Carnahan. Conf. on Crime Countermeasures*, pp. 52–64, 1975.

[Quatieri 94] T. F. Quatieri, C. R. Jankowski, Jr., and D. A. Reynols. Energy onset times for speaker identification. *IEEE Signal Processing Letters*, vol. 1, no. 11, November 1994.

[Rabiner 78] L. R. Rabiner and R. W. Schafer. *Digital processing of speech signals.* P-H signal processing series. Prentice-Hall, 1978.

[Rabiner 89] L. R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE*, vol. 77, no. 2, pp. 257–285, February 1989.

[Rabiner 93] L. R. Rabiner and B-H. Juang. *Fundamentals of speech recognition.* PTR Prentice-Hall, 1993.

[Redner 84] R. A. Redner and H. F. Walker. Mixture densities, maximum likelihood and the EM algorithm. *SIAM Review*, vol. 26, no. 2, pp. 195–239, April 1984.

[Reich 79] A. R. Reich and J. E. Duke. Effects of selected vocal disguises upon speaker identification by listening. *J. Acoust. Soc. Am.*, vol. 66, no. 4, pp. 1023–1028, 1979.

[Reich 81] A. R. Reich. Detecting the presence of vocal disguise in the male voice. *J. Acoust. Soc. Am.*, vol. 69, no. 5, pp. 1458–1461, 1981.

[Reynolds 92] D. A. Reynolds. *A gaussian mixture modeling approach to text-independent speaker identification.* PhD thesis, Georgia Institute of Technology, August 1992.

[Reynolds 94a] D. A. Reynolds. Experimental evaluation of features for robust speaker identification. *IEEE Trans. Speech and Audio Processing*, vol. 2, no. 4, pp. 639–643, 1994.

[Reynolds 94b] D. A. Reynolds. Speaker identification and verification using gaussian mixture speaker models. In *Proc. ESCA Workshop on Automatic Speaker Recognition Identification and Verification ([ESCA 94])*, pp. 27–30, 1994.

[Rose 90] R. C. Rose and D. A. Reynolds. Text independent speaker identification using automatic acoustic segmentation. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 293–296, 1990.

[Rose 91] R. C. Rose, J. Fitzmaurice, E. M. Hofstetter, and D. A. Reynolds. Robust speaker identification in noisy environments using noise adaptive speaker models. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 401–404, 1991.

[Rose 94] R. C. Rose, E. M. Hofstetter, and D. A. Reynolds. Integrated models of signal and background with application to speaker identification in noise. *IEEE Trans. Speech and Audio Processing*, vol. 2, no. 2, pp. 245–257, April 1994.

[Rosenberg 75] A. E. Rosenberg and M. R. Sambur. New techniques for automatic speaker verification. *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 23, no. 2, pp. 169–176, 1975.

[Rosenberg 76] A. E. Rosenberg. Automatic speaker verification: a review. *Proc. IEEE*, vol. 64, no. 4, pp. 475–487, 1976.

[Rosenberg 86] A. E. Rosenberg and F. K. Soong. Evaluation of a vector quantization talker recognition system in text independent and text dependent modes. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 873–876, 1986.

[Rosenberg 90] A. E. Rosenberg, C. H. Lee, and F. K. Soong. Sub-word unit talker verification using hidden Markov models. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 269–272, 1990.

[Rosenberg 91] A. E. Rosenberg and F. K. Soong. Recent research in automatic speaker recognition. S. Furui and M. Sondhi, editors, In *Advances in Speech Signal Processing*, chapter 22, pp. 701–738. Marcel Dekker, 1991.

[Sambur 75] M. R. Sambur. Selection of acoustic features for speaker identification. *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 23, no. 2, pp. 176–182, 1975.

[Sankar 94] A. Sankar and C.-H. Lee. Stochastic matching for robust speech recognition. *IEEE Signal Processing Letters*, vol. 1, no. 8, August 1994.

[Savic 90] M. Savic and S. K. Gupta. Variable parameter speaker verification system based on hidden Markov modeling. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 281–284, 1990.

[Shridhar 82] M. Shridhar and N. Mohankrishnan. Text-independent speaker recognition: a review and some new results. *Speech Commun.*, vol. 1, no. 3-4, pp. 257–267, 1982.

[Soong 85] F. K. Soong, A. E. Rosenberg, L. R. Rabiner, and B. H. Huang. A vector quantization approach to speaker recognition. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 387–390, 1985.

[Soong 87] F. K. Soong, A. E. Rosenberg, and B-H. Juang. A vector quantization approach to speaker recognition. *AT&T Technical Journal*, vol. 66, no. 2, pp. 14–26, 1987.

[Soong 88] F. K. Soong and A. E. Rosenberg. On the use of instantaneous and transitional spectral information in speaker recognition. *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 36, no. 6, pp. 871–879, 1988.

[Tishby 91] N. Z. Tishby. On the application of mixture AR hidden Markov models to text independent speaker recognition. *IEEE Trans. Signal Processing*, vol. 39, no. 3, pp. 563–570, 1991.

[Tseng 92] B. L. Tseng, F. K. Soong, and A. E. Rosenberg. Continuous probabilistic acoustic map for speaker recognition. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. II–161–II–164, 1992.

[Ventsel 73] H. Ventsel. *Théorie des probabilités (traduit du russe)*. Editions Mir, 1973.

[Vincent 93] R. Vincent. La bataille des voix. *France-Soir*, jeudi 18 novembre 1993.

[Wagner 94] M. Wagner, F. Chen, I. Macleod, B. Millar, S. ran, A. Tridgell, and X. Zhu. Analysis of type-II errors for VQ-distorsion based speaker verification. In *Proc. ESCA Workshop on Automatic Speaker Recognition Identification and Verification ([ESCA 94])*, pp. 83–86, 1994.

[Webb 93] J.J. Webb and E. L. Rissanen. Speaker identification experiments using HMMs. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. II–387–II–390, 1993.

[Wolf 72] J. J. Wolf. Efficient acoustic parameters for speaker recognition. *J. Acoust. Soc. Am.*, vol. 51, no. 6 (Part 2), pp. 2044–2056, 1972.

[Xu 89a] L. Xu and J. S. Mason. Instantaneous and transitional perceptually-based features in speaker identification. In *Proc. EUROSPEECH*, pp. 271–274, 1989.

[Xu 89b] L. Xu, J. Oglesby, and J. S. Mason. The optimization of perceptually-based features for speaker identification. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 520–523, 1989.

[Yu 93] G. Yu and H. Gish. Identification of speakers engaged in dialog. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. II–383–II–386, 1993.