APPROCHES STATISTIQUES ET FILTRAGE VECTORIEL DE TRAJECTOIRES SPECTRALES POUR L'IDENTIFICATION DU LOCUTEUR INDÉPENDANTE DU TEXTE

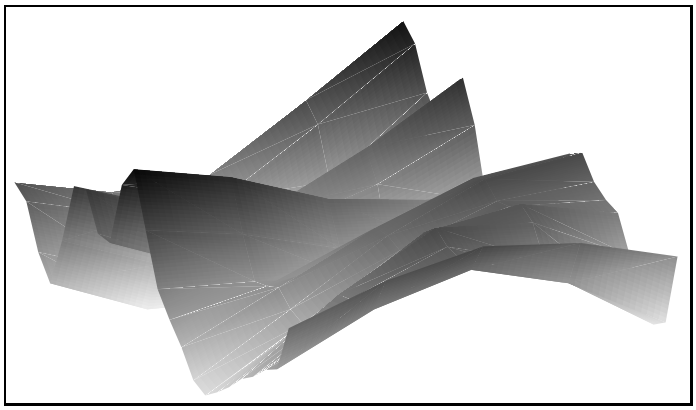
Ivan Magrin-Chagnolleau

APPROCHES STATISTIQUES

ET

FILTRAGE VECTORIEL DE TRAJECTOIRES SPECTRALES

POUR L'IDENTIFICATION DU LOCUTEUR INDÉPENDANTE DU TEXTE



Ivan Magrin-Chagnolleau

Thèse de l'E.N.S.T. sous la direction de Frédéric Bimbot

Soutenue le 17 Janvier 1997 devant un jury composé de :

Jean-Paul Haton	Président
Lori Lamel	Rapporteurs
Henri Méloni	
Frédéric Bimbot	Examinateurs
Gérard Chollet	
John Mason	
Chafic Mokbel	
Christian Wellekens	

A la mémoire de mon père.

A Diana.

It is better to be lucky. But I would rather be exact. Then when luck comes you are ready.

Ernest Hemingway, *The Old Man and the Sea.*

Si on fermait la porte à toutes les erreurs, la vérité resterait dehors.

Rabindranath Tagore.

Résumé

Approches statistiques et filtrage vectoriel de trajectoires spectrales pour l'identification du locuteur indépendante du texte

L'identification du locuteur consiste à attribuer une identité au locuteur d'un énoncé. Cette identité sera celle du locuteur d'une base de référence qui est le plus proche de ce locuteur inconnu, au sens d'une mesure de similarité donnée. Le mode indépendant du texte signifie qu'il n'y a aucune contrainte sur le contenu des phrases prononcées.

Au cours de cette thèse, nous développons un ensemble de mesures de similarité reposant sur une modélisation statistique Gaussienne de vecteurs de paramètres obtenus à l'issue d'une analyse spectrale. Ces mesures reposent essentiellement sur les matrices de covariance de ces vecteurs de paramètres. Une symétrisation de ces mesures est également proposée. Toutes ces mesures, sous leurs différentes formes, sont systématiquement testées sur les bases TIMIT et NTIMIT.

Une fois ces mesures de référence établies, nous tentons de prendre en compte les aspects dynamiques des séquences de vecteurs de paramètres. Ceci nous conduit à étudier les modèles auto-régressifs vectoriels dans le cadre de l'identification du locuteur. Nous testons systématiquement différentes façons de combiner les erreurs résiduelles de prédiction obtenues à l'aide de ces modèles, et comparons les résultats aux mesures de référence précédentes. Nous mettons en œuvre également un protocole expérimental qui permet de mesurer l'efficacité des modèles AR-vectoriels en identification du locuteur après avoir détruit la structure temporelle des vecteurs de paramètres.

Nous établissons alors un formalisme beaucoup plus général pour le filtrage des séquences de vecteurs de paramètres, que nous appelons filtrage vectoriel de trajectoires spectrales. Ce formalisme englobe un grand nombre d'approches classiques en traitement de la parole, parmi lesquelles on trouve les modèles AR-vectoriels, l'analyse cepstrale, les paramètres Δ et $\Delta\Delta$, la paramétrisation RASTA, la transformée en cosinus de trajectoires spectrales, ... L'avantage de ce type de filtrage est qu'il opère simultanément dans les dimensions temporelle et fréquentielle.

Nous présentons finalement un autre filtrage particulier, qui entre dans le cadre du formalisme précédent. Ce filtrage repose sur une analyse en composantes principales temps-fréquence de parole multi-locuteur.

En conclusion, il apparaît que le filtrage vectoriel de trajectoires spectrales est très prometteur, puisqu'il permet de prendre en compte une évolution temporelle des vecteurs de paramètres, tout en filtrant ces mêmes vecteurs dans la dimension fréquentielle. Il permet aussi d'unifier de nombreuses approches différentes. En outre, ce travail suggère de nouvelles approches au niveau de la représentation du signal de parole, et plus particulièrement dans le cadre de la reconnaissance de la parole et du locuteur. Ce type de filtrage peut enfin s'appliquer à d'autres familles de signaux.

Mots-clés : Reconnaissance du locuteur, identification du locuteur, mode indépendant du texte, mesures statistiques du second-ordre, modèles AR-vectoriels, filtrage vectoriel de trajectoires spectrales, filtrage vectoriel temps-fréquence, composantes principales temps-fréquence.

Abstract

Statistical Approaches and Filtering of Spectral Trajectories for Text-Independent Speaker Identification

Speaker identification consists of attributing an identity to a speaker of a utterance. This identity will correspond to the speaker of a reference set who is closest to the unknown speaker with respect to a given similarity measure. The text-independent mode means that there is no constraint on the text content of the utterance.

In this thesis, we develop a set of similarity measures based on multivariate Gaussian modeling of parameter vectors obtained by spectral analysis. These measures essentially use the covariance matrices of the parameter vectors. A symmetrization of these measures is also proposed. All these measures, with all their different forms, are systematically tested on the TIMIT and NTIMIT databases.

Once these reference measures are defined, we try to take into consideration the dynamics of the sequences of the parameter vectors. For this purpose, auto-regressive vector models are studied in the framework of speaker identification. Different ways of combining the prediction residual errors obtained by these models are systematically tested, and compared with the previous reference measures. We also propose an experimental protocol to measure the efficiency of AR-vector models in speaker identification after the destruction of the temporal structure of parameter vectors.

We then establish a more general formalism for the filtering of the sequences of parameter vectors, that we call vector filtering of spectral trajectories. This formalism includes a lot of classical approaches in speech processing, like for instance AR-vector models, Δ and $\Delta\Delta$ parameters, RASTA parameters, cosine transform of spectral trajectories,... The main advantage of this vector filtering is the possibility of filtering simultaneously in both the temporal and frequency domains.

Finally, a new filtering method is presented, which is a special case of the previous formalism. This filtering is based on a time-frequency principal component analysis of utterances recorded from many speakers.

In conclusion, it appears that the vector filtering of spectral trajectories is very promising, because it allows us to take into account a temporal evolution of the parameter vectors, and, at the same time, to filter these vectors in the frequency domain. It is also a way of unifying many different approaches. Moreover, this thesis suggests new approaches for representing speech signals, particularly in the framework of speech and speaker recognition. Finally, this type of filtering can be applied to other signal families.

Keywords : Speaker recognition, Speaker identification, Text-independent mode, Second-order statistical measures, AR-vector models, Filtering of spectral trajectories, Time-frequency filtering, Time-frequency principal components.

Remerciements

Un travail de thèse est une entreprise de longue haleine, qui nécessite l'accompagnement et le soutien de nombreuses personnes. Je voudrais donc leur consacrer quelques pages de ma thèse, afin de leur manifester ma très vive reconnaissance.

Je voudrais commencer par remercier ceux qui ont rendu possible ce travail, et en particulier ceux qui lui ont trouvé un financement : M. Daniel Claude, *responsable du DEA Automatique et Traitement du Signal, à Orsay ;* M. Yves Grenier, *correspondant de ce DEA au département Signal de l'ENST ;* M. Henri Barral *et* M. Pierre Duhamel, *chefs successifs du département Signal de l'ENST.*

Je voudrais également remercier les différents membres de mon jury de thèse, en commençant par M. Jean-Paul Haton, *qui a accepté de présider ce jury, et qui m'a accueilli dans son équipe de recherche nancéenne pour une journée de présentation de mon travail. Merci pour son accueil, sa gentillesse, et son excellent café.*

Merci à Mme Lori Lamel *et à* M. Henri Méloni, *qui ont accepté d'être rapporteurs de ce travail, malgré leur emploi du temps bien rempli et la charge que cette fonction représente.*

Merci enfin à Messieurs Frédéric Bimbot, Gérard Chollet, John Mason, Chafic Mokbel, *et* Christian Wellekens, *examinateurs de ce travail.*

Au cours de ma thèse, j'ai eu l'occasion de passer plusieurs semaines

dans deux laboratoires de recherche. Tout d'abord au Laboratoire d'Informatique d'Avignon, et je remercie pour cela M. Henri MÉLONI, *responsable de l'équipe qui m'a accueilli, ainsi que* M. Jean-François BONASTRE, *dont la collaboration a été très fructueuse. Je voudrais également remercier* Jean-François *pour m'avoir hébergé durant mes deux séjours en Avignon, ainsi que* Thierry, Laurence, Gillou, Fred, Stéphane, Georges, Pierre, Marie *et* Tajine, *pour leur gentillesse et leur accueil. Merci enfin au* Monbar, *pour ses délicieux sandwiches.*

Le deuxième laboratoire qui m'a accueilli est le "speech group" du département d'Electrical Engineering de l'Université de Swansea, au Pays de Galles. Un grand merci à M. John MASON, *qui dirige ce groupe, et qui a rendu cette expérience possible à travers un programme d'échange Européen. Merci à* Hywel, *qui a mis son logement à ma disposition. Merci à* Kin, Tim, John, Jonathan, Marc et Claude, *qui ont accompagné mes premiers balbutiements de conversation anglaise.*

Merci à l'équipe Parole de l'ENST, et plus particulièrement aux habitants de la pièce C125, parce qu'ils ont su toujours y faire régner une ambiance de détente et de bonne humeur. Merci à Christoph, Gilles, Jean-Benoit, Petr, Premek, Sabine, Shigeki. *Et encore merci pour toutes les réponses que vous avez données à toutes mes "petites questions".*

Merci à tous les thésards du département Signal pour avoir su créer un groupe sympathique et agréable. Merci à tous ceux qui ont déjà soutenu leur thèse pour leurs délicieux "pots de thèse", et bon courage à tous ceux qui sont encore en plein dedans.

Merci à M. Gilles DAUPHIN, *le monsieur "réseau" du département, pour avoir maintenu pendant plus de trois ans le réseau du département en état de fonctionnement, malgré toutes mes tentatives pour faire "planter" les machines.*

Merci à toute l'équipe de la documentation de l'école, parce que je les ai souvent mis à contribution, notamment lors de recherches d'articles ou de livres un peu anciens. Ils ont toujours trouvé ce que je leur demandais, dans des délais brefs, et toujours avec bonne humeur.

Merci aux différents relecteurs de ce document.

Il est une personne que je voudrais tout particulièrement remercier, c'est la personne qui est à l'initiative de tout ce travail. Merci à Frédéric, *pour ton enthousiame de chercheur, pour la confiance que tu as mise en moi, pour avoir toujours été capable de me redonner foi en la recherche dans mes moments de découragements, pour m'avoir appris la rigueur de raisonnement et surtout la rigueur dans l'expérimentation scientifique, pour tes quelques rares "coups de semonce" qui ont su me remettre en route, pour ton exigence, pour ton aide précieuse dans certaines "chasses aux bugs", pour m'avoir permis de découvrir l'Irlande, pour toutes les discussions enrichissantes que nous avons eues ensemble à propos de sujets divers, ...*

Un grand merci à tous ceux de mes proches qui m'ont soutenu par leur amour ou leur amitié. Merci à mes amis. Merci à ma famille. Ils savent déjà depuis longtemps combien leur soutien m'est important.

Merci à tous ceux que j'ai oubliés dans ces pages, et dont j'ai croisé la route d'une façon ou d'une autre au cours de ces trois années riches d'expériences.

Merci enfin à Diana, *qui illumine ma vie depuis quelque temps :)*

Ivan Magrin-Chagnolleau

Table des matières

Liste des tableaux

Liste des figures

Notations

Par convention, les lettres minuscules ou majuscules, quand elles ne sont pas en gras, désignent des scalaires (S, p, ...) ; les lettres minuscules en gras désignent des vecteurs ($\mathbf{x}_t$, $\bar{\mathbf{x}}$, ...) ; et les lettres majuscules en gras des matrices ($\boldsymbol{\mathcal{X}}_0$, $\mathbf{X}_k$, ...).

S	Nombre de locuteurs dans la base de référence.
p	Dimension des vecteurs fournis par l'analyse acoustique.
$\mathcal{X}$	Un des locuteurs de la base de référence.
M	Nombre de vecteurs issus de l'analyse acoustique du signal de parole prononcé par le locuteur $\mathcal{X}$.
$\{\mathbf{x}_t\}_{1\leq t\leq M}$	Séquence de M vecteurs issus d'une analyse acoustique de dimension p du signal de parole prononcé par le locuteur $\mathcal{X}$.
$\{\mathbf{x}_t^*\}_{1\leq t\leq M}$	Séquence centrée correspondante.
$\bar{\mathbf{x}}$	Vecteur moyen de la séquence $\{\mathbf{x}_t\}_{1\leq t\leq M}$.
$\boldsymbol{\mathcal{X}}_0$	Matrice de covariance de la séquence $\{\mathbf{x}_t\}_{1\leq t\leq M}$.
$\boldsymbol{\mathcal{X}}_k$	Matrice de covariance décalée d'ordre k de la séquence $\{\mathbf{x}_t\}_{1\leq t\leq M}$.
$\mathbf{X}_k$	Matrice bloc-Toeplitz d'ordre k constituée des différentes matrices de covariance décalées de la séquence $\{\mathbf{x}_t\}_{1\leq t\leq M}$.

$\mathcal{Y}$	Locuteur de test, i.e. locuteur dont on cherche l'identité.
N	Nombre de vecteurs issus de l'analyse acoustique du signal de parole prononcé par le locuteur $\mathcal{Y}$.
$\{\mathbf{y}_t\}_{1\leq t\leq N}$	Séquence de N vecteurs issus d'une analyse acoustique de dimension p du signal de parole prononcé par le locuteur $\mathcal{Y}$.
$\{\mathbf{y}_t^*\}_{1\leq t\leq M}$	Séquence centrée correspondante.
$\bar{\mathbf{y}}$	Vecteur moyen de la séquence $\{\mathbf{y}_t\}_{1\leq t\leq M}$.
$\boldsymbol{\mathcal{Y}}_0$	Matrice de covariance de la séquence $\{\mathbf{y}_t\}_{1\leq t\leq M}$.
$\boldsymbol{\mathcal{Y}}_k$	Matrice de covariance décalée d'ordre k de la séquence $\{\mathbf{y}_t\}_{1\leq t\leq M}$.
$\mathbf{Y}_k$	Matrice bloc-Toeplitz d'ordre k constituée des différentes matrices de covariance décalées de la séquence $\{\mathbf{y}_t\}_{1\leq t\leq M}$.

Introduction

Introduction

Préliminaires

La **Reconnaissance Automatique du Locuteur (R.A.L.)** consiste à utiliser les caractéristiques de la voix d'un locuteur pour l'identifier, ou pour vérifier que son identité est bien celle qu'il a proclamée.

Un système de reconnaissance automatique du locuteur se divise généralement en quatre modules : un module de **paramétrisation** du signal de parole, qui est généralement constitué d'une analyse spectrale vectorielle ; un module de **modélisation**, qui détermine les caractéristiques d'un modèle à partir des paramètres extraits ; un module de **comparaison**, qui consiste à utiliser des mesures de similarité entre modèles ou entre paramètres, voire entre paramètres et modèles ; et enfin un module de **décision**, qui fournit finalement la réponse du système.

Plusieurs éléments nous permettent d'espérer de bonnes performances dans ce domaine : le fait qu'il n'existe pas deux conduits vocaux identiques ; la capacité pour quelqu'un de reconnaître assez facilement les personnes de son entourage, même à travers le téléphone.

Cependant, on ne peut pas parler pour autant d'empreinte vocale. Le terme de signature vocale semble plus approprié. En effet, on trouve d'abord une grande variabilité dans la voix d'un même locuteur, due par exemple à son état de fatigue, à son état émotionnel, à son état de santé, ... Ensuite, un locuteur peut également modifier intentionnellement sa voix. Pour finir, une autre personne peut tenter d'imiter ce locuteur, et pourrait éventuellement se faire passer pour lui.

Malgré ces inconvénients, cette discipline est actuellement en plein essor. Elle offre en effet de nombreuses applications potentielles (validation de transactions par le téléphone, contrôle supplémentaire au niveau d'une application sur site comme l'accès sécurisé à un bâtiment, remplacement du mot de passe sur les ordinateurs, ...). De plus, le domaine semble suffisamment avancé scientifiquement pour espérer des progrès rapides dans les années à venir, en termes de performances et de développement d'applications.

Motivations et objectifs

Le travail exposé dans cette thèse cherche à atteindre plusieurs objectifs. Nous voulons dans un premier temps établir une famille de mesures de similarité de référence, reposant sur une modélisation du locuteur simple, qui puisse servir de point de départ à de nombreuses directions d'étude. Cette famille de mesures répond également à un besoin, qui selon nous est primordial en reconnaissance automatique du locuteur, qui est de fournir une méthode de référence pouvant être systématiquement utilisée pour l'évaluation. Nous proposons donc de modéliser un locuteur par les caractéristiques du second ordre (vecteur moyen et matrice de covariance) d'une séquence de vecteurs de paramètres extraite d'un énoncé de ce locuteur. Puis nous détaillons des mesures de similarité qui utilisent uniquement ces caractéristiques du second ordre.

L'inconvénient majeur de cette approche est qu'elle repose sur une modélisation globale, et qu'elle ne permet donc pas de capter l'information dynamique de la séquence de vecteurs. En particulier, cette limitation ne permet pas d'espérer de bonnes performances lorsque la qualité de la parole se dégrade, et notamment dans le cas de la parole téléphonique. En effet, dans ce cas, les caractéristiques globales sont fortement influencées par les conditions de transmission. D'autre part, chaque locuteur adopte des stratégies différentes au niveau articulatoire, et ce niveau se retrouve essentiellement dans les caractéristiques dynamiques, bien plus que dans les caractéristiques statiques.

Nous proposons donc une approche qui permet de modéliser les caractéristiques dynamiques du locuteur. Cette approche repose sur un filtrage vectoriel de trajectoires spectrales, c'est-à-dire opérant sur une séquence de vecteurs de paramètres. Nous développons un formalisme qui englobe une grande partie des filtrages déjà existant. Puis nous étudions quelques filt-

rages particuliers pour l'identification du locuteur.

La possibilité d'interpréter ces différents filtrages comme des masques temps-fréquence appliqués à des séquences de vecteurs de paramètres nous apparaît comme un avantage certain de cette approche, alors que la plupart des filtrages présentés jusqu'à maintenant n'étaient appliqués que dans une dimension temporelle ou fréquentielle.

Au fil des pages

Le chapitre 1 présente le domaine de la reconnaissance du locuteur de manière générale. On y trouve les définitions importantes, la description de la structure d'un système de reconnaissance, quelques applications, et l'explication des choix que nous avons faits a priori pour ce travail.

Le chapitre 2 présente une bibliographie chronologique non-exhaustive sur la reconnaissance automatique du locuteur en général, en décrivant notamment les différentes familles d'approches, et en mettant l'accent sur les méthodes de type statistique. Cette étude bibliographique expose pour finir les tendances actuelles.

Le chapitre 3 est consacré aux conditions expérimentales. Il décrit les bases de données que nous avons utilisées pour nos expériences, et discute en particulier les différentes propriétés d'une base de données, ainsi que l'adéquation des bases actuelles au problème de la reconnaissance automatique du locuteur. Nous y décrivons également notre analyse acoustique, ainsi que les différents protocoles expérimentaux adoptés.

Le chapitre 4 est une description détaillée de plusieurs méthodes statistiques du second ordre. Ces méthodes reposent sur une modélisation des vecteurs de paramètres par leurs caractéristiques du second ordre (moyennes et matrices de covariances). Comme elles ne sont pas symétriques dans leurs formes de base, nous proposons quelques symétrisations empiriques de ces méthodes. Nous consacrons également une section de ce chapitre à l'utilisation des méthodes statistiques du second ordre dans le cadre d'une approche analytique en reconnaissance automatique du locuteur, c'est-à-dire une méthode prenant en compte un certain nombre d'informations a priori sur les segments de parole.

Le chapitre 5 constitue une étude systématique de différentes mesures et diverses normalisations appliquées à des résiduels de prédiction linéaire vectorielle. En particulier, on discute la pertinence de l'hypothèse selon laquelle les modèles AR-vectoriels modélisent des caractéristiques dynamiques du locuteur. Pour cela, nous utilisons un protocole expérimental original consistant à détruire l'ordre temporel des vecteurs acoustiques en les mélangeant aléatoirement.

Le chapitre 6 introduit le concept de filtrage vectoriel de trajectoires spectrales, et en donne un formalisme mathématique. Le formalisme énoncé est commun à différents filtrages linéaires déjà existants (il englobe en particulier les modèles AR-vectoriels, l'analyse cepstrale, ou encore les coefficients Δ et $\Delta\Delta$). Les filtres appliqués aux séquences de vecteurs de paramètres sont interprétés en termes de masques temps-fréquence.

Le chapitre 7 présente un filtrage particulier, qui s'inscrit dans le formalisme du chapitre précédent, et qui repose sur une analyse en composantes principales temps-fréquence de parole multi-locuteur.

La partie III rassemble les principales conclusions de ce travail, en soulignant ses contributions majeures. On y met en évidence, en particulier, les avantages d'une méthode de référence pour l'évaluation, ainsi que l'intérêt d'un filtrage vectoriel opérant conjointement dans les domaines temporel et fréquentiel. Le filtrage vectoriel est riche en perspectives, et nous en présentons quelques unes pour clore ce document.

On trouve finalement différentes annexes, parmi lesquelles une étude sur les conditions d'analyse (annexe A), quelques résultats supplémentaires sur les mesures statistiques du second ordre (annexe B), le détail du calcul des coefficients matriciels d'un modèle AR-vectoriel (annexe C), la démonstration de l'invariance des mesures statistiques du second ordre par filtrage inversible (annexe D), la représentation visuelle d'un grand nombre de composantes principales (annexe E, annexe F et annexe G), et enfin un exemplaire des publications produites pendant cette thèse (annexe H).

Les références bibliographiques pour l'ensemble de la thèse sont rassemblées à la fin du document.

Approches statistiques et filtrage vectoriel de trajectoires spectrales pour l'identification du locuteur indépendante du texte

Chapitre 1

Contexte

Ce premier chapitre est destiné à placer le travail de cette thèse dans son contexte. Nous donnons entre autres la définition des termes les plus usuels, nous proposons quelques applications de la reconnaissance automatique du locuteur, nous réfléchissons sur la dualité méthodes globales – méthodes analytiques, et nous justifions finalement les choix a priori que nous avons faits au début de ce travail.

1.1 La communication Homme-Machine

Les recherches en reconnaissance automatique du locuteur (R.A.L.) font partie du domaine plus large de la communication Homme-Machine (cf. Fig. 1.1). Dans ce contexte en effet, il est souhaitable qu'une machine puisse identifier automatiquement la personne qui lui parle, comme un locuteur le fait naturellement au cours d'une conversation. Cela peut-être nécessaire pour une authentification vocale (mot de passe vocal par exemple), ou pour aider à d'autres tâches (reconnaissance de la parole, synthèse de la parole, ...).

1.2 Terminologie

Nous définissons dans cette section un certain nombre de termes fréquemment utilisés dans le domaine de la R.A.L.

1.2.1 Identification et vérification

Commençons par définir les deux principales tâches que l'on distingue en reconnaissance automatique du locuteur, ainsi que les différentes phases

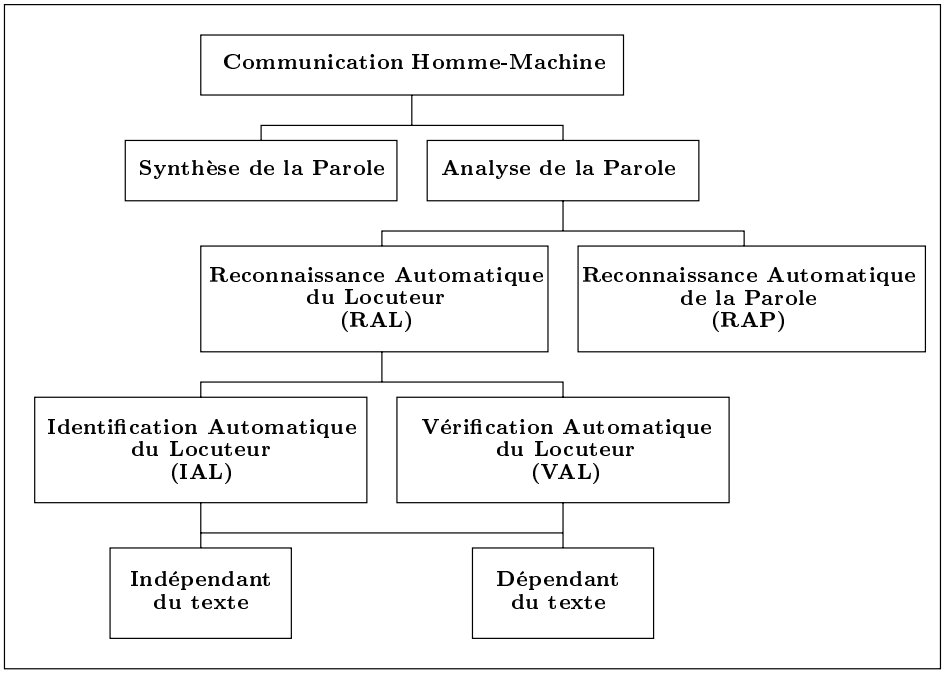


FIG. 1.1: *Contexte d'étude : la reconnaissance automatique du locuteur dans le contexte de la communication homme-machine.*

qui constituent chacune de ces tâches.

La première phase d'un système de reconnaissance du locuteur est la phase d'apprentissage (cf. FIG. 1.2). Au cours de cette phase, on construit une base de référence contenant des données (signaux, paramètres, modèles) relatives à un nombre de locuteurs fixé S. Remarquons qu'il n'est pas toujours nécessaire d'avoir constitué une importante base de données pour pouvoir commencer l'identification ou la vérification.

La phase de test dépend alors de la tâche qui est réalisée. On en distingue essentiellement trois :

- **Identification en ensemble fermé** [5], [34], [117] : on dispose uniquement d'un échantillon de parole du locuteur inconnu. Le système fournit en sortie l'identité du locuteur de la base de référence dont le locuteur inconnu est le plus "proche". Cette décision est prise après avoir comparé le locuteur inconnu à tous les locuteurs de la base de référence. Il s'agit d'une décision de type 1 parmi S. On suppose en

fait que le locuteur inconnu fait nécessairement partie de la base de référence (cf. FIG. 1.3).

- **Vérification** [5], [137], [34], [117] : on dispose d'un échantillon de parole du locuteur inconnu ainsi que d'une identité proclamée, identité qui est celle de l'un des locuteurs de la base de référence. Le système doit alors vérifier si cette identité est correcte. On dit que le système rejette le locuteur si cette identité est considérée comme erronée, et qu'il l'accepte s'il juge cette identité correcte. Il s'agit cette fois d'une décision binaire (cf. FIG. 1.4).
- **Identification en ensemble ouvert** [34] : c'est une combinaison des deux tâches précédentes. Le système commence par faire une identification, et choisit donc le locuteur de la base de référence qui est le plus proche du locuteur inconnu. Puis il décide finalement si c'est bien ce locuteur-là (cf. FIG. 1.5).

1.2.2 Typologie des erreurs

Chaque tâche possède ses propres erreurs. Nous rappelons ici la typologie de chacune d'elles.

Les performances d'un système d'identification en ensemble fermé se mesurent par son **taux de mauvaise identification**.

Celles d'un système de vérification se mesurent par son **taux de fausse acceptation** et par son **taux de faux rejet**. La fausse acceptation correspond au cas où le système accepte le locuteur inconnu alors que celui-ci n'est pas la personne qu'il prétend être. Le faux rejet correspond au cas où le système rejette le locuteur inconnu alors qu'il est vraiment la personne dont il a donné l'identité au système.

Enfin, les performances d'un système d'identification en ensemble ouvert se mesurent par son taux de **mauvaise identification**, c'est-à-dire un locuteur faisant partie de la base de référence est reconnu comme un autre locuteur de cette base, son **taux de fausse acceptation**, c'est-à-dire un imposteur est accepté comme l'un des locuteurs de la base de référence, et son **taux de faux rejet**, dans le cas où un locuteur faisant partie de la base de référence est rejeté.

Cette typologie des erreurs est très clairement présentée dans les articles suivant : [5], [17], [14].

Enfin, un très gros travail a été fait au niveau de la terminologie et de l'évaluation en reconnaissance du locuteur, dans le cadre du projet Européen EAGLES [28]. Je ne peux que renvoyer à cette référence indispensable, en espérant qu'elle contribuera à uniformiser les procédures d'évaluation.

1.2.3 Dépendance au texte

On distingue classiquement en reconnaissance automatique du locuteur deux types de contraintes par rapport au texte, l'une que l'on appelle **dépendante du texte**, et l'autre que l'on nomme **indépendante du texte** [5], [137], [34], [117]. Mais cette terminologie ne rend pas bien compte des différentes dépendances au texte possible, comme le remarquent les auteurs du rapport du projet Européen SAM-A [17], [14]. Les différents systèmes y sont classés, du plus contraignant au moins contraignant, de la façon suivante :

- **Système à texte fixé dépendant du locuteur (user-specific text-dependent)** : pour un locuteur donné, le texte est toujours le même d'une session à l'autre. Mais chaque locuteur a un texte différent.
- **Système dépendant du vocabulaire (vocabulary-dependent)** : l'utilisateur du système prononce une séquence de mots, issus d'un vocabulaire limité (des séquences de chiffres par exemple), mais dont l'ordre peut varier d'une session à l'autre.
- **Système dépendant d'événements phonétiques (speech-event-dependent)** : le vocabulaire n'est pas directement imposé, mais certains événements phonétiques doivent être présents dans la séquence de parole prononcée (présence de certaines voyelles, de certaines nasales, ...). Les phrases à prononcer peuvent éventuellement être affichées sur l'écran à chaque session.
- **Système à texte imposé par la machine (text prompted)** : le texte est différent pour chaque session et pour chaque locuteur, mais affiché à chaque fois par la machine. Le texte est choisi de manière imprédictible pour éviter l'utilisation d'enregistrements par un imposteur.
- **Système indépendant du texte (independent or free-text)** : le locuteur est entièrement libre de ce qu'il dit à chaque session.

Cette classification rend bien mieux compte des différents systèmes que l'on peut effectivement trouver dans des articles, ou dans des applications, et cela sans ambiguïté.

1.3 Les différents modules de la reconnaissance du locuteur

La reconnaissance automatique du locuteur peut aussi être interprétée comme une tâche particulière de reconnaissance des formes, comme nous l'avons déjà signalé dans la première partie de l'introduction. C'est une succession de modules (paramétrisation, modélisation, comparaison et décision), dont l'étape finale est de reconnaître une forme particulière, le signal de parole que l'on met à l'entrée de cette chaîne. Ces différents modules sont redétaillés sous forme de schémas pour les différentes phases et les différentes tâches. La FIG. 1.2 présente un schéma modulaire de la phase d'apprentissage, qui est la même quelque soit la tâche effectuée. La FIG. 1.3, la FIG. 1.4 et la FIG. 1.5 présentent respectivement des schémas modulaires pour l'identification du locuteur en ensemble fermé, la vérification du locuteur, et l'identification du locuteur en ensemble ouvert.

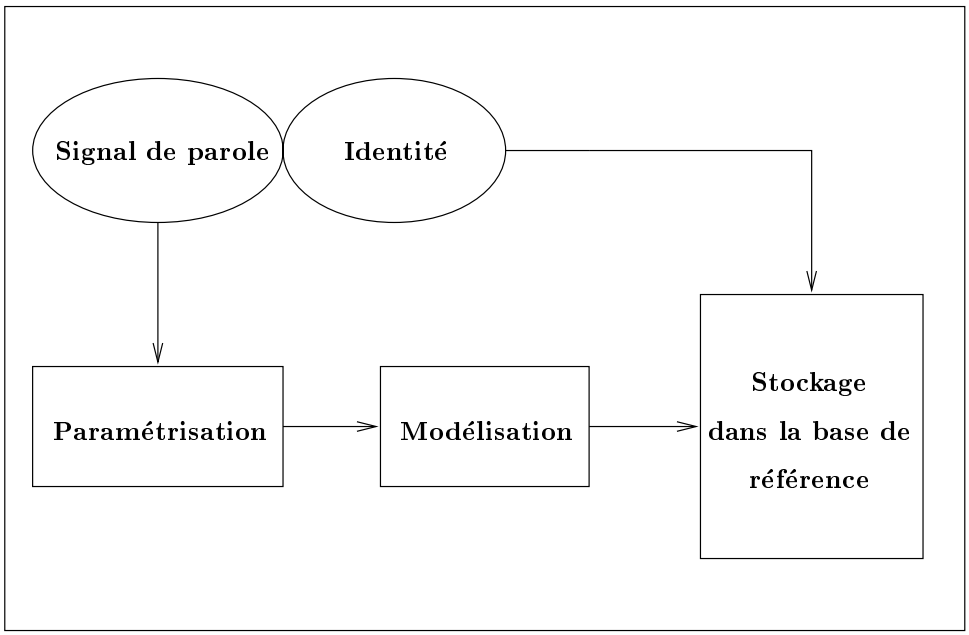


FIG. 1.2: *Schéma modulaire de la phase d'apprentissage en reconnaissance du locuteur.*

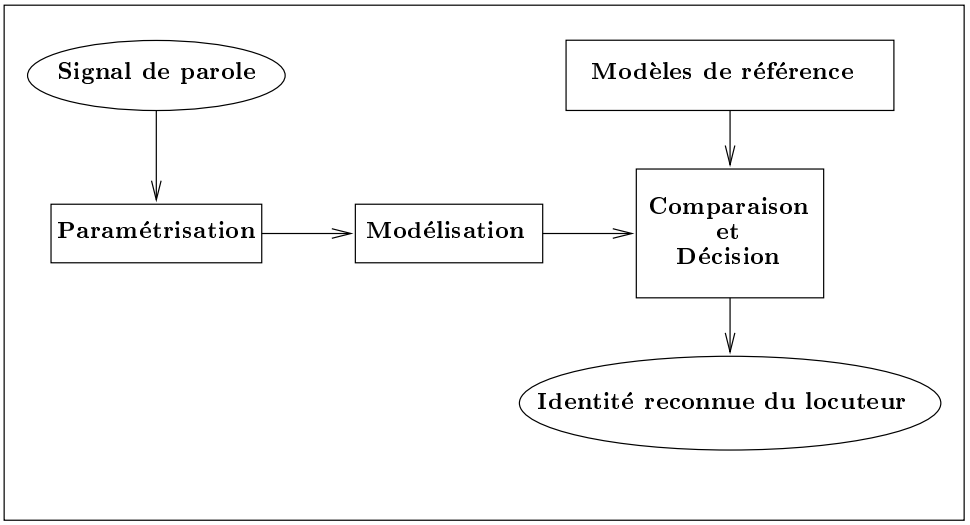


FIG. 1.3: *Schéma modulaire de l'identification du locuteur en ensemble fermé.*

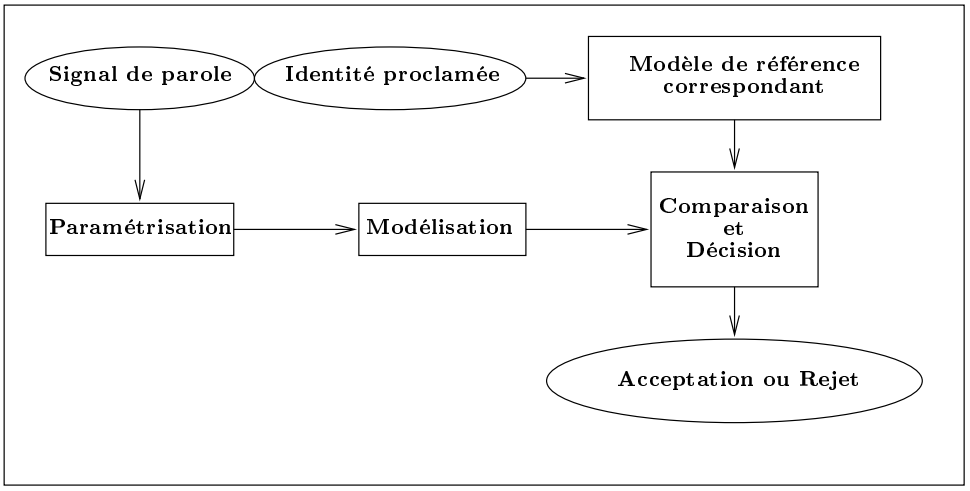


FIG. 1.4: *Schéma modulaire de la vérification du locuteur.*

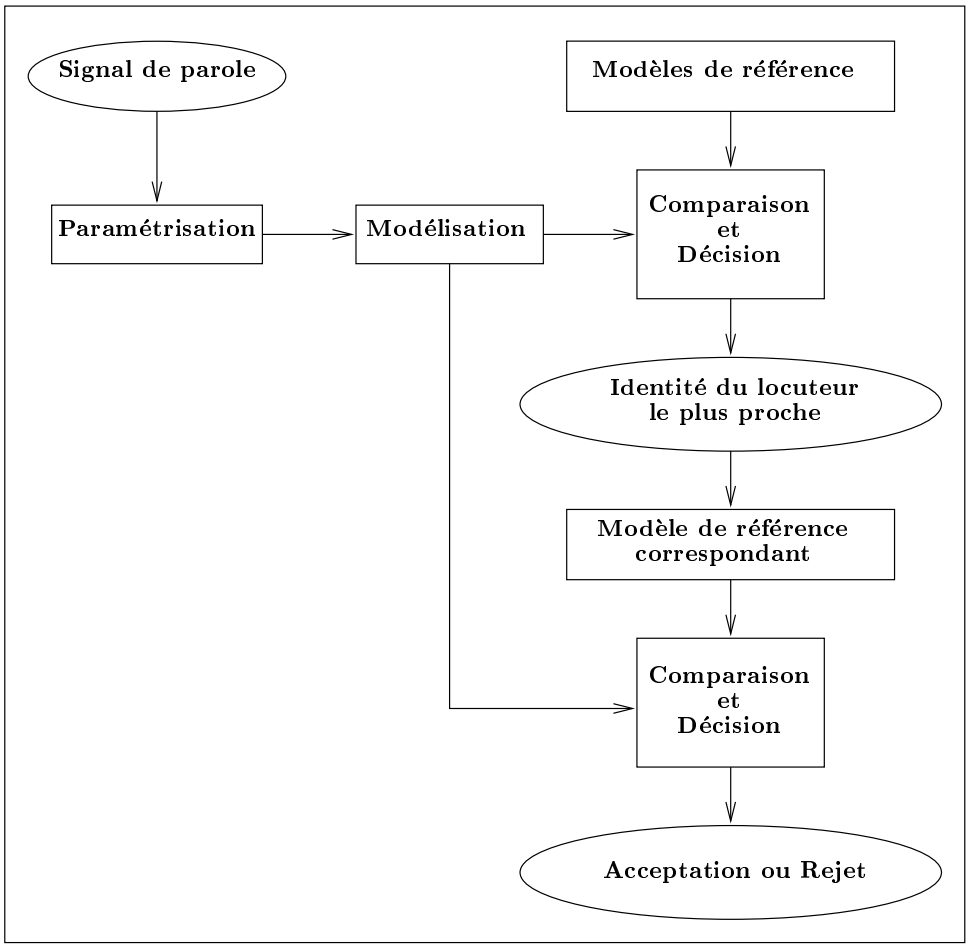


FIG. 1.5: *Schéma modulaire de l'identification du locuteur en ensemble ouvert.*

1.4 Applications

On peut regrouper les applications de la R.A.L. en trois catégories. Nous donnons pour chacune d'elles quelques exemples, ainsi qu'une brève description des contraintes d'environnement. Remarquons que la plupart des applications présentées reposent sur la vérification du locuteur.

1.4.1 Applications sur sites

La première catégorie concerne les applications qui se trouvent sur un site géographique particulier, à l'entrée d'un bâtiment, d'une salle, à un distributeur d'argent. Voici quelques exemples de ce type d'applications :

- ❑ Validation de transactions sur sites (comme contrôle supplémentaire au niveau des distributeurs bancaires par exemple).
- ❑ Applications domestiques (protection de domiciles ou de garages par verrous électroniques).
- ❑ Remplacement des mots de passe sur les ordinateurs (login vocal).

Dressons maintenant un inventaire des différentes contraintes, et des avantages et inconvénients de cette catégorie d'applications :

- ☞ l'environnement peut facilement être contrôlé.
- ☞ l'utilisation de la reconnaissance du locuteur a surtout un rôle dissuasif.
- ☞ l'utilisateur peut avoir son modèle de locuteur sur lui (sur la puce d'une carte par exemple).
- ☞ la reconnaissance vocale peut être associée à une autre technique de reconnaissance d'identité.

On se trouve donc dans un cas plutôt favorable d'utilisation de la reconnaissance du locuteur.

1.4.2 Applications téléphoniques

Ce sont toutes les applications liées à l'usage d'un téléphone, comme par exemple :

- ❑ Contrôle d'accès à des données (sécurité militaire, protection industrielle, accès à des banques de données, ...).
- ❑ Validation de transactions bancaires par téléphone (pour valider légalement la transaction effectuée).

Les conditions d'utilisation ne sont bien sûr pas les mêmes que le type d'applications précédentes :

- ☞ l'environnement est difficilement contrôlable car les lignes téléphoniques peuvent varier considérablement d'un appel à un autre, ainsi que le micro.
- ☞ il n'y a aucun effet dissuasif.
- ☞ ce type d'application requiert de stocker toutes les données de manière centralisée.
- ☞ il est impossible d'utiliser d'autres techniques de reconnaissance (excepté un code numérique tapé sur des touches à fréquences vocales).

Ce type d'applications pose donc de nombreux problèmes supplémentaires. Le projet Européen CAVE (CAller VErification) porte d'ailleurs sur cet aspect applicatif de la RAL [72].

1.4.3 Applications judiciaires

Enfin, on trouve le domaine des applications judiciaires, qui pose actuellement le plus de problèmes. La reconnaissance du locuteur est par exemple utilisée pour :

- ❑ les orientations d'enquêtes.
- ❑ la constitution d'éléments de preuve au cours d'un procès.

Nous sommes cette fois-ci dans le cas le plus défavorable :

- ☞ la quantité de parole à disposition est en général très réduite.
- ☞ les conditions d'environnement sont très mauvaises.
- ☞ les locuteurs impliqués sont très rarement coopératifs.
- ☞ en revanche, on a souvent d'autres techniques pour effectuer une reconnaissance (empreintes digitales, ...).

Il faut être très prudent quant à l'utilisation de la RAL en criminologie. On pourra suivre à ce sujet les réflexions du groupe de travail GT1 du GdR-PRC Communication Homme-Machine, dont le sujet est la caractérisation de la langue et du locuteur, et dont l'un des thèmes de réflexion est l'utilisation de la RAL dans les milieux judiciaires.

Concernant les applications judiciaires de la RAL, on peut aussi consulter Chollet 1991 [27], qui met en garde contre l'application abusive d'une technique, alors que la fiabilité de celle-ci ne permet pas de se reposer sur elle

dans des domaines aussi cruciaux que le domaine judiciaire. Les problèmes d'éthique posés par l'utilisation de la reconnaissance du locuteur dans un domaine aussi sensible sont effectivement difficiles à résoudre.

L'article de Künzel 1994 [80] présente quant à lui un bon état de l'art sur les méthodes de reconnaissance du locuteur qui sont utilisées dans le milieu judiciaire allemand.

Voir enfin plus récemment Chollet 1997 [28].

1.5 Méthodes globales ou analytiques

Maintenant que nous avons présenté les principales applications de la reconnaissance du locuteur, il nous reste à choisir le type d'approches que nous allons utiliser.

Nous écartons de notre étude toute approche qui nécessite l'intervention d'un "expert", pour segmenter les signaux de parole ou pour prendre directement des décisions par l'écoute des segments de parole ou la lecture des spectrogrammes, tant la subjectivité du dit expert occupe une part importante dans la décision finale. On ne pourrait, dans ce cas-là, parler de reproductibilité des expériences.

Nous nous plaçons donc d'emblée dans la famille des méthodes entièrement automatiques. Il reste encore à dissocier deux familles de méthodes, les méthodes globales et les méthodes analytiques.

Les méthodes analytiques consistent à utiliser des informations a priori sur le signal de parole, ces informations ayant été extraites de manière entièrement automatique au cours d'une phase préalable à la reconnaissance. Cette dernière se fait ensuite également à l'aide d'un système automatique, mais après une classification préalable des segments initiaux en fonction de certains types d'information. Dans ce cas-là, la reproductibilité des expériences est préservée, mais le processus mis en oeuvre pour extraire cette information a priori génère ses propres erreurs, et il est difficile de les dissocier des erreurs dues au système de reconnaissance lui-même.

Une telle approche présente néanmoins de nombreux avantages. Elle permet de rester relié à des interprétations physiologiques, articulatoires, ...

Elle permet en outre de mieux comprendre sur quel type de données une méthode automatique extrait le plus d'information sur le locuteur.

Cette démarche a été abordée dans [91], où il est surtout question de caractériser un peu mieux les méthodes statistiques du second-ordre, au regard de la durée des segments de parole utilisés, et de leur contenu phonétique. Nous présentons plus en détails ces expériences dans le chapitre consacré aux méthodes statistiques du second ordre. Pour en savoir plus sur les méthodes analytiques, on peut aussi consulter la thèse de Bonastre [19], ainsi que [20] et [21].

Le reste de notre travail ne fait pas du tout intervenir de phase analytique préalable. Nous utilisons donc uniquement des méthodes automatiques globales. La dernière partie de notre travail présente cependant une alternative aux méthodes analytiques, en proposant la prise en compte de la dynamique des séquences de vecteurs de paramètres par des techniques de filtrages vectoriels.

1.6 Choix effectués

Une fois que nous avons choisi d'adopter une démarche automatique et globale, il nous reste encore un certain nombre de décisions à prendre, tant est importante la variété des configurations possibles. En particulier, il nous faut choisir la tâche de reconnaissance sur laquelle nous allons travailler, le type de dépendance au texte que nous allons privilégier, et la famille de techniques que nous allons employer, pour mener à bien nos différents travaux.

1.6.1 Nature de la tâche

Bien que la plupart des applications utilisent la vérification du locuteur, nous avons choisi de traiter **l'identification du locuteur en ensemble fermé**. En effet, nous souhaitons étudier différentes approches, et pouvoir les comparer objectivement entre elles, sans avoir recours à une stratégie de décision particulière. Cette première étape permet de comparer des méthodes de classification. En outre, les résultats obtenus en terme de performances sont généralement transposables en vérification du locuteur, c'est-à-dire qu'une amélioration des performances en identification du locuteur, avec une méthode donnée, conduit la plupart du temps à une

amélioration en vérification avec cette même méthode.

En vérification du locuteur, la méthode de classification joue un rôle certes non négligeable, mais le critère de décision utilisé a aussi son importance, ainsi que d'autres facteurs, et il est difficile de déterminer si le succès d'une méthode de vérification est dû à l'un ou à l'autre. D'autres problèmes délicats tels que le choix des imposteurs, le choix du modèle d'imposteur pour chaque locuteur de la base, la détermination des seuils *a priori*, constituent à eux seuls des sujets de recherche.

1.6.2 Nature de la dépendance au texte

Nous avons choisi de travailler **en mode indépendant du texte**, le locuteur étant entièrement libre de ce qu'il prononce. C'est le mode qui contraint le moins les utilisateurs dans une application. Ce choix présente un certain nombre d'autres avantages. Tout d'abord, avec ce type de contrainte au texte, on fait l'économie d'un alignement temporel, ce qui peut aussi être le cas en mode dépendant du texte, à condition de ne pas utiliser la technique de programmation dynamique (DTW). L'intérêt de ne pas avoir d'alignement temporel à faire est aussi d'augmenter la robustesse du système aux effets de parole spontanée (hésitations, mot remplacé par un autre, ...). Ce mode permet enfin de poursuivre identification ou vérification en cours de transaction.

Cette approche présente néanmoins un inconvénient majeur. Si un imposteur possède un enregistrement d'un des locuteurs de la base de référence, il pourra facilement duper le système, puisque ce dernier ne lui imposera aucun texte particulier. Dans le cas d'une application particulière, il faudra prendre en compte cette éventualité. Un autre inconvénient de ce type de dépendance au texte est qu'il nécessite souvent plus de parole que lorsqu'on travaille avec un texte fixé pour chaque utilisateur. Dans ce cas-là, la durée du segment de parole peut effectivement être très courte. Néanmoins, le fait d'avoir besoin de davantage de parole quand on travaille en mode indépendant du texte vient surtout de la complexité des modèles statistiques utilisés. En effet, plus les modèles sont complexes et plus la quantité de parole nécessaire pour les caractériser est importante. Il y a là un compromis à trouver entre complexité du modèle et contrainte sur la durée du texte.

1.6.3 Techniques choisies

Nous avons enfin choisi de travailler sur des approches de type statistique, en postulant en particulier une répartition Gaussienne des vecteurs de paramètres acoustiques. Les locuteurs sont donc représentés par les caractéristiques du second ordre de ces vecteurs (moyennes, matrices de covariance, matrices de covariances décalées). Cependant, cette hypothèse de Gaussianité ne sera pas toujours nécessaire.

Si nous avons fait le choix d'une technique relativement simple, c'est aussi avec la conviction que toute amélioration apportée au classificateur Gaussien peut se transposer à d'autres techniques plus complexes comme les mélanges de Gaussiennes ou les modèles de Markov cachés. L'extension des résultats de cette thèse à d'autres techniques fait d'ailleurs partie des perspectives de notre travail.

Nous avons maintenant clairement indiqué nos choix, quant à la nature de la tâche, à la dépendance au texte, et à la technique choisie. La présentation d'autres travaux dans le domaine fait l'objet du chapitre suivant, qui fournit quelques repères bibliographiques sur la reconnaissance du locuteur. On y trouve entre autres choses une brève présentation des autres techniques utilisées en R.A.L.

Chapitre 2

Repères bibliographiques en reconnaissance du locuteur

Faire une bibliographie exhaustive en reconnaissance du locuteur est une tâche quasi-insurmontable tant est grand le nombre de travaux portant sur ce sujet. Je me suis donc attaché, au cours de ce survol, à citer d'une part les quelques articles incontournables du domaine, d'autre part les articles présentant pour la première fois des techniques particulières ou leur utilisation en reconnaissance du locuteur. Enfin, il m'a paru inévitable de citer quelques articles fétiches, ceux que j'ai pris plaisir à lire, ceux qui m'ont apportés quelques lumières sur certains aspects de la reconnaissance du locuteur, et ceux de quelques personnes dont j'apprécie plus particulièrement la clarté, la rigueur scientifique, l'imagination ou l'esprit créatif.

2.1 Quelques points d'entrée

Lorsqu'on aborde un sujet comme la reconnaissance automatique du locuteur, on cherche avant tout à se faire une idée globale du sujet, à se familiariser avec le vocabulaire propre à ce domaine. Nous renvoyons pour cela aux quelques articles qui nous ont aidés à démarrer : [5], [137], [34], [117], [118] (Chapitre 11), [140], [47]. Ces articles constituent en outre de très bons points d'entrée bibliographiques, puisqu'ils renvoient eux-même à de nombreux travaux.

2.2 Les précurseurs

Avant de parler des méthodes automatiques, et de présenter les quelques techniques "traditionnelles", évoquons brièvement quelques travaux précurseurs dans le domaine de la reconnaissance de l'identité par la voix.

Les premiers travaux dans ce domaine remontent aux années 40. A cette époque, la reconnaissance de la voix se faisait surtout par auditeurs-experts, et dans le domaine judiciaire. C'est le cas notamment dans [100] et [101]. D'autres travaux sur la reconnaissance de la voix par auditeurs ont suivi, quelques années plus tard, parmi lesquels on peut citer [119], [32], [163], [23], [69], [159], [29], [62], ou encore [136].

Parmi les approches utilisées très tôt, on trouve aussi la reconnaissance du locuteur à partir de spectrogrammes. C'est le cas notamment dans [77], [81], [159] ou [18]. Mais cette approche, même si elle a recueilli l'adhésion de nombreux chercheurs dans les années 60, a été remise assez vite en question quelques années plus tard par quelques auteurs, dont notamment Ladefoged et Stevens.

Suite à cette remise en question de l'approche par lecture de spectrogrammes, et à cause de la subjectivité de l'approche par auditeurs-experts, il est très vite apparu qu'il était nécessaire de réaliser cette tâche de reconnaissance de manière automatique. Nous allons maintenant aborder ces différentes méthodes automatiques.

2.3 Spectres et cepstres moyens

La première approche automatique très largement répandue a été l'utilisation du spectre moyen à long terme. Pour un locuteur de référence donné, on extrait d'une phrase prononcée un ensemble de vecteurs de paramètres (spectraux, cepstraux, ...), et on les modélise par leur vecteur moyen. Chaque locuteur de référence est ainsi modélisé par un spectre (ou cepstre) moyen global, ou à long terme. On calcule alors un vecteur moyen pour le locuteur de test, puis une distance spectrale (ou cepstrale) entre ce vecteur et un vecteur de référence.

On trouve ce type d'approche dans [44], [94], [95], [154], [153].

2.4 La programmation dynamique

La programmation dynamique est une technique exclusivement utilisée en mode dépendant du texte. Elle permet d'aligner temporellement une phrase de test avec une phrase d'apprentissage, ce qui permet de prendre

en compte les différences de débit qui peuvent survenir entre deux énoncés d'une même phrase par un même locuteur.

Le premier algorithme de programmation dynamique a été proposé par Sakoe et Chiba [144]. Dans cet article, l'algorithme est appliqué en reconnaissance de mots isolés. La programmation dynamique trouve parfaitement son utilité dans ce type d'application, puisqu'elle permet d'aligner temporellement des mots les uns avec les autres. Sur le principe de la programmation dynamique, on peut aussi consulter [124] (Chapitre 4, pages 200 et suivantes). On doit l'une des premières utilisations de la programmation dynamique en reconnaissance du locuteur à Furui 1981 [45]. On trouve également l'utilisation de la programmation dynamique en reconnaissance du locuteur dans les articles suivants : [46], [22], [58], [169] ou encore [170].

2.5 La quantification vectorielle

La quantification vectorielle peut être utilisée indifféremment en mode dépendant ou indépendant du texte. En mode dépendant du texte, elle représente une alternative intéressante à la programmation dynamique. Elle a été utilisée un peu plus tard en mode indépendant du texte. Puis elle a été un peu mise à l'écart comme méthode en tant que telle, mais intervient quelques fois dans la phase de paramétrisation, suivie de l'utilisation d'une modélisation statistique. On l'utilise en particulier comme initialisation pour l'algorithme EM, lorsque l'on utilise les mélanges de Gaussiennes.

La quantification vectorielle consiste à représenter l'ensemble des vecteurs de paramètres extraits le long d'une phrase par un petit nombre de vecteurs représentatifs, appelés généralement centroïdes. On appelle dictionnaire l'ensemble des centroïdes extraits le long d'une phrase. Il existe plusieurs algorithmes pour établir ces dictionnaires. Cette méthode est particulièrement avantageuse quand le signal sur lequel on l'applique présente naturellement une structure en "segments", ce qui est justement le cas du signal de parole.

On trouve une bonne description de cette approche dans [54], qui présente de nombreuses applications où cette technique peut être utilisée. De nombreux articles proposent l'emploi de la quantification vectorielle en reconnaissance du locuteur.

Les Bell Labs se sont beaucoup intéressés à cette technique : [156], [155], [141], [142], [158], [157]. Tous ces articles permettent d'acquérir une bonne connaissance de cette méthode.

Plusieurs articles plus récents ont également proposé l'utilisation de la quantification vectorielle. C'est le cas notamment des articles suivants : [96], [35], [97], [140], [99]. On trouve enfin une bonne description de cette technique dans [124].

2.6 Les modèles de Markov cachés

L'inconvénient majeur de toutes les techniques déjà présentées est qu'elles ne prennent pas en compte la façon dont les vecteurs de paramètres se succèdent. Les modèles de Markov sont l'une des premières tentatives pour résoudre ce problème. Ce modèle a été initialement introduit en reconnaissance de la parole. Puis son utilisation s'est étendue peu à peu au domaine de la reconnaissance du locuteur.

L'un des premiers articles publié sur l'utilisation des modèles de Markov cachés en traitement de la parole est celui de Poritz 1982 [120].

Pour une présentation très claire de la technique, nous renvoyons à Rabiner : [121], [123] ou [124]. Pour une présentation plus mathématique des modèles de Markov cachés, on peut aussi lire [122].

Un modèle de Markov caché est constitué de plusieurs états, chaque état étant caractérisé par une distribution de probabilité. On connait en outre les probabilités de passage d'un état à l'autre. Enfin, les vecteurs de paramètres sont en fait les observations de ce modèle probabiliste, c'est-à-dire chaque état possède une densité de probabilité d'émission de ces différents vecteurs de paramètres. On caractérise alors entièrement un modèle de markov caché par la donnée des différentes probabilités de se trouver à l'instant initial dans chaque état, par la donnée des différentes probabilités de transitions entre les différents états, et par la donnée des différentes densités de probabilités d'émissions.

Quant à l'utilisation des modèles de Markov cachés en reconnaissance du locuteur, on peut se référer à : [172], [147], [139], [162], [138], [99], [98], [165].

2.7 Les mélanges de Gaussiennes

Les modèles de Markov cachés sont l'une des nombreuses méthodes statistiques pour modéliser les vecteurs de paramètres. Parmi ces méthodes, on trouve aussi la modélisation des vecteurs par une densité de probabilité Gaussienne multi-dimensionnelle. Cette technique est détaillée dans le chapitre 4, nous ne revenons donc pas dessus ici. L'une des extensions de cette modélisation Gaussienne est la modélisation par un mélange de densités Gaussiennes. Cette technique a été utilisée assez récemment en reconnaissance du locuteur, et elle fournit actuellement les meilleurs résultats en reconnaissance du locuteur indépendante du texte. On utilise en général un algorithme EM pour estimer les différents paramètres du mélange. La plupart des auteurs prennent d'ailleurs des matrices diagonales pour simplifier un peu les calculs.

On trouve ce type d'approches dans : [135], [134], [128], [133], [129], [132], [130], [131].

2.8 Les réseaux de neurones

Les réseaux de neurones ont été assez largement utilisés en reconnaissance du locuteur. Ils offrent en effet une bonne alternative au problème de la discrimination entre les locuteurs. Ces outils de classification permettent en effet de séparer des classes, dans un espace de représentation donné, de façon non linéaire. On peut lire notamment [78], [111], [110], [143], [59], [3], [60], [37], [66], [164], [4], [61], [88], [70], [38], [79], [173].

L'inconvénient important de l'application de cette technique en reconnaissance du locuteur est le coût important dû à l'ajout d'un nouveau locuteur dans la base de référence. Plusieurs tentatives ont vu le jour pour tenter de remédier à cela, parmi lesquelles on peut citer [40], [41], [39], [10].

On peut aussi utiliser les réseaux de neurones en les couplant à d'autres techniques, comme par exemple les modèles de Markov cachés [109]. On parle alors de méthodes hybrides.

2.9 Les approches analytiques

Enfin, pour terminer ce tour d'horizon, il existe une famille de méthodes reposant sur une extraction préalable de certaines caractéristiques du signal de parole. On va choisir par exemple de caractériser les locuteurs sur des segments de paroles possédant une énergie importante entre 1000 Hz et 2000 Hz, ou bien ayant une énergie importante dans les hautes fréquences. On peut toujours parler de méthodes automatiques, à condition que ces techniques de localisations soient elles aussi entièrement automatiques.

Nous avons déjà dit un mot de ces méthodes dans le chapitre 1. Nous renvoyons ègalement à [20], [21], [19], [91].

2.10 Où en est la reconnaissance du locuteur aujourd'hui

De nombreux travaux s'intéressent actuellement aux techniques de reconnaissance du locuteur permettant une meilleure robustesse aux conditions dégradées. La parole téléphonique fait notamment l'objet de nombreuses études en reconnaissance du locuteur. On peut lire [112], [113], [7], [108], [125], [114], [93], [116].

La paramétrisation et l'extraction de caractéristiques fortement dépendantes du locuteur sont également des sujets très abordés : [126], [63], [78], [9], [168], [160], [161], [164], [64], [67], [68], [75], [79], [89], [11]. En particulier, on recherche des techniques qui permettent de prendre en compte les caractéristiques dynamiques du locuteur. Nous revenons en détail sur cet aspect-là au cours de notre travail de thèse.

2.11 Quelques remarques sur cette étude bibliographique

Nous avons choisi délibérément de mélanger dans cette étude bibliographique des travaux traitant ou bien de l'identification du locuteur, ou bien de la vérification du locuteur, car bien que les applications soient différentes, les méthodes utilisées sont les mêmes, et des progrès apportés dans l'une ou l'autre bénéficient généralement aux deux.

Nous avons également mélangé certains travaux sur des expériences dépendantes du texte aux nombreux autres travaux sur les expériences indépendantes du texte, car les méthodes sont également conjointes aux deux types d'expérimentations, exception faite pour la programmation dynamique, qui ne peut être utilisée qu'en mode dépendant du texte.

Maintenant que nous avons présenté quelques repères bibliographiques sur la reconnaissance automatique du locuteur, nous nous proposons d'entrer définitivement dans le vif du sujet de notre travail de thèse, en détaillant en premier lieu nos différents protocoles expérimentaux, puis les différentes méthodes que nous avons étudiées, et en proposant finalement un formalisme pour le filtrage vectoriel de trajectoires spectrales, ainsi qu'une application de ce filtrage en reconnaissance du locuteur.

Chapitre 3

Conditions expérimentales

Avant de décrire en détails toutes les techniques que nous avons étudiées, nous avons pris le parti de rassembler dans un même chapitre tout ce qui concernait les conditions expérimentales. En effet, elles sont communes, à quelques variantes près, à toutes nos expériences. Nous avons donc choisi de les placer en tête de notre travail, afin de pouvoir ensuite y référer très simplement.

3.1 Bases de données

Nous commençons par décrire les bases de données sur lesquelles nous avons travaillé, après avoir énuméré les différentes caractéristiques d'une base de données en général. Cette partie se termine par un regard quelque peu critique sur les bases utilisées.

3.1.1 Caractéristiques d'une base de données

Une base de données se caractérise par différents facteurs, qui vont de la qualité de la parole qu'elle contient, à la prise en compte ou non de la dérive temporelle, en passant par le degré de coopération des locuteurs qui la constituent.

3.1.1.a Qualité de la parole

Conditions d'enregistrement Plusieurs facteurs interviennent dans les conditions d'enregistrement : l'acoustique de la salle, le bruit ambiant, le type de micro choisi, les caractéristiques de la ligne de transmission (téléphonique par exemple). Il est donc important de bien décrire tous ces aspects.

Type de parole Une base de données peut également contenir différents types de parole, qui peut être spontanée, si le locuteur dit ce qui lui passe par la tête, ou bien lue, si le texte est écrit, ou affiché sur un écran.

3.1.1.b Coopération des locuteurs

La coopération des locuteurs est aussi un important facteur. La plupart des bases de données qui sont utilisées en reconnaissance du locuteur comportent des locuteurs dits "coopératifs", c'est-à-dire qui n'introduisent volontairement aucune modification de leur voix naturelle, ce qui n'empêche pas des modifications involontaires telles que celles qui sont dues par exemple à un rhume. Ce type de base n'est toutefois pas très réaliste car, dans une application réelle, la coopération d'un locuteur n'est jamais innocente. Ou bien il est "très coopératif", car il tient vraiment à se faire reconnaître, ce qui influera nécessairement sur sa façon de parler, ou bien il est non-coopératif, et ne tient absolument pas à se faire reconnaître, voire à se faire passer pour quelqu'un d'autre (ce qui est le cas notamment dans le domaine judiciaire).

3.1.1.c Prise en compte de la dérive temporelle

Nous savons, d'après plusieurs études [44, 6], que les performances en RAL diminuent lorsque la durée qui sépare une session de test et la session d'apprentissage augmente. C'est ce qu'on appelle la **dérive temporelle de la voix**. Il existe actuellement peu de bases de données qui prennent en compte ce facteur, alors que c'est également un facteur très important en termes de performances, et qu'il est à prendre en compte dans la plupart des applications de la reconnaissance du locuteur.

Pour illustrer le rôle joué par ce facteur sur les dégradations des performances d'un système d'identification du locuteur, nous reproduisons ici le résultat d'une expérience menée lors d'un workshop sur la reconnaissance du locuteur. Il a eu lieu en Juin 1994, à Martigny, en Suisse. Les participants ont pu pendant toute la semaine tester une démonstration d'un système d'identification du locuteur, dont le principe reposait sur une mesure de sphéricité (cf. chapitre 4). Environ 80 personnes ont fait un apprentissage de 15 secondes le premier jour du workshop, à travers un micro de bonne qualité, relié directement à une station SUN Sparc 10. Le milieu était plutôt bruité puisque de nombreuses personnes discutaient en permanence dans la même salle. Le signal était échantillonné à 16 kHz, et codé sur 16 bits. L'analyse acoustique était la même que celle présentée dans la section 3.2. Sur 124 tests

de locuteurs coopératifs effectués le même jour que celui de l'apprentissage, nous avons obtenu un taux d'erreur d'identification de 22.6 %. Sur 37 tests réalisés le lendemain du jour d'apprentissage, nous avons obtenus un taux d'erreur de 40.5 %. Sur 29 tests réalisés le surlendemain du jour d'apprentissage, nous avons obtenus un taux d'erreur de 44.8 %. Et enfin, sur 19 tests réalisés 3 jours après celui d'apprentissage, le taux d'erreur est monté à 57.9 %. Le nombre décroissant de tests de jour en jour ne permet pas de tirer de conclusions définitives, mais ces résultats illustrent néanmoins une chute importante des performances lorsque l'écart entre l'apprentissage et le test augmente.

Nous voyons donc le rôle crucial joué par ce facteur dans la dégradation des performances. La plupart des applications potentielles de la RAL sont rendues impossible essentiellement à cause de ce problème. Une solution réaliste permet toutefois de s'en sortir en vérification du locuteur. Après avoir constitué une première référence au cours d'une session d'apprentissage, nous pouvons adapter cette référence lors de chaque test fructueux, i.e. à chaque fois qu'un locuteur a été accepté par le système, son échantillon de parole est aussitôt utilisé pour mettre à jour sa référence [140].

Le problème de la constitution, et surtout de l'utilisation, d'une base de données prenant en compte ce facteur se pose également. Il n'est pas difficile de constituer une base multi-sessions. Encore faut-il définir un protocole expérimental permettant cette adaptation, et surtout l'évaluation des gains de performances qu'elle apporte.

3.1.1.d Problème de l'imposture

Un autre problème qui se pose souvent en reconnaissance du locuteur est la définition des imposteurs. Ce problème est directement lié à la constitution des bases de données ou à l'utilisation de celles qui existent.

En effet, à l'heure actuelle, on utilise généralement une partie des bases de données pour les clients d'un système, et l'autre partie pour les imposteurs. Cependant, cette façon de procéder ne prend absolument pas en compte la capacité qu'ont certains locuteurs de modifier leur voix, et d'augmenter ainsi leurs chances de tromper la machine. En particulier, le problème de l'imitation reste posé.

A notre connaissance, l'un des seuls travaux dans ce domaine est celui

présenté dans Rosenberg 1973 [136], et il concerne la reconnaissance du locuteur par auditeurs.

3.1.2 Bases de données utilisées

Après avoir décrit les différents aspects d'une base de données, nous présentons maintenant les bases que nous avons utilisées.

3.1.2.a TIMIT, NTIMIT, et FTIMIT

La base TIMIT [42, 48, 150, 82] contient 630 locuteurs (192 femmes et 438 hommes), ayant prononcé 10 phrases différentes chacun. Deux de ces phrases ("sa1" et "sa2") sont les mêmes pour tous les locuteurs. Les huit autres phrases (préfixes "si" et "sx") sont différentes d'un locuteur à l'autre[1]. Les phrases "sa" et "si" ont une durée moyenne de 2.9 secondes, les phrases "sx" ont une durée moyenne de 3.2 secondes. La parole est enregistrée à travers un microphone de très haute qualité, dans un environnement très calme, avec une bande passante 0-8 kHz [42]. Le signal est échantillonné à 16 kHz, codé sur 16 bits, avec une échelle d'amplitude linéaire. Les 10 phrases d'un locuteur sont enregistrées au cours de la même session.

La base FTIMIT, qui est une base que nous avons dérivée de TIMIT, est simplement une version filtrée de TIMIT. Nous avons simulé, au niveau de l'analyse acoustique, un filtrage passe-bas proche du filtrage téléphonique. Nous décrivons en détail la procédure utilisée pour simuler ce filtrage dans la section 3.2. Cette base FTIMIT constitue un intermédiaire entre TIMIT (parole de bonne qualité) et NTIMIT (parole de qualité téléphonique).

Enfin, la base NTIMIT [73] est obtenue à partir de la base TIMIT, en prenant chacune des dix phrases d'un locuteur donné, et en la faisant passer artificiellement par une ligne téléphonique différente. En plus de la réduction de la bande passante, nous avons de la distortion. Le signal est toujours échantillonné à 16 kHz, mais sa partie utile se limite cette fois-ci à la bande 0-4 kHz (la bande passante téléphonique est en fait approximativement égale à 300-3400 Hz).

[1] Il existe en fait un nombre limité de phrases. Mais ce nombre est suffisamment grand pour qu'il y ait au plus une seule phrase commune entre deux locuteurs donnés. Le biais introduit sur l'aspect "indépendant du texte" est donc très minime, à condition bien entendu de ne pas utiliser les phrases "sa" en apprentissage.

3.1.2.b TIMIT63, FTIMIT63 et NTIMIT63

Nous faisons en outre référence aux bases de données TIMIT63, FTIMIT63 et NTIMIT63, qui sont les bases TIMIT, FTIMIT et NTIMIT pour lesquelles nous nous sommes limités à 63 locuteurs (19 femmes et 44 hommes pour respecter la proportion des bases initiales) : nous avons conservé toutes les femmes et tous les hommes de "/train/dr1" et "/test/dr1", ainsi que la première locutrice et les 13 premiers locuteurs de "/train/dr2".

3.1.3 Critiques sur la pertinence des bases de données

Les bases TIMIT, FTIMIT et NTIMIT présentent quelques limites qu'il est bon de rappeler. Tout d'abord, il n'y a que 10 phrases par locuteur, ce qui limite les tests après l'apprentissage. D'autre part, toutes les phrases d'un même locuteur sont issues de la même session, il n'y a donc aucune prise en compte de la dérive temporelle. La parole initiale est de très bonne qualité et enregistrée dans un environnement très calme, ce qui est finalement peu réaliste pour une application réelle, excepté peut-être pour les applications sur sites, dont certaines d'entre elles permettent un environnement très calme (accès sécurisé se faisant par un sas). Enfin, il est bon de noter que le micro utilisé pour la constitution de NTIMIT est le même pour tous les enregistrements, seules varient les lignes téléphoniques par lesquelles passent les différentes phrases.

Néanmoins, ces bases permettent une première évaluation des systèmes, et une comparaison plus rigoureuse. Elles sont en outre largement utilisées par de nombreux laboratoires [127], [38], [90], [75].

Quelques nouvelles bases commencent également à être massivement utilisées par les laboratoires :

- Switchboard : [52, 149, 131, 148].
- Yoho : [24], [88], [25], [30], [151].
- Spidre, Polycost, Sesp, Polyvar, ...

Mais la plupart de ces bases nécessitent généralement une mise en forme assez importante.

3.2 Analyse acoustique

Passons maintenant à la description de l'analyse acoustique appliquée aux différentes phrases.

Chaque phrase des différentes bases de données est analysée de la façon décrite ci-après. Le signal est décomposé en trames de 504 échantillons (31,5 ms) avec un décalage de 160 échantillons (10 ms). Puis nous appliquons une fenêtre de Hamming à chaque trame. Le signal n'est pas pré-accentué.

Pour chaque trame, une transformée de Fourier de Winograd[2] est calculée, et elle fournit 252 valeurs représentant la première moitié des valeurs du module de la densité spectrale de puissance à court-terme dans la bande passante 0-8 kHz. La longueur de la fenêtre (504 échantillons) a été choisie de telle sorte que ce soit un nombre de Winograd, c'est-à-dire un nombre de la forme $N = 2^p\, 3^q\, 5^r\, 7^s$, avec $0 \leq p \leq 4$, $0 \leq q \leq 2$, $0 \leq r \leq 1$ et $0 \leq s \leq 1$. Ici, $N = 2^3\, 3^2\, 5^0\, 7^1$. D'autres largeurs de fenêtres ont également été testées. Les résultats de ces tentatives sont relatés dans l'annexe A.

Ce spectre est alors utilisé pour calculer 24 coefficients de banc de filtres. Tous les filtres sont triangulaires (à l'exception du premier et du dernier, qui sont trapézoïdaux). Ils sont placés sur une échelle de fréquence non-uniforme, similaire à une échelle Bark/Mel. L'échelle linéaire a également été testée (cf. annexe A). Les fréquences centrales des 24 filtres sont, en Hz : 47, 147, 257, 378, 510, 655, 813, 987, 1178, 1386, 1615, 1866, 2141, 2442, 2772, 3133, 3529, 3964, 4440, 4961, 5533, 6159, 6845, et 7597. Chaque filtre couvre un intervalle fréquentiel qui part de la fréquence centrale du filtre précédent et va jusqu'à la fréquence centrale du filtre suivant, avec une valeur maximale de 1 pour sa propre fréquence centrale. Remarquons pour finir que, pour chaque fréquence, il y a au maximum 2 filtres avec des valeurs non nulles, et que la somme de ces valeurs vaut toujours 1 (cf. Fig. 3.1).

Nous prenons finalement le logarithme en base 10 de chaque coefficient, que nous multiplions par 10, pour former un vecteur de dimension 24 comportant des valeurs en dB. L'ensemble de cette analyse acoustique est représentée schématiquement Fig. 3.2.

Pour la base TIMIT (et TIMIT63), nous gardons la totalité des 24 co-

[2] Il s'agit d'une transformée de Fourier rapide particulière.

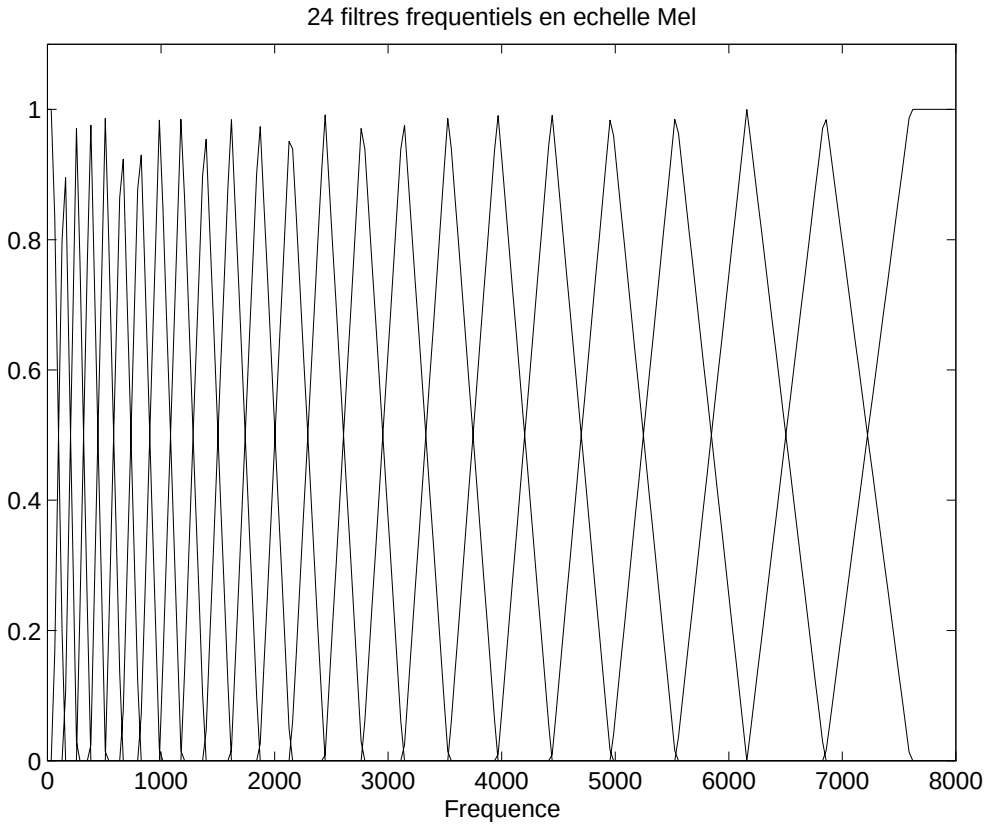


FIG. 3.1: *Banc de filtres appliqué aux 252 coefficients issus de la transformée de Winograd. Les filtres sont triangulaires, repartis sur une échelle similaire à l'échelle Bark/Mel. Pour chaque fréquence, la somme des valeurs vaut 1.*

efficients ainsi calculés. Pour la base NTIMIT (et NTIMIT63), puisque l'information utile se trouve dans la bande passante téléphonique, nous n'avons gardé que les 17 premiers coefficients sur les 24 calculés, ce qui correspond approximativement à la bande 0-3600 kHz, qui est la meilleure approximation qu'on puisse faire de la bande passante téléphonique (330Hz-3400Hz) avec ces coefficients. En effet, le 17ème filtre a une fréquence centrale de 3529 Hz, et devient nul à la fréquence 3964 Hz. Finalement, la base FTIMIT (et FTIMIT63) s'obtient à partir de TIMIT, en ne prenant que les 17 premiers coefficients des 24 calculés sur TIMIT. Ceci permet de simuler un filtrage passe-bas proche de celui du téléphone, mais sans avoir la distortion. On pourra ainsi mesurer les dégradations qui sont dues uniquement au fil-

trage de type téléphonique, et celles qui sont dues à la fois au filtrage, à la distortion, et à la variabilité du canal de transmission.

3.3 Choix des vecteurs de paramètres

Nous voudrions maintenant revenir un peu sur le choix des vecteurs de paramètres.

Plusieurs auteurs se sont intéressés à ce choix de vecteurs de paramètres : [167], [53], [145], [5], [94], [26], ou encore [95]. La plupart s'accordent à penser qu'on trouve une grande partie de l'information sur le locuteur dans ses caractéristiques spectrales à court-terme (coefficients de banc de filtres ou cepstre) [45]. Certains utilisent en outre les variations dynamiques de ces caractéristiques spectrales [46, 43, 78, 171]. Enfin, on peut aussi rajouter comme composantes supplémentaires le pitch ou l'énergie. Ces facteurs sont généralement caractéristiques du locuteur, mais facilement modifiables par celui-ci, même involontairement.

Nous avons choisi de travailler sur des vecteurs spectraux, puisqu'il apparaît qu'une grande partie de l'information sur le locuteur s'y trouve. Nous le faisons par une approche de type coefficients de banc de filtres puisque cette approche est plus simple et plus rapide que l'approche cepstrale (elle requiert une transformation de Fourier ou une transformation en cosinus en moins). Nous montrons d'ailleurs ultérieurement que les approches de base que nous utilisons sont invariantes par transformations linéaires inversibles (cf. annexe D). Cela revient à dire que, pour les approches que nous utilisons, les coefficients spectraux ou les coefficients cepstraux donnent les mêmes résultats, à condition néanmoins que nous gardions le même nombre de coefficients à chaque fois, c'est-à-dire que la transformation entre les deux soit inversible.

Nous n'avons pas choisi de composantes supplémentaires de type pitch ou énergie, car ces dernières ne sont pas tout à fait de même nature. Nous préférons utiliser des vecteurs dont les coordonnées sont toutes homogènes. Intégrer de l'information sur le pitch ou l'énergie nous paraît plus du ressort de la fusion de données. On pourrait néanmoins le faire par le biais d'une normalisation des vecteurs spectraux, ce que l'on fait parfois lorsqu'on utilise ensuite la modélisation Markovienne. D'autre part, en ce qui concerne l'énergie, il n'est pas évident qu'elle apporte de l'information supplémentaire

sur le locuteur. Enfin, le pitch n'est pas une grandeur facile à extraire, surtout dans le cas de parole téléphonique.

Pour finir, notons que nous avons choisi de ne pas retirer les silences de début et de fin de phrases, car cela introduirait dans notre démarche un élément plus difficilement reproductible. Toutefois, il est fort probable que l'introduction d'un outil de cette nature améliorerait significativement les résultats (cf. [128]).

3.4 Protocoles expérimentaux

Nous terminons ce chapitre sur nos conditions expérimentales par la description des protocoles expérimentaux.

Pour chaque locuteur, nous avons 10 phrases. Les 2 premières phrases étant les mêmes pour tous les locuteurs, si nous voulons travailler rigoureusement en mode indépendant du texte, nous ne pouvons pas les utiliser en apprentissage. Nous choisissons en conséquence 2 apprentissages différents :

- **apprentissage court (2)** : nous utilisons les 2 premières phrases "sx" pour l'apprentissage. La durée totale moyenne est alors de **5,7 secondes**, incluant les éventuels silences de début et de fin de phrases.
- **apprentissage long (5)** : les 5 phrases "sx" sont utilisées, ce qui représente une durée totale moyenne de **14,4 secondes**.

Nous optons également pour deux types de tests :

- **test court (1)** : nous utilisons, une par une, les phrases "sa" et "si", ce qui fait 5 tests par locuteur. La durée moyenne d'un test, qui dans ce cas est celle d'une phrase, est de **3,2 secondes**.
- **test long (5)** : les 2 phrases "sa" et les 3 phrases "si" sont utilisées en un seul test. La durée totale moyenne d'un test est alors de **15,9 secondes**.

Nous pouvons alors combiner apprentissages et tests de façons différentes. Dans la suite, nous référerons aux protocoles expérimentaux de la manière suivante : le **protocole 2.1** par exemple correspond à l'apprentissage court (2 phrases) combiné au test court (1 phrase) tandis que le **protocole 5.5** correspond à l'apprentissage long (5 phrases) et au test long (5 phrases).

Maintenant que nous avons décrit toutes nos conditions expérimentales, nous pouvons aborder la description des différentes méthodes étudiées, ainsi que la présentation du filtrage vectoriel de trajectoires spectrales. Tout ceci fait l'objet des chapitres suivants.

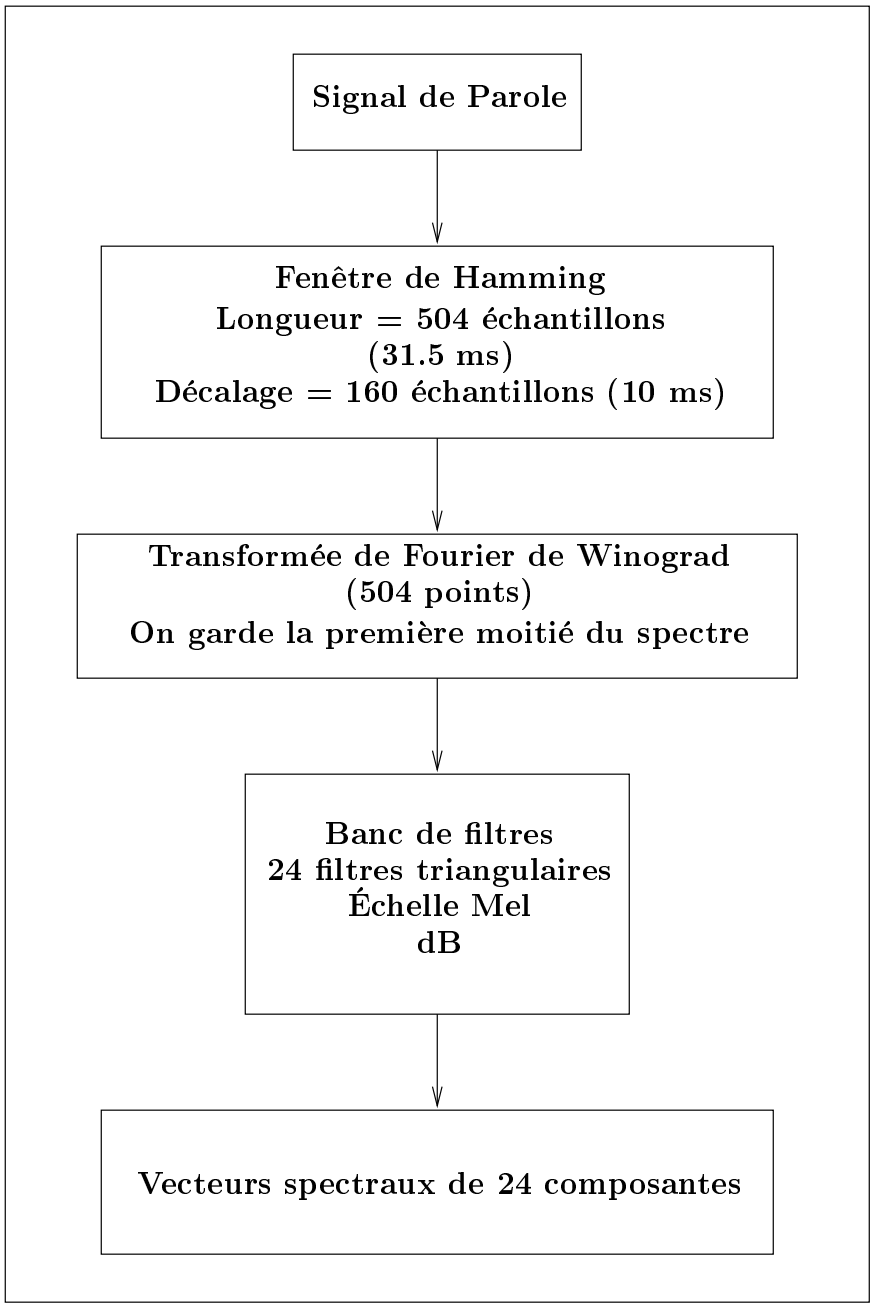


FIG. 3.2: *Schéma-bloc de l'analyse acoustique appliquée au signal de parole.*

Chapitre 4

Méthodes statistiques du second ordre

Après avoir passé en revue les différentes conditions expérimentales, nous décrivons dans les chapitres à venir différentes méthodes que nous avons utilisées pour faire de l'identification du locuteur indépendante du texte.

Ce chapitre présente du point de vue théorique une famille de mesures de similarité entre locuteurs. Ces mesures reposent sur les caractéristiques du second ordre d'une séquence de vecteurs, c'est-à-dire sur le vecteur moyen et la matrice de covariance de cette séquence. Une procédure de symétrisation de ces mesures, pour l'instant empirique, est également abordée. Enfin, l'influence de la durée et du contenu phonétique sur les performances de certaines de ces mesures sont étudiées[1].

4.1 Motivations

De nombreuses expériences en identification du locuteur ont été menées récemment sur la base TIMIT [143], [8], [106], [59], [129], [38], [90] et sur la base NTIMIT [74], [83], [127], [75]. Mais en dépit du fait que toutes ces expériences sont faites à partir des mêmes bases de données, les résultats sont difficilement comparables. Parmi les facteurs de variabilité, on trouve le prétraitement du signal (suppression des silences, pré-accentuation, longueur de la fenêtre, type de fenêtre, longueur du décalage, ...), le type d'analyse acoustique (MFCC, LPCC, Coefficients de banc de filtres, ...), les longueurs des

[1] Ce chapitre reprend en grande partie le contenu d'un article publié dans la revue *Speech Communication* [15], à l'exception de la section 4.9 sur l'influence de la durée et du contenu phonétique qui, elle, reproduit l'essentiel d'un article de conférence à *EUROSPEECH 95* [91]. Une copie de ces deux articles se trouve en annexe H de ce document.

apprentissages et des tests, le nombre de locuteurs sur lequel les résultats sont obtenus, ... Ainsi, une comparaison systématique de toute nouvelle approche avec les autres approches, sous le même protocole expérimental, est théoriquement possible mais pratiquement irréalisable. Ceci est dû aux différences mentionnées précédemment, mais également au fait qu'il est extrêmement difficile de reproduire en détail un algorithme particulier, parce que toutes les informations ne sont pas forcément publiquement accessibles, ou bien parce que l'algorithme peut être très sensible aux conditions initiales, ...

Une façon possible de régler ce problème d'évaluation est l'utilisation d'un algorithme de référence sous un protocole donné. On peut alors évaluer la complexité d'une nouvelle base de données étant donnée une tâche particulière, ou bien mesurer les bénéfices d'un nouveau modèle de locuteur associé à une nouvelle mesure de similarité. Un tel système de référence doit être relativement efficace et robuste, mais surtout présenter une implantation facile et une absolue reproductibilité. C'est dans ce sens que nous proposons cette étude sur les méthodes statistiques du second ordre. C'est aussi le moyen de constituer pour nos expériences une méthode de référence à laquelle nous pourrons comparer toute nouvelle approche.

Une autre motivation importante pour l'utilisation des méthodes statistiques du second-ordre est l'hypothèse selon laquelle l'information sur le locuteur se trouve essentiellement dans les corrélations entre différentes bandes de fréquence, i.e. essentiellement sur les coefficients non-diagonaux de la matrice de covariance [87], [86].

4.2 Notations, définitions, propriétés

4.2.1 Un modèle mono-Gaussien par locuteur

Considérons $\{\mathbf{x}_t\}_{1\leq t\leq M}$ une séquence de M vecteurs résultant de l'analyse acoustique p-dimensionnelle d'un signal de parole prononcé par un locuteur $\mathcal{X}$ (coefficients de banc de filtres, coefficients de prédiction linéaire, coefficients cepstraux, ...). Si on suppose que ces vecteurs suivent une distribution Gaussienne, ils peuvent être représentés plus succintement par leur moyenne $\bar{\mathbf{x}}$ et par leur matrice de covariance $\boldsymbol{\mathcal{X}}_0$:

$$\bar{\mathbf{x}} = \frac{1}{M}\sum_{t=1}^{M}\mathbf{x}_t \quad \text{et} \quad \boldsymbol{\mathcal{X}}_0 = \frac{1}{M}\sum_{t=1}^{M}(\mathbf{x}_t - \bar{\mathbf{x}})\cdot(\mathbf{x}_t - \bar{\mathbf{x}})^T \tag{4.1}$$

En fait, il s'agit ici d'estimateurs de cette moyenne et de cette matrice de covariance. On peut d'ailleurs représenter aussi ces vecteurs par leur moyenne et leur matrice de covariance, même si leur distribution n'est pas Gaussienne. On sait simplement que plus leur distribution est Gaussienne, moins il manque d'information dans leur vecteur moyen et leur matrice de covariance.

De même, à partir d'un signal de parole prononcé par un locuteur $\mathcal{Y}$, on peut extraire une séquence $\{\mathbf{y}_t\}_{1\leq t\leq N}$ de N vecteurs de dimension p, lesquels peuvent être représentés par leur moyenne $\bar{\mathbf{y}}$ et par leur matrice de covariance $\boldsymbol{\mathcal{Y}}_0$:

$$\bar{\mathbf{y}} = \frac{1}{N}\sum_{t=1}^{N}\mathbf{y}_t \quad \text{et} \quad \boldsymbol{\mathcal{Y}}_0 = \frac{1}{N}\sum_{t=1}^{N}(\mathbf{y}_t - \bar{\mathbf{y}})\cdot(\mathbf{y}_t - \bar{\mathbf{y}})^T \tag{4.2}$$

Les vecteurs $\bar{\mathbf{x}}$ et $\bar{\mathbf{y}}$ sont des vecteurs de dimension p. Les matrices $\boldsymbol{\mathcal{X}}_0$ et $\boldsymbol{\mathcal{Y}}_0$ sont des matrices symétriques définies positives de dimension $p \times p$.

Tout au long de ce chapitre, un locuteur $\mathcal{X}$ (respectivement $\mathcal{Y}$) sera représenté par le triplet $\{\bar{\mathbf{x}},\ \boldsymbol{\mathcal{X}}_0,\ M\}$ (respectivement $\{\bar{\mathbf{y}},\ \boldsymbol{\mathcal{Y}}_0,\ N\}$). Nous aurons aussi besoin de la matrice $\boldsymbol{\Gamma}_0$ définie par :

$$\boldsymbol{\Gamma}_0 = \boldsymbol{\mathcal{X}}_0^{-\frac{1}{2}}\ \boldsymbol{\mathcal{Y}}_0\ \boldsymbol{\mathcal{X}}_0^{-\frac{1}{2}} \tag{4.3}$$

où $\boldsymbol{\mathcal{X}}_0^{\frac{1}{2}}$ est la racine carrée symétrique de la matrice $\boldsymbol{\mathcal{X}}_0$. $\boldsymbol{\Gamma}_0$ est en fait une matrice ayant les mêmes valeurs propres que la matrice $\boldsymbol{\mathcal{Y}}_0\ \boldsymbol{\mathcal{X}}_0^{-1}$, mais qui est en outre symétrique, ce qui n'est généralement pas le cas de $\boldsymbol{\mathcal{Y}}_0\ \boldsymbol{\mathcal{X}}_0^{-1}$.

Remarquons pour finir que la matrice $\boldsymbol{\Gamma}_0$ est définie positive par définition, car $\boldsymbol{\mathcal{X}}_0^{-\frac{1}{2}}$ est symétrique et $\boldsymbol{\mathcal{Y}}_0$ est définie positive. En effet, $\forall \mathbf{x}$, on a :

$$\mathbf{x}^T\boldsymbol{\mathcal{X}}_0^{-\frac{1}{2}}\ \boldsymbol{\mathcal{Y}}_0\ \boldsymbol{\mathcal{X}}_0^{-\frac{1}{2}}\ \mathbf{x} = \left(\boldsymbol{\mathcal{X}}_0^{-\frac{1}{2}}\ \mathbf{x}\right)^T\ \boldsymbol{\mathcal{Y}}_0\ \left(\boldsymbol{\mathcal{X}}_0^{-\frac{1}{2}}\ \mathbf{x}\right) > 0$$

4.2.2 Une famille de mesures de similarité

Dans la suite de ce chapitre, nous nous intéressons à une famille particulière de mesures de similarité entre les locuteurs $\mathcal{X}$ et $\mathcal{Y}$:

$$\mu\ (\mathcal{X},\ \mathcal{Y}) = \phi\ (\bar{\mathbf{x}},\ \boldsymbol{\mathcal{X}}_0,\ M,\ \bar{\mathbf{y}},\ \boldsymbol{\mathcal{Y}}_0,\ N) \tag{4.4}$$

La fonction ϕ est choisie la plupart du temps en dérivant des tests d'hypothèses statistiques. Elle est en outre généralement choisie de sorte que la mesure μ soit définie positive :

$$\forall\, \mathcal{X},\, \forall\, \mathcal{Y}, \quad \mu\,(\mathcal{X},\, \mathcal{Y}) \;\geq\; 0 \tag{4.5}$$

$$\forall\, \mathcal{X}, \quad \mu\,(\mathcal{X},\, \mathcal{X}) \;=\; 0 \tag{4.6}$$

Dans leur forme de base, les mesures μ étudiées sont rarement symétriques, mais nous proposerons une façon systématique d'obtenir des versions symétriques de ces mesures, c'est-à-dire telles que :

$$\forall\, \mathcal{X},\, \forall\, \mathcal{Y}, \quad \mu\,(\mathcal{X},\, \mathcal{Y}) \;=\; \mu\,(\mathcal{Y},\, \mathcal{X}) \tag{4.7}$$

4.2.3 Valeurs propres de la matrice $\mathbf{\Gamma}_0$

Notons $\{\lambda_i\}_{1\leq i\leq p}$ les valeurs propres de la matrice $\mathbf{\Gamma}_0$, i.e. les racines de l'équation :

$$det\,[\mathbf{\Gamma}_0 \;-\; \lambda \mathbf{I}] = 0 \tag{4.8}$$

Elles sont strictement positives puisque la matrice $\mathbf{\Gamma}_0$ est définie positive.

Ce sont aussi les valeurs propres de la matrice $\boldsymbol{\mathcal{Y}}_0 \boldsymbol{\mathcal{X}}_0^{-1}$, et on les appelle valeurs propres de $\boldsymbol{\mathcal{Y}}_0$ relativement à $\boldsymbol{\mathcal{X}}_0$. Rappelons ici que la matrice $\mathbf{\Gamma}_0$ a été introduite car elle possède les mêmes valeurs propres que $\boldsymbol{\mathcal{Y}}_0 \boldsymbol{\mathcal{X}}_0^{-1}$, et qu'elle est en outre symétrique, ce qui n'est pas nécessairement le cas de $\boldsymbol{\mathcal{Y}}_0 \boldsymbol{\mathcal{X}}_0^{-1}$. La matrice $\mathbf{\Gamma}_0$ peut donc être décomposée de la manière suivante :

$$\mathbf{\Gamma}_0 \;=\; \mathbf{\Theta}_0\, \mathbf{\Lambda}_0\, \mathbf{\Theta}_0^T \;\text{ avec }\; \mathbf{\Theta}_0^T = \mathbf{\Theta}_0^{-1} \tag{4.9}$$

où $\mathbf{\Lambda}_0$ est la matrice diagonale de dimension $p \times p$ contenant les valeurs propres de $\mathbf{\Gamma}_0$, et $\mathbf{\Theta}_0$ la matrice orthogonale de dimension $p \times p$ dont les vecteurs colonnes sont les vecteurs propres associés aux valeurs propres correspondantes de $\mathbf{\Lambda}_0$.

Remarquons que les valeurs propres de $\boldsymbol{\mathcal{X}}_0$ relativement à $\boldsymbol{\mathcal{Y}}_0$ (i.e. les valeurs propres de $\mathbf{\Gamma}_0^{-1}$) sont les $\{1/\lambda_i\}_{1\leq i\leq p}$. [2]

[2] Notons que, contrairement à ce qui a été dit dans [15], l'interversion de $\boldsymbol{\mathcal{X}}_0$ et $\boldsymbol{\mathcal{Y}}_0$ ne transforme pas $\mathbf{\Gamma}_0$ en $\mathbf{\Gamma}_0^{-1}$, mais la propriété énoncée sur les valeurs propres suffit pour justifier le reste des développements théoriques.

4.2.4 Différentes fonctions de ces valeurs propres

Trois fonctions particulières des $\{\lambda_i\}$ sont utilisées :

$$\begin{aligned}
\text{la moyenne arithmétique :} \quad a\ (\lambda_1, ..., \lambda_p) &= \frac{1}{p} \sum_{i=1}^{p} \lambda_i \\
\text{la moyenne géométrique :} \quad g\ (\lambda_1, ..., \lambda_p) &= \left(\prod_{i=1}^{p} \lambda_i \right)^{1/p} \qquad (4.10) \\
\text{la moyenne harmonique :} \quad h\ (\lambda_1, ..., \lambda_p) &= \left(\frac{1}{p} \sum_{i=1}^{p} \frac{1}{\lambda_i} \right)^{-1}
\end{aligned}$$

Comme toutes les valeurs propres λ_i sont positives, on peut montrer que :

$$a \geq g \geq h \qquad (4.11)$$

avec égalité si et seulement si toutes les λ_i sont égales.

De plus, intervertir $\mathcal{X}$ et $\mathcal{Y}$ revient à remplacer a par $1/h$, g par $1/g$ et h par $1/a$. En d'autres termes,

$$a\ (\frac{1}{\lambda_1}, ..., \frac{1}{\lambda_p}) = \frac{1}{h\ (\lambda_1, ..., \lambda_p)} \qquad (4.12)$$

$$g\ (\frac{1}{\lambda_1}, ..., \frac{1}{\lambda_p}) = \frac{1}{g\ (\lambda_1, ..., \lambda_p)} \qquad (4.13)$$

$$h\ (\frac{1}{\lambda_1}, ..., \frac{1}{\lambda_p}) = \frac{1}{a\ (\lambda_1, ..., \lambda_p)} \qquad (4.14)$$

4.2.5 Calcul de a, g et h

La trace (notée tr) vérifie la propriété $tr(AB) = tr(BA)$ et le déterminant (noté det) $det\ (AB) = det\ A \cdot det\ B$. Nous avons alors les propriétés suivantes :

$$a\ (\lambda_1, ..., \lambda_p) = \frac{1}{p}\ tr\ \mathbf{\Lambda}_0 = \frac{1}{p}\ tr\ \mathbf{\Gamma}_0 = \frac{1}{p}\ tr\ (\boldsymbol{\mathcal{Y}}_0 \boldsymbol{\mathcal{X}}_0^{-1}) \qquad (4.15)$$

$$g\ (\lambda_1, ..., \lambda_p) = (det\ \mathbf{\Lambda}_0)^{1/p} = (det\ \mathbf{\Gamma}_0)^{1/p} = \left(\frac{det\ \boldsymbol{\mathcal{Y}}_0}{det\ \boldsymbol{\mathcal{X}}_0} \right)^{1/p} \qquad (4.16)$$

$$h\ (\lambda_1, ..., \lambda_p) = \frac{p}{tr\ (\mathbf{\Lambda}_0^{-1})} = \frac{p}{tr\ (\mathbf{\Gamma}_0^{-1})} = \frac{p}{tr\ (\boldsymbol{\mathcal{X}}_0 \boldsymbol{\mathcal{Y}}_0^{-1})} \qquad (4.17)$$

Ces équations montrent que a, g et h peuvent être obtenues directement à partir de $\boldsymbol{\mathcal{X}}_0$, $\boldsymbol{\mathcal{Y}}_0$, $\boldsymbol{\mathcal{X}}_0^{-1}$, $\boldsymbol{\mathcal{Y}}_0^{-1}$, $det\ \boldsymbol{\mathcal{X}}_0$ et $det\ \boldsymbol{\mathcal{Y}}_0$, sans extraire explicitement les valeurs propres λ_i, et sans calculer les racines carrées de $\boldsymbol{\mathcal{X}}_0$ et $\boldsymbol{\mathcal{Y}}_0$. De plus, $tr\ (\boldsymbol{\mathcal{Y}}_0\boldsymbol{\mathcal{X}}_0^{-1})$ et $tr\ (\boldsymbol{\mathcal{X}}_0\boldsymbol{\mathcal{Y}}_0^{-1})$ peuvent être calculées sans effectuer le produit matriciel complet, mais seulement pour les éléments diagonaux du produit.

4.3 Maximum de vraisemblance

4.3.1 Définition

En supposant que tous les vecteurs acoustiques extraits d'un signal de parole prononcé par le locuteur $\mathcal{X}$ sont distribués statistiquement selon une densité Gaussienne (hypothèse $\mathcal{H}_1$), la vraisemblance d'un vecteur acoustique $\mathbf{y}_t$ issu d'un signal de parole prononcé par le locuteur $\mathcal{Y}$ est classiquement :

$$G\ (\mathbf{y}_t \mid \mathcal{X})\ =\ \frac{1}{(2\pi)^{\frac{p}{2}}\ (det\ \boldsymbol{\mathcal{X}}_0)^{\frac{1}{2}}}\ e^{-\frac{1}{2}(\mathbf{y}_t-\bar{\mathbf{x}})^T\boldsymbol{\mathcal{X}}_0^{-1}(\mathbf{y}_t-\bar{\mathbf{x}})} \tag{4.18}$$

Si nous supposons en outre que tous les vecteurs $\mathbf{y}_t$ sont les observations indépendantes d'un même processus (hypothèse $\mathcal{H}_2$), la log-vraisemblance moyenne de $\{\mathbf{y}_t\}_{1\leq t\leq N}$ s'écrit :

$$\begin{aligned}\overline{G_{\mathcal{X}}}\ (y_1^N)\ &=\ \frac{1}{N}\ log\ G\ (y_1\ ...\ y_n \mid \mathcal{X})\ =\ \frac{1}{N}\sum_{t=1}^{N} log\ G\ (\mathbf{y}_t \mid \mathcal{X}) \\ &=\ -\frac{1}{2}\left[\ p\ log\ 2\pi\ +\ log\ (det\ \boldsymbol{\mathcal{X}}_0)\ +\ \frac{1}{N}\sum_{t=1}^{N}(\mathbf{y}_t-\bar{\mathbf{x}})^T\boldsymbol{\mathcal{X}}_0^{-1}(\mathbf{y}_t-\bar{\mathbf{x}})\ \right]\end{aligned} \tag{4.19}$$

En remplaçant alors $\mathbf{y}_t-\bar{\mathbf{x}}$ par $\mathbf{y}_t-\bar{\mathbf{y}}+\bar{\mathbf{y}}-\bar{\mathbf{x}}$ et en utilisant la propriété :

$$\frac{1}{N}\sum_{t=1}^{N}(\mathbf{y}_t-\bar{\mathbf{y}})^T\boldsymbol{\mathcal{X}}_0^{-1}(\mathbf{y}_t-\bar{\mathbf{y}})\ =\ tr\ (\boldsymbol{\mathcal{Y}}_0\boldsymbol{\mathcal{X}}_0^{-1}) \tag{4.20}$$

nous obtenons :

$$\begin{aligned}&\overline{G_{\mathcal{X}}}\ (y_1^N)\ +\ \frac{p}{2}\ log\ 2\pi\ = \\ &\quad -\frac{1}{2}\ \left[\ log\ (det\ \boldsymbol{\mathcal{X}}_0)\ +\ tr\ (\boldsymbol{\mathcal{Y}}_0\boldsymbol{\mathcal{X}}_0^{-1})\ +\ (\bar{\mathbf{y}}-\bar{\mathbf{x}})^T\boldsymbol{\mathcal{X}}_0^{-1}(\bar{\mathbf{y}}-\bar{\mathbf{x}})\ \right]\end{aligned} \tag{4.21}$$

et

$$\frac{2}{p}\,\overline{G_{\mathcal{X}}}\,(y_1^N) \;+\; log\,2\pi \;+\frac{1}{p}\,log\,(det\,\boldsymbol{\mathcal{Y}}_0) \;+\; 1 \qquad (4.22)$$
$$= \;\frac{1}{p}\left[\,log\left(\frac{det\,\boldsymbol{\mathcal{Y}}_0}{det\,\boldsymbol{\mathcal{X}}_0}\right) \;-\; tr\,(\boldsymbol{\mathcal{Y}}_0\boldsymbol{\mathcal{X}}_0^{-1}) \;-\; (\bar{\mathbf{y}}-\bar{\mathbf{x}})^T\boldsymbol{\mathcal{X}}_0^{-1}(\bar{\mathbf{y}}-\bar{\mathbf{x}})\right] \;+\; 1$$

Ainsi, nous définissons la mesure μ_G par :

$$\mu_G\,(\mathcal{X},\,\mathcal{Y}) \;=\; \frac{1}{p}\,tr\,(\boldsymbol{\mathcal{Y}}_0\boldsymbol{\mathcal{X}}_0^{-1}) \;-\frac{1}{p}\;log\,\left(\frac{det\,\boldsymbol{\mathcal{Y}}_0}{det\,\boldsymbol{\mathcal{X}}_0}\right) \qquad (4.23)$$
$$+\frac{1}{p}\,(\bar{\mathbf{y}}-\bar{\mathbf{x}})^T\boldsymbol{\mathcal{X}}_0^{-1}(\bar{\mathbf{y}}-\bar{\mathbf{x}}) \;-\; 1$$
$$= \;\frac{1}{p}\left[\;tr\,\boldsymbol{\Gamma}_0 \;-\; log\,(det\,\boldsymbol{\Gamma}_0) \;+\; \boldsymbol{\delta}^T\boldsymbol{\mathcal{X}}_0^{-1}\boldsymbol{\delta}\;\right] \;-\; 1 \qquad (4.24)$$

$$\boxed{\mu_G\,(\mathcal{X},\,\mathcal{Y}) = a \;-\; log\;g \;+\; \frac{1}{p}\,\boldsymbol{\delta}^T\boldsymbol{\mathcal{X}}_0^{-1}\boldsymbol{\delta} \;-\; 1} \qquad (4.25)$$

avec :

$$\boldsymbol{\delta} = \bar{\mathbf{y}} - \bar{\mathbf{x}} \qquad (4.26)$$

Nous avons enfin :

$$\underset{\mathcal{X}}{Argmax}\quad \overline{G_{\mathcal{X}}}\,(y_1^N) \;=\; \underset{\mathcal{X}}{Argmin}\quad \mu_G\,(\mathcal{X},\,\mathcal{Y}) \qquad (4.27)$$

4.3.2 Propriétés de μ_G

La matrice $\boldsymbol{\mathcal{X}}_0^{-1}$ étant définie positive comme $\boldsymbol{\mathcal{X}}_0$, on a $\boldsymbol{\delta}^T\boldsymbol{\mathcal{X}}_0^{-1}\boldsymbol{\delta} \geq 0$. De plus, nous avons $log\,g \leq g-1$ et $a \geq g$. Ainsi, $a - log\,g - 1 \geq 0$. D'où finalement, $\mu_G\,(\mathcal{X},\,\mathcal{Y}) \geq 0$.

D'autre part, $\mu_G\,(\mathcal{X},\,\mathcal{Y}) \;=\; 0$ si et seulement si toutes les valeurs propres λ_i sont égales à 1 et $\boldsymbol{\delta}$ est le vecteur nul, i.e. si et seulement si $\boldsymbol{\mathcal{X}}_0 = \boldsymbol{\mathcal{Y}}_0$ et $\bar{\mathbf{x}} = \bar{\mathbf{y}}$.

Enfin, $\mu_G\,(\mathcal{X},\,\mathcal{Y})$ n'est pas symétrique, et le terme dual correspondant est :

$$\mu_G\,(\mathcal{Y},\,\mathcal{X}) \;=\; \frac{1}{h} \;+\; log\,g \;+\frac{1}{p}\,\boldsymbol{\delta}^T\boldsymbol{\mathcal{Y}}_0^{-1}\boldsymbol{\delta} \;-\; 1 \;\neq\; \mu_G\,(\mathcal{X},\,\mathcal{Y}) \qquad (4.28)$$

4.3.3 Une variante de μ_G

Lorsque les données de parole utilisées sont bruitées ou que la distortion est importante, les vecteurs moyens $\bar{\mathbf{x}}$ et $\bar{\mathbf{y}}$ peuvent être fortement influencés par les caractéristiques du canal de transmission, alors que les matrices de covariance $\boldsymbol{\mathcal{X}}_0$ et $\boldsymbol{\mathcal{Y}}_0$ sont habituellement plus robustes aux variations des conditions d'enregistrement et des canaux de transmission [51], puisqu'on soustrait le vecteur moyen à chaque vecteur pour les calculer. Dans ce cas là, $\boldsymbol{\delta} = \bar{\mathbf{y}} - \bar{\mathbf{x}}$ peut être un facteur pénalisant dans la mesure μ_G.

Une mesure de vraisemblance Gaussienne sur les matrices de covariance seules s'obtient alors de la façon suivante :

$$\mu_{Gc}\,(\mathcal{X},\,\mathcal{Y}) \;=\; \frac{1}{p}\left[\; tr\,(\boldsymbol{\mathcal{Y}}_0\boldsymbol{\mathcal{X}}_0^{-1}) \;-\; log\,\left(\frac{det\,\boldsymbol{\mathcal{Y}}_0}{det\,\boldsymbol{\mathcal{X}}_0}\right)\right] \;-\; 1 \qquad (4.29)$$

$$=\; \frac{1}{p}\,[\; tr\,\boldsymbol{\Gamma}_0 \;-\; log\,(det\,\boldsymbol{\Gamma}_0)\,] \;-\; 1 \qquad (4.30)$$

$$\boxed{\mu_{Gc}\,(\mathcal{X},\,\mathcal{Y}) = a \;-\; log\;g \;-\; 1} \qquad (4.31)$$

Cette mesure peut s'exprimer en fonction des valeurs propres λ_i de la matrice $\boldsymbol{\Gamma}_0$. Cependant, il n'est pas obligatoire d'extraire explicitement ces valeurs propres.

Cette mesure a de plus les mêmes propriétés que la mesure μ_G. En particulier, cette mesure n'est toujours pas symétrique puisque :

$$\mu_{Gc}\,(\mathcal{Y},\,\mathcal{X}) \;=\; \frac{1}{h} \;+\; log\;g \;-\; 1 \;\neq\; \mu_{Gc}\,(\mathcal{X},\,\mathcal{Y}) \qquad (4.32)$$

4.4 Test de sphéricité

4.4.1 Définition

Comme cela a été présenté dans [1], nous pouvons construire un test de proportionalité entre deux matrices de covariances $\boldsymbol{\mathcal{Y}}_0$ et $\boldsymbol{\mathcal{X}}_0$ à l'aide d'une fonction de vraisemblance :

$$S\,(\boldsymbol{\mathcal{Y}}_0 \mid \boldsymbol{\mathcal{X}}_0) \;=\; \left[\frac{det\,(\boldsymbol{\mathcal{X}}_0^{-\frac{1}{2}}\boldsymbol{\mathcal{Y}}_0\boldsymbol{\mathcal{X}}_0^{-\frac{1}{2}})}{\left(\frac{1}{p}\;tr\,(\boldsymbol{\mathcal{X}}_0^{-\frac{1}{2}}\boldsymbol{\mathcal{Y}}_0\boldsymbol{\mathcal{X}}_0^{-\frac{1}{2}})\right)^p}\right]^{\frac{N}{2}} \;=\; \left[\frac{det\,\boldsymbol{\Gamma}_0}{\left(\frac{1}{p}\;tr\,\boldsymbol{\Gamma}_0\right)^p}\right]^{\frac{N}{2}} \qquad (4.33)$$

Cette expression peut être interprétée comme la combinaison de deux critères : l'un sur la diagonalité de la matrice $\mathbf{\Gamma}_0$, et l'autre sur l'égalité des éléments diagonaux de cette matrice, sachant que $\mathbf{\Gamma}_0$ est diagonale.

En notant $\overline{S_{\mathcal{X}}}\ (y_1^N)$ la fonction de vraisemblance moyenne pour le test de sphéricité :

$$\overline{S_{\mathcal{X}}}\ (y_1^N) \ = \ \frac{1}{N}\ log\ S\ (\boldsymbol{\mathcal{Y}}_0 \mid \boldsymbol{\mathcal{X}}_0) \tag{4.34}$$

Nous définissons alors une nouvelle mesure de similarité :

$$\mu_{Sc}\ (\mathcal{X},\ \mathcal{Y}) \ = \ log\ \left[\frac{\frac{1}{p}\ tr\ \mathbf{\Gamma}_0}{(det\ \mathbf{\Gamma}_0)^{1/p}}\right] \tag{4.35}$$

$$= \ log\ \left[\frac{\frac{1}{p}\ tr\ (\boldsymbol{\mathcal{Y}}_0 \boldsymbol{\mathcal{X}}_0^{-1})}{\left(\frac{det\ \boldsymbol{\mathcal{Y}}_0}{det\ \boldsymbol{\mathcal{X}}_0}\right)^{1/p}}\right] \tag{4.36}$$

$$\boxed{\mu_{Sc}\ (\mathcal{X},\ \mathcal{Y}) = log\ \left(\frac{a}{g}\right)} \tag{4.37}$$

Et nous avons :

$$\underset{\mathcal{X}}{Argmax}\ \ \overline{S_{\mathcal{X}}}\ (y_1^N) \ = \ \underset{\mathcal{X}}{Argmin}\ \ \mu_{Sc}\ (\mathcal{X},\ \mathcal{Y}) \tag{4.38}$$

La mesure μ_{Sc} est en fait le logarithme du rapport entre la moyenne arithmetique et la moyenne géométrique des valeurs propres de $\boldsymbol{\mathcal{Y}}_0$ relativement à $\boldsymbol{\mathcal{X}}_0$, d'où son nom de mesure de sphéricité arithmético-géométrique. Comme pour la mesure μ_{Gc}, la mesure μ_{Sc} constitue un test sur les matrices de covariances seules. Elle peut être exprimée comme une fonction des valeurs propres λ_i, mais ne requiert pas l'extraction explicite de ces valeurs propres.

L'utilisation d'un test de sphéricité arithmético-géométrique pour la reconnaissance du locuteur a été initialement proposée dans [55], dans le cadre d'expériences indépendantes du texte.

4.4.2 Propriétés de μ_{Sc}

Comme $a \geq g$, on a immédiatement $\mu_{Sc}(\mathcal{X}, \mathcal{Y}) \geq 0$.

D'autre part, $\mu_{Sc}(\mathcal{X}, \mathcal{Y}) = 0$ si et seulement si toutes les valeurs propres λ_i sont égales, i.e. si et seulement si $\boldsymbol{\mathcal{X}}_0$ et $\boldsymbol{\mathcal{Y}}_0$ sont proportionnelles. En particulier, $\mu_{Sc}(\mathcal{X}, \mathcal{X}) = 0$, mais $\boldsymbol{\mathcal{X}}_0 = \boldsymbol{\mathcal{Y}}_0$ n'est pas une condition nécessaire.

Finalement, μ_{Sc} n'est pas symétrique, et :

$$\mu_{Sc}(\mathcal{Y}, \mathcal{X}) = log\left(\frac{g}{h}\right) \neq \mu_{Sc}(\mathcal{X}, \mathcal{Y}) \tag{4.39}$$

4.4.3 Interprétation géométrique

La mesure de sphéricité permet en fait de mesurer à quel point les valeurs propres d'une matrice sont toutes égales, c'est-à-dire à quel point cette matrice est proportionnelle à l'identité. Or la représentation de la matrice identité dans un espace multidimensionnel est une sphère. La mesure de sphéricité est donc une façon de mesurer à quel point la représentation d'une matrice est sphérique, d'où son nom. Dans notre cas, nous nous intéressons à la sphéricité de la matrice $\boldsymbol{\mathcal{Y}}_0 \boldsymbol{\mathcal{X}}_0^{-1}$.

4.5 Déviation absolue des valeurs propres

4.5.1 Définition

Les expressions de μ_{Gc} et de μ_{Sc} en fonction des valeurs propres λ_i sont :

$$\mu_{Gc}(\mathcal{X}, \mathcal{Y}) = \frac{1}{p}\sum_{i=1}^{p}(\lambda_i - log\ \lambda_i - 1) \tag{4.40}$$

$$\mu_{Sc}(\mathcal{X}, \mathcal{Y}) = log\left(\frac{1}{p}\sum_{i=1}^{p}\lambda_i\right) - \frac{1}{p}\sum_{i=1}^{p} log\ \lambda_i \tag{4.41}$$

Sur le même modèle, on peut construire d'autres mesures de similarités entre locuteurs, à partir de leurs matrices de covariance, et plus particulièrement en fonction des valeurs propres de la matrice $\boldsymbol{\Gamma}_0$. Toute fonction des λ_i, qui est positive, et qui prend la valeur zéro lorsque toutes les valeurs propres sont égales à 1, est un choix possible.

Cette approche a été proposée dans [50], où l'on trouve une mesure reposant sur la déviation absolue des valeurs propres par rapport à 1. L'expression de cette mesure, notée μ_{Dc}, est :

$$\mu_{Dc}\,(\mathcal{X},\,\mathcal{Y}) \;=\; \frac{1}{p}\,\sum_{i=1}^{p}\;|\;\lambda_i\;-\;1\;| \qquad (4.42)$$

Dans cette formulation, μ_{Dc} est la déviation absolue moyenne des valeurs propres λ_i par rapport à l'unité. Gish montre que la robustesse de cette mesure peut être améliorée en retirant les plus grandes valeurs propres, car elles correspondent à des "*anomalies dans des espaces de petite dimension*".

4.5.2 Propriétés de μ_{Dc}

La mesure μ_{Dc} est une mesure positive par définition, et on vérifie facilement qu'elle est nulle si et seulement si les matrices de covariance $\boldsymbol{\mathcal{X}}_0$ et $\boldsymbol{\mathcal{Y}}_0$ sont égales. Cette mesure n'est pas symétrique car :

$$\mu_{Dc}\,(\mathcal{Y},\,\mathcal{X}) \;=\; \frac{1}{p}\,\sum_{i=1}^{p}\;|\;\frac{1}{\lambda_i}\;-\;1\;|\neq\;\mu_{Dc}\,(\mathcal{X},\,\mathcal{Y}) \qquad (4.43)$$

4.5.3 Autres mesures inspirées de μ_{Dc}

En suivant la même idée que dans [50], nous avons testé deux autres mesures qui s'expriment de façon similaire à μ_{Dc}. La première est obtenue à partir de μ_{Dc} en ne prenant pas en compte la plus grande valeur propre, ni la plus petite. Si nous notons μ_{D1c} cette nouvelle mesure, son expression est alors :

$$\mu_{D1c}\,(\mathcal{X},\,\mathcal{Y}) \;=\; \frac{1}{p-2}\,\sum_{i=2}^{p-1}\;|\;\lambda_i\;-\;1\;| \qquad (4.44)$$

La deuxième mesure, notée μ_{D2c} est la mesure complémentaire de la précédente, c'est-à-dire qu'on ne prend cette fois-ci que la plus grande et la plus petite valeur propre. Ainsi, nous avons :

$$\mu_{D2c}\,(\mathcal{X},\,\mathcal{Y}) \;=\; \frac{1}{2}\;(\;|\;\lambda_1\;-\;1\;|\;+\;|\;\lambda_p\;-\;1\;|\;) \qquad (4.45)$$

Nous nous sommes limités à ces deux variantes, mais on aurait pu imaginer d'autres combinaisons. Les résultats concernant ces deux mesures sont donnés en annexe B, afin de ne pas surcharger cette partie.

4.6 Symétrisation

4.6.1 Motivations

Toutes les mesures présentées précédemment ont en commun de ne pas être symétriques. En d'autres termes, les rôles joués par les données d'apprentissage et par les données de test ne sont pas interchangeables. Pourtant, si nous raisonnons par analogie avec les propriétés d'une distance, nous souhaiterions que cette propriété de symétrie soit vérifiée.

L'asymétrie des mesures μ_G, μ_{Gc} et μ_{Sc} peut s'expliquer par le fait qu'elles se fondent sur des tests statistiques qui supposent que le modèle du locuteur de référence $\mathcal{X}$ est exact, tandis que le modèle du locuteur de test $\mathcal{Y}$ est estimé. En pratique, les deux modèles sont estimés, bien que celui du locuteur de référence le soit sur un plus grand nombre de données, et soit donc plus proche du modèle exact. C'est dans le but de prendre en compte le fait que les deux modèles sont estimés que nous cherchons à symétriser les mesures.

De plus, on peut penser que la fiabilité d'un modèle de référence est liée au nombre de données utilisées pour son estimation. C'est d'ailleurs ce que nous observons sur les résultats expérimentaux, puisque $\mu\ (\mathcal{X},\ \mathcal{Y})$ et $\mu\ (\mathcal{Y},\ \mathcal{X})$ donnent des performances d'autant plus différentes que $\rho = N/M$ est différent de 1. C'est ce qui nous a donné l'idée de tester également des symétrisations dépendant du nombre de vecteurs utilisés pour l'estimation des différents modèles.

4.6.2 Procédures empiriques de symétrisation

Une première possibilité pour symétriser une mesure $\mu\ (\mathcal{X},\ \mathcal{Y})$ est de construire la moyenne entre cette mesure et son terme dual :

$$\mu_{[0.5]}\ (\mathcal{X},\ \mathcal{Y}) \ = \ \frac{1}{2}\,\mu\ (\mathcal{X},\ \mathcal{Y}) \ + \ \frac{1}{2}\,\mu\ (\mathcal{Y},\ \mathcal{X}) \ = \ \mu_{[0.5]}\ (\mathcal{Y},\ \mathcal{X}) \tag{4.46}$$

Par exemple, la mesure de vraisemblance Gaussienne, symétrisée de cette façon, devient :

$$\mu_{G_{[0.5]}}\ (\mathcal{X},\ \mathcal{Y}) \ = \ \frac{1}{2}\left[\ a \ + \ \frac{1}{h} \ + \ \frac{1}{p}\,\boldsymbol{\delta}^T\,(\boldsymbol{\mathcal{X}}_0^{-1} + \boldsymbol{\mathcal{Y}}_0^{-1})\,\boldsymbol{\delta}\ \right] \ - \ 1 \tag{4.47}$$

ce qui se simplifie, pour la mesure sur les covariances seules, en :

$$\mu_{Gc_{[0.5]}}\ (\mathcal{X},\ \mathcal{Y}) \ = \ \frac{1}{2}\left(\ a \ + \ \frac{1}{h}\right) \ - \ 1 \tag{4.48}$$

tandis que la mesure de sphéricité arithmético-géometrique symétrisée de la même façon devient proportionnelle à la mesure de sphéricité arithmético-harmonique [16] :

$$\mu_{Sc_{[0.5]}}(\mathcal{X},\,\mathcal{Y}) \;=\; \frac{1}{2}\,log\left(\frac{a}{h}\right) \tag{4.49}$$

Cette procédure de symétrisation peut améliorer les performances de classification, comparé aux termes asymétriques pris individuellement. C'est le cas lorsque référence et test ont des tailles comparables. Cependant, si les tailles diffèrent significativement ($\rho \not\approx 1$), cette symétrisation peut dégrader les performances par rapport à la meilleure des deux mesures asymétriques.

D'autre part, lorsque les tailles sont significativement différentes, la mesure $\mu(\mathcal{X},\,\mathcal{Y})$ donne de meilleurs résultats que la mesure $\mu(\mathcal{Y},\,\mathcal{X})$ lorsque $\rho \leq 1$, et inversement. En d'autres termes, il vaut mieux faire jouer le rôle du modèle de référence au locuteur (référence ou test) pour lequel on dispose de plus de données. Ceci suggère donc d'utiliser une procédure de symétrisation qui tienne compte explicitement de la quantité de données utilisée pour l'estimation des différents modèles.

En l'abscence d'un cadre théorique rigoureux, nous avons limité nos investigations à des essais empiriques, en postulant en outre une forme particulière de symétrisation, i.e. des combinaisons linéaires des termes asymétriques, pondérés par des coefficients qui sont fonction du nombre de vecteurs d'apprentissage et de test (respectivement M et N) :

$$\boxed{\begin{array}{l}\mu_{[\psi_{MN}]}(\mathcal{X},\,\mathcal{Y}) \;=\; \psi_{MN}\cdot\mu(\mathcal{X},\,\mathcal{Y}) \;+\; \psi_{NM}\cdot\mu(\mathcal{Y},\,\mathcal{X}) \\ \qquad\qquad\qquad\qquad\qquad\qquad \text{avec}\quad \psi_{MN} \;+\; \psi_{NM} \;=\; 1\end{array}} \tag{4.50}$$

Nous avons limité nos essais à deux fonctions ψ_{MN} et ψ_{NM} particulières :

$$\begin{array}{rcccccl}\psi_{MN} &=& \alpha_{MN} &=& \frac{\sqrt{M}}{\sqrt{M}+\sqrt{N}} &=& \frac{1}{1+\sqrt{\rho}} \\ \psi_{NM} &=& 1-\alpha_{MN} &=& \frac{\sqrt{N}}{\sqrt{M}+\sqrt{N}} &=& \frac{\sqrt{\rho}}{1+\sqrt{\rho}}\end{array} \tag{4.51}$$

et

$$\begin{array}{rcccccl}\psi_{MN} &=& \beta_{MN} &=& \frac{M}{M+N} &=& \frac{1}{1+\rho} \\ \psi_{NM} &=& 1-\beta_{MN} &=& \frac{N}{M+N} &=& \frac{\rho}{1+\rho}\end{array} \tag{4.52}$$

Une approche similaire a été utilisée par Montacié sur les résiduels de modèles AR-vectoriels [106]. Remarquons que, lorsque $M \geq N$, nous avons

$\rho \leq 1$, ce qui entraîne $0.5 \leq \alpha_{MN} \leq \beta_{MN}$.

Nous n'allons pas donner les expressions détaillées de chaque mesure, pour chaque ensemble de poids. Nous donnons juste un exemple. La mesure μ_G pondérée par β_{MN} devient :

$$\mu_{G_{[\beta_{MN}]}}(\mathcal{X},\,\mathcal{Y}) = \frac{M \cdot \mu_G(\mathcal{X},\,\mathcal{Y}) + N \cdot \mu_G(\mathcal{Y},\,\mathcal{X})}{M+N} \tag{4.53}$$

$$= \frac{1}{1+\rho}\, a - \frac{1-\rho}{1+\rho}\, log\; g + \frac{\rho}{1+\rho}\,\frac{1}{h} + \frac{1}{p}\,\boldsymbol{\delta}^T \left(\frac{\boldsymbol{\mathcal{X}}_0^{-1} + \rho\,\boldsymbol{\mathcal{Y}}_0^{-1}}{1+\rho} \right) \boldsymbol{\delta} - 1 \tag{4.54}$$

$$= \frac{1}{p}\left[\frac{1}{1+\rho}\, tr\,(\boldsymbol{\mathcal{Y}}_0 \boldsymbol{\mathcal{X}}_0^{-1}) + \frac{\rho}{1+\rho}\, tr\,(\boldsymbol{\mathcal{X}}_0 \boldsymbol{\mathcal{Y}}_0^{-1}) \right] - \frac{1}{p}\left[\frac{1-\rho}{1+\rho}\, log\, \left(\frac{det \boldsymbol{\mathcal{Y}}_0}{det \boldsymbol{\mathcal{X}}_0} \right) \right] + \frac{1}{p}\left[(\bar{\mathbf{y}} - \bar{\mathbf{x}})^T \left(\frac{\boldsymbol{\mathcal{X}}_0^{-1} + \rho\,\boldsymbol{\mathcal{Y}}_0^{-1}}{1+\rho} \right) (\bar{\mathbf{y}} - \bar{\mathbf{x}}) \right] - 1 \tag{4.55}$$

La symétrie de cette expression se vérifie aisément.

Bien qu'elles soient empiriques, les symétrisations utilisant α_{MN} et β_{MN} donnent généralement de meilleurs résultats que les symétrisations avec des poids égaux à $\frac{1}{2}$. L'expression optimale des mesures symétrisées peut très certainement être obtenue à partir de la théorie de l'estimation, mais ce n'est pas un problème évident.

Pour la mesure μ_{Dc}, nous avons procédé un peu différemment, car nous avons observé qu'il était plus efficace d'appliquer la symétrisation précédente à $log\ \mu_{Dc}$ plutôt qu'à μ_{Dc} elle-même[3]. La symétrisation de μ_{Dc} se fait donc

[3]Cependant, $log\ \mu_{Dc}$ ne peut pas être considérée comme une mesure au sens mathématique du terme, puisqu'elle n'est pas positive.

selon :

$$log\left[\mu_{Dc_{[\psi_{MN}]}}(\mathcal{X},\,\mathcal{Y})\right] = \psi_{MN} \cdot log\left[\mu_{Dc}(\mathcal{X},\,\mathcal{Y})\right] + \psi_{NM} \cdot log\left[\mu_{Dc}(\mathcal{Y},\,\mathcal{X})\right] \tag{4.56}$$

ce qui est équivalent à :

$$\mu_{Dc_{[\psi_{MN}]}}(\mathcal{X},\,\mathcal{Y}) = \mu_{Dc}(\mathcal{X},\,\mathcal{Y})^{\,\psi_{MN}} \cdot \mu_{Dc}(\mathcal{Y},\,\mathcal{X})^{\,\psi_{NM}} \tag{4.57}$$

4.7 Expériences et résultats

4.7.1 Description des expériences

4.7.1.a Tâche

La tâche évaluée ici est l'identification du locuteur en ensemble fermé indépendante du texte. Pour chaque mesure, sous une forme symétrique ou non, nous proposons pour un test donné l'identité du locuteur de la base de référence qui est le plus proche au sens de la mesure testée. Les résultats sont donnés en pourcentages d'erreurs d'identification.

4.7.1.b Bases de données

Les résultats sont reportés pour les bases TIMIT, FTIMIT et NTIMIT (cf. chapitre 3).

4.7.1.c Analyse acoustique du signal

L'analyse acoustique est décrite avec précision dans le chapitre 3. Nous avons choisi une fenêtre de 504 échantillons (31,5 ms).

4.7.1.d Protocoles expérimentaux

Nous avons testé 4 protocoles expérimentaux différents : le protocole 5.5 (apprentissage long - test long), le protocole 2.5 (apprentissage court - test long), le protocole 5.1 (apprentissage long - test court), et le protocole 2.1 (apprentissage court - test court). Pour chaque protocole, nous donnons les valeurs moyennes de la durée totale des données de parole par locuteur $T = M + N$, du rapport ρ entre la durée du test et de l'apprentissage, ainsi que les valeurs correspondantes des coefficients de normalisation α_{MN} et β_{MN} :

- **Protocole 5.5** : apprentissage long × test long
  $\bar{T} \approx 3000\ cs$, $\bar{\rho} \approx 1.10$, $\bar{\alpha}_{MN} \approx 0.48$, $\bar{\beta}_{MN} \approx 0.49$
- **Protocole 2.5** : apprentissage court × test long
  $\bar{T} \approx 2150\ cs$, $\bar{\rho} \approx 2.79$, $\bar{\alpha}_{MN} \approx 0.26$, $\bar{\beta}_{MN} \approx 0.37$
- **Protocole 5.1** : apprentissage long × test court
  $\bar{T} \approx 1750\ cs$, $\bar{\rho} \approx 0.22$, $\bar{\alpha}_{MN} \approx 0.82$, $\bar{\beta}_{MN} \approx 0.68$
- **Protocole 2.1** : apprentissage court × test court
  $\bar{T} \approx\ \ 900\ cs$, $\bar{\rho} \approx 0.56$, $\bar{\alpha}_{MN} \approx 0.64$, $\bar{\beta}_{MN} \approx 0.57$

4.7.2 Résultats

Les résultats sont organisés en trois tableaux. Les résultats concernant la base TIMIT sont donnés TAB. 4.3, ceux concernant la base FTIMIT sont donnés TAB. 4.4, et ceux concernant NTIMIT TAB. 4.5. Chaque tableau est lui-même subdivisé en 4 parties. La première partie de chaque tableau correspond au protocole 5.5, la seconde au protocole 2.5, la troisième au protocole 5.1, et enfin la quatrième partie de chaque tableau donne les résultats du protocole 2.1. Chaque partie de chaque tableau est organisée de la même façon. Les résultats relatifs à une famille de mesures donnée sont présentés en colonnes. La première ligne correspond aux scores des deux versions asymétriques de la mesure, tandis que les trois autres lignes fournissent les résultats de trois symétrisations différentes. Tous les résultats sont donnés en pourcentages d'erreurs d'identification.

Nous trouvons TAB. 4.1 la demie largeur des intervalles de confiance à 95 %, étant donné le pourcentage d'erreurs d'identification P, et le nombre de tests n. Le calcul est le suivant [146](Chapitre 14) :

$$\pm 2 \sqrt{\frac{P \cdot (100 - P)}{n}}$$

Pour chaque mesure, $\mu\ (\mathcal{X},\ \mathcal{Y})$ et $\mu\ (\mathcal{Y},\ \mathcal{X})$ donnent des résultats différents. Le terme $\mu\ (\mathcal{X},\ \mathcal{Y})$ donne de meilleurs résultats lorsque la durée de l'apprentissage est plus longue que la durée du test, et vice versa. La différence de performances entre les deux termes asymétriques est particulièrement visible pour la mesure μ_{Dc}.

Avec de la parole sans distortion (TIMIT et FTIMIT), la mesure μ_G est meilleure que la mesure μ_{Gc} et que toutes les autres mesures sur les matrices de covariances seules. En revanche, lorsque la variabilité du canal de

P | 100 - P	95 | 5	85 | 15	75 | 25	65 | 35	55 | 45
$n = 630$	$\pm$ 1.7 %	$\pm$ 2.8 %	$\pm$ 3.5 %	$\pm$ 3.8 %	$\pm$ 4.0%
$n = 3150$	$\pm$ 0.8 %	$\pm$ 1.3 %	$\pm$ 1.5 %	$\pm$ 1.7 %	$\pm$ 1.8%

TAB. 4.1: *Demie-largeur des intervalles de confiance à 95 % pour différentes valeurs du pourcentage P d'erreurs d'identification, et pour les deux durées de test. $n = 630$ correspond au nombre de tests longs, $n = 3150$ au nombre de tests courts.*

transmission est présente (NTIMIT), l'utilisation explicite du vecteur moyen est, comme prévu, un facteur de dégradation.

Dans leurs formes asymétriques, la mesure la plus efficace parmi les mesures sur les covariances seules est la mesure μ_{Sc}. Cependant, lorsque la symétrisation est appliquée, les performances ont tendances à se niveler, avec un très léger avantage pour la mesure μ_{Dc}.

Parmi les différentes symétrisations testées, la plus efficace semble être la symétrisation en β_{MN} et β_{NM} pour les mesures μ_G, μ_{Gc} et μ_{Sc}, tandis que la symétrisation α_{MN} et α_{NM} semble la plus adéquate pour *log* μ_{Dc}.

L'effet positif de la symétrisation est important lorsque la quantité de parole disponible est petite. Les différences les plus significatives sont obtenues pour le protocole 2.1. Nous trouvons TAB. 4.2 les ordres de grandeur des réductions de taux d'erreur entre les mesures sous leur forme asymétrique et la meilleure des versions symétriques. Si P est le pourcentage d'erreurs d'identification pour la forme asymétrique, et P' celui de la meilleure version symétrique, nous calculons la réduction du taux d'erreur selon :

$$\frac{P' - P}{P}$$

Cette réduction relative des taux d'erreur est donnée uniquement dans le cas du protocole 2.1 et 5.1, c'est-à-dire lorsque la durée du test est courte. Pour les deux autres protocoles, les écarts observés ne sont pas significatifs compte-tenu des intervalles de confiance calculés précédemment.

Les résultats présentés TAB. 4.2 montrent que la symétrisation améliore les performances des mesures sur les covariances seules (μ_{Gc}, μ_{Sc} and μ_{Dc}) d'autant plus que la base de données est intrinsèquement plus facile (TIMIT > FTIMIT > NTIMIT), et que la mesure originale est moins performante

Mesures	μ_G	μ_{Gc}	μ_{Sc}	μ_{Dc}
TIMIT	~ 30 %	~ 40 %	~ 10 %	~ 75 %
FTIMIT	~ 15 %	~ 10 %	~ 5 %	~ 40 %
NTIMIT	~ 1 %	~ 1 %	< 1 %	~ 10 %

TAB. 4.2: *Ordre de grandeur de la réduction relative du taux d'erreur entre les formes asymétriques et symétriques des mesures. Les résultats sont donnés pour une durée de test courte (protocoles 5.1 et 2.1).*

($\mu_{Dc} > \mu_{Gc} > \mu_{Sc}$). Néanmoins, lorsque le modèle Gaussien est moins pertinent (NTIMIT), ou lorsque la mesure initiale est déjà assez efficace (μ_{Sc}), la symétrisation est moins utile.

Mesures		μ_G		μ_{Gc}		μ_{Sc}		μ_{Dc}	
$\mu(\mathcal{X},\mathcal{Y})$	$\mu(\mathcal{Y},\mathcal{X})$	0.0 %	0.0 %	0.0 %	0.2 %	0.0 %	0.0 %	0.5 %	0.2 %
$\mu_{[0.5]}(\mathcal{X},\mathcal{Y})$		0.0 %		0.0 %		0.0 %		0.0 %	
$\mu_{[\alpha_{MN}]}(\mathcal{X},\mathcal{Y})$		0.0 %		0.0 %		0.0 %		0.0 %	
$\mu_{[\beta_{MN}]}(\mathcal{X},\mathcal{Y})$		0.0 %		0.0 %		0.0 %		0.0 %	

Protocole 5.5
apprentissage long (5 phrases $\approx$ 14.4 s) – test long (5 phrases $\approx$ 15.9 s)
$\bar{T} \approx 3000\ cs$, $\bar{\rho} \approx 1.10$, $\bar{\alpha}_{MN} \approx 0.48$, $\bar{\beta}_{MN} \approx 0.49$

Mesures		μ_G		μ_{Gc}		μ_{Sc}		μ_{Dc}	
$\mu(\mathcal{X},\mathcal{Y})$	$\mu(\mathcal{Y},\mathcal{X})$	6.8 %	0.6 %	13.3 %	2.9 %	5.1 %	3.6 %	26.7 %	7.9 %
$\mu_{[0.5]}(\mathcal{X},\mathcal{Y})$		1.9 %		5.4 %		4.3 %		4.9 %	
$\mu_{[\alpha_{MN}]}(\mathcal{X},\mathcal{Y})$		1.3 %		4.3 %		3.8 %		3.0 %	
$\mu_{[\beta_{MN}]}(\mathcal{X},\mathcal{Y})$		0.8 %		3.5 %		4.0 %		3.0 %	

Protocole 2.5
apprentissage court (2 phrases $\approx$ 5.7 s) – test long (5 phrases $\approx$ 15.9 s)
$\bar{T} \approx 2150\ cs$, $\bar{\rho} \approx 2.79$, $\bar{\alpha}_{MN} \approx 0.26$, $\bar{\beta}_{MN} \approx 0.37$

Mesures		μ_G		μ_{Gc}		μ_{Sc}		μ_{Dc}	
$\mu(\mathcal{X},\mathcal{Y})$	$\mu(\mathcal{Y},\mathcal{X})$	2.1 %	10.3 %	3.8 %	21.2 %	2.7 %	6.4 %	16.4 %	40.6 %
$\mu_{[0.5]}(\mathcal{X},\mathcal{Y})$		2.8 %		6.1 %		3.0 %		2.7 %	
$\mu_{[\alpha_{MN}]}(\mathcal{X},\mathcal{Y})$		1.6 %		2.9 %		2.7 %		2.4 %	
$\mu_{[\beta_{MN}]}(\mathcal{X},\mathcal{Y})$		1.6 %		2.4 %		2.4 %		5.2 %	

Protocole 5.1
apprentissage long (5 phrases $\approx$ 14.4 s) – test court (1 sentence $\approx$ 3.2 s)
$\bar{T} \approx 1750\ cs$, $\bar{\rho} \approx 0.22$, $\bar{\alpha}_{MN} \approx 0.82$, $\bar{\beta}_{MN} \approx 0.68$

Mesures		μ_G		μ_{Gc}		μ_{Sc}		μ_{Dc}	
$\mu(\mathcal{X},\mathcal{Y})$	$\mu(\mathcal{Y},\mathcal{X})$	14.2 %	21.8 %	26.5 %	35.1 %	18.1 %	22.3 %	47.1 %	54.6 %
$\mu_{[0.5]}(\mathcal{X},\mathcal{Y})$		10.3 %		17.8 %		17.3 %		15.6 %	
$\mu_{[\alpha_{MN}]}(\mathcal{X},\mathcal{Y})$		9.9 %		16.6 %		17.0 %		15.8 %	
$\mu_{[\beta_{MN}]}(\mathcal{X},\mathcal{Y})$		10.3 %		16.4 %		16.7 %		19.9 %	

Protocole 2.1
apprentissage court (2 phrases $\approx$ 5.7 s) – test court (1 sentence $\approx$ 3.2 s)
$\bar{T} \approx\ 900\ cs$, $\bar{\rho} \approx 0.56$, $\bar{\alpha}_{MN} \approx 0.64$, $\bar{\beta}_{MN} \approx 0.57$

TAB. 4.3: *Méthodes statistiques du second ordre. Identification du locuteur indépendante du texte. Base de données TIMIT. Les résultats sont donnés en pourcentages d'erreurs d'identification.*

Mesures		μ_G		μ_{Gc}		μ_{Sc}		μ_{Dc}	
$\mu(\mathcal{X},\mathcal{Y})$	$\mu(\mathcal{Y},\mathcal{X})$	1.6 %	0.6 %	3.9 %	1.7 %	2.4 %	1.9 %	9.5 %	4.6 %
$\mu_{[0.5]}(\mathcal{X},\mathcal{Y})$		0.6 %		2.1 %		2.1 %		1.7 %	
$\mu_{[\alpha_{MN}]}(\mathcal{X},\mathcal{Y})$		0.5 %		2.1 %		2.1 %		1.4 %	
$\mu_{[\beta_{MN}]}(\mathcal{X},\mathcal{Y})$		0.5 %		2.1 %		2.2 %		1.6 %	

Protocole 5.5
apprentissage long (5 phrases $\approx$ 14.4 s) – test long (5 phrases $\approx$ 15.9 s)
$\bar{T} \approx 3000\ cs$, $\bar{\rho} \approx 1.10$, $\bar{\alpha}_{MN} \approx 0.48$, $\bar{\beta}_{MN} \approx 0.49$

Mesures		μ_G		μ_{Gc}		μ_{Sc}		μ_{Dc}	
$\mu(\mathcal{X},\mathcal{Y})$	$\mu(\mathcal{Y},\mathcal{X})$	21.3 %	11.4 %	36.8 %	23.0 %	27.1 %	23.6 %	55.9 %	34.4 %
$\mu_{[0.5]}(\mathcal{X},\mathcal{Y})$		13.0 %		23.2 %		23.6 %		23.3 %	
$\mu_{[\alpha_{MN}]}(\mathcal{X},\mathcal{Y})$		12.1 %		22.5 %		23.8 %		22.4 %	
$\mu_{[\beta_{MN}]}(\mathcal{X},\mathcal{Y})$		11.0 %		22.2 %		23.6 %		23.8 %	

Protocole 2.5
apprentissage court (2 phrases $\approx$ 5.7 s) – test long (5 phrases $\approx$ 15.9 s)
$\bar{T} \approx 2150\ cs$, $\bar{\rho} \approx 2.79$, $\bar{\alpha}_{MN} \approx 0.26$, $\bar{\beta}_{MN} \approx 0.37$

Mesures		μ_G		μ_{Gc}		μ_{Sc}		μ_{Dc}	
$\mu(\mathcal{X},\mathcal{Y})$	$\mu(\mathcal{Y},\mathcal{X})$	18.6 %	32.1 %	30.0 %	50.2 %	29.3 %	33.7 %	51.9 %	66.7 %
$\mu_{[0.5]}(\mathcal{X},\mathcal{Y})$		18.2 %		32.7 %		29.6 %		27.8 %	
$\mu_{[\alpha_{MN}]}(\mathcal{X},\mathcal{Y})$		15.8 %		28.2 %		28.3 %		26.9 %	
$\mu_{[\beta_{MN}]}(\mathcal{X},\mathcal{Y})$		16.4 %		27.4 %		28.0 %		35.6 %	

Protocole 5.1
apprentissage long (5 phrases $\approx$ 14.4 s) – test court (1 sentence $\approx$ 3.2 s)
$\bar{T} \approx 1750\ cs$, $\bar{\rho} \approx 0.22$, $\bar{\alpha}_{MN} \approx 0.82$, $\bar{\beta}_{MN} \approx 0.68$

Mesures		μ_G		μ_{Gc}		μ_{Sc}		μ_{Dc}	
$\mu(\mathcal{X},\mathcal{Y})$	$\mu(\mathcal{Y},\mathcal{X})$	45.3 %	50.3 %	60.2 %	67.8 %	57.4 %	58.8 %	76.9 %	79.4 %
$\mu_{[0.5]}(\mathcal{X},\mathcal{Y})$		38.6 %		56.1 %		55.6 %		53.5 %	
$\mu_{[\alpha_{MN}]}(\mathcal{X},\mathcal{Y})$		38.2 %		54.7 %		55.5 %		53.2 %	
$\mu_{[\beta_{MN}]}(\mathcal{X},\mathcal{Y})$		38.6 %		54.2 %		55.6 %		56.4 %	

Protocole 2.1
apprentissage court (2 phrases $\approx$ 5.7 s) – test court (1 sentence $\approx$ 3.2 s)
$\bar{T} \approx\ 900\ cs$, $\bar{\rho} \approx 0.56$, $\bar{\alpha}_{MN} \approx 0.64$, $\bar{\beta}_{MN} \approx 0.57$

TAB. 4.4: *Méthodes statistiques du second ordre. Identification du locuteur indépendante du texte. Base de données FTIMIT. Les résultats sont donnés en pourcentages d'erreurs d'identification.*

Mesures		μ_G		μ_{Gc}		μ_{Sc}		μ_{Dc}	
$\mu(\mathcal{X},\mathcal{Y})$	$\mu(\mathcal{Y},\mathcal{X})$	54.7 %	49.7 %	40.5 %	37.0 %	34.0 %	35.1 %	59.0 %	49.0 %
$\mu_{[0.5]}(\mathcal{X},\mathcal{Y})$		50.6 %		37.0 %		33.6 %		32.1 %	
$\mu_{[\alpha_{MN}]}(\mathcal{X},\mathcal{Y})$		51.0 %		37.0 %		33.5 %		31.6 %	
$\mu_{[\beta_{MN}]}(\mathcal{X},\mathcal{Y})$		51.6 %		36.4 %		33.5 %		31.4 %	

Protocole 5.5
apprentissage long (5 phrases $\approx$ 14.4 s) – test long (5 phrases $\approx$ 15.9 s)
$\bar{T} \approx 3000\ cs$, $\bar{\rho} \approx 1.10$, $\bar{\alpha}_{MN} \approx 0.48$, $\bar{\beta}_{MN} \approx 0.49$

Mesures		μ_G		μ_{Gc}		μ_{Sc}		μ_{Dc}	
$\mu(\mathcal{X},\mathcal{Y})$	$\mu(\mathcal{Y},\mathcal{X})$	82.4 %	75.1 %	77.8 %	69.0 %	71.6 %	70.3 %	87.6 %	77.5 %
$\mu_{[0.5]}(\mathcal{X},\mathcal{Y})$		75.6 %		70.5 %		70.2 %		69.7 %	
$\mu_{[\alpha_{MN}]}(\mathcal{X},\mathcal{Y})$		75.2 %		69.5 %		70.0 %		69.2 %	
$\mu_{[\beta_{MN}]}(\mathcal{X},\mathcal{Y})$		74.3 %		68.7 %		69.5 %		69.2 %	

Protocole 2.5
apprentissage court (2 phrases $\approx$ 5.7 s) – test long (5 phrases $\approx$ 15.9 s)
$\bar{T} \approx 2150\ cs$, $\bar{\rho} \approx 2.79$, $\bar{\alpha}_{MN} \approx 0.26$, $\bar{\beta}_{MN} \approx 0.37$

Mesures		μ_G		μ_{Gc}		μ_{Sc}		μ_{Dc}	
$\mu(\mathcal{X},\mathcal{Y})$	$\mu(\mathcal{Y},\mathcal{X})$	79.3 %	86.2 %	74.6 %	86.5 %	73.7 %	77.0 %	85.9 %	94.8 %
$\mu_{[0.5]}(\mathcal{X},\mathcal{Y})$		80.7 %		76.6 %		74.8 %		74.8 %	
$\mu_{[\alpha_{MN}]}(\mathcal{X},\mathcal{Y})$		78.9 %		74.6 %		73.9 %		73.6 %	
$\mu_{[\beta_{MN}]}(\mathcal{X},\mathcal{Y})$		78.6 %		73.9 %		73.3 %		79.5 %	

Protocole 5.1
apprentissage long (5 phrases $\approx$ 14.4 s) – test court (1 sentence $\approx$ 3.2 s)
$\bar{T} \approx 1750\ cs$, $\bar{\rho} \approx 0.22$, $\bar{\alpha}_{MN} \approx 0.82$, $\bar{\beta}_{MN} \approx 0.68$

Mesures		μ_G		μ_{Gc}		μ_{Sc}		μ_{Dc}	
$\mu(\mathcal{X},\mathcal{Y})$	$\mu(\mathcal{Y},\mathcal{X})$	89.9 %	91.0 %	88.0 %	91.1 %	86.3 %	87.2 %	93.6 %	96.9 %
$\mu_{[0.5]}(\mathcal{X},\mathcal{Y})$		88.3 %		86.3 %		85.7 %		85.6 %	
$\mu_{[\alpha_{MN}]}(\mathcal{X},\mathcal{Y})$		88.3 %		85.8 %		85.6 %		85.2 %	
$\mu_{[\beta_{MN}]}(\mathcal{X},\mathcal{Y})$		88.4 %		85.0 %		85.6 %		87.3 %	

Protocole 2.1
apprentissage court (2 phrases $\approx$ 5.7 s) – test court (1 sentence $\approx$ 3.2 s)
$\bar{T} \approx\ 900\ cs$, $\bar{\rho} \approx 0.56$, $\bar{\alpha}_{MN} \approx 0.64$, $\bar{\beta}_{MN} \approx 0.57$

TAB. 4.5: *Méthodes statistiques du second ordre. Identification du locuteur indépendante du texte. Base de données NTIMIT. Les résultats sont donnés en pourcentages d'erreurs d'identification.*

4.8 Discussion

Notre évaluation montre qu'on peut atteindre de remarquables performances sur TIMIT avec les méthodes statistiques du second ordre, c'est-à-dire avec des méthodes reposant sur un modèle de locuteur sous-jacent très simple. Cependant, la base TIMIT ne constitue pas une base de données particulièrement difficile pour faire de la reconnaissance du locuteur. En particulier, on remarque déjà une dégradation significative des performances sur la base FTIMIT, ce qui montre qu'une partie importante de l'information sur le locuteur se trouve dans la bande de fréquence 4-8 kHz. Si on dégrade encore les conditions en ajoutant la distortion et la variabilité du canal de transmission (NTIMIT), les pourcentages d'erreurs d'identification augmentent encore dans de larges proportions. Les bases de données issues de TIMIT ne permettent pas de prendre en compte la dérive temporelle, mais on peut penser que ce facteur supplémentaire dégraderait également les performances.

Les méthodes statistiques du second ordre constituent une famille de méthodes simples et efficaces pour des tâches relativement limitées. Elles ne sont cependant pas la solution ultime au problème de la reconnaissance du locuteur.

4.8.1 Au delà des performances

Cependant, ces méthodes offrent plusieurs avantages.

Elles sont très simples à implanter et facile à reproduire.

En particulier, les mesures issues du critère de maximum de vraisemblance (μ_G et μ_{Gc}) dans leur forme asymétrique sont des cas particuliers d'approches plus générales fréquemment utilisées en reconnaissance du locuteur indépendante du texte. Un modèle mono-Gaussien du locuteur est équivalent à un dictionnaire de quantification vectorielle à une entrée, associé à une distance de Mahalanobis. Il est aussi équivalent à tous type de modèle de Markoc caché (droite-gauche, ergodique, ...) à un état et à une distribution Gaussienne par état. Il est aussi un cas particulier du mélange de Gaussiennes pleines, où le mélange se réduit à une Gaussienne. Enfin, on peut aussi l'interpréter comme un modèle de prédiction vectorielle d'ordre 0 (modèles prédictifs linéaires ou modèles prédictifs connexionnistes).

Les mesures μ_G et μ_{Gc} sont à l'intersection de plusieurs approches classiques, qui sont des extensions de ce modèle de base dans différentes directions (différences au niveau de la mesure de similarité, utilisation de contraintes temporelles plus ou moins fortes, raffinement du modèle de locuteur, filtrage des paramètres acoustiques, ...).

Etant donnée l'extrême simplicité des mesures statistiques du second ordre, nous proposons d'étalonner systématiquement une tâche de reconnaissance du locuteur par une ou deux de ces mesures, dans le but d'obtenir un score de référence indiquant la complexité intrinsèque de la base de données choisie et du protocole. En particulier, le prétraitement du signal, son analyse acoustique, et la composition des bases d'apprentissage et de test, ainsi que la stratégie de décision devraient être les mêmes que ceux de la nouvelle méthode testée.

4.8.2 Une méthode de référence

Même si les versions asymétriques des mesures reposant sur le critère du maximum de vraisemblance ne donnent pas systématiquement de meilleurs résultats que les autres mesures statistiques du second ordre, μ_G et μ_{Gc} semblent être des choix préférables comme mesures de référence dans deux cas : lorsqu'elles sont comparées à d'autres approches asymétriques (ce qui est le cas par exemple de la quantification vectorielle, des modèles de Markov cachés, ou des mélanges de Gaussiennes), et quand la longueur de l'apprentissage est significativement plus grande que celle du test.

Le choix entre μ_G et μ_{Gc} dépend alors du traitement appliqué aux données du système en cours d'évaluation, c'est-à-dire du fait que, par exemple, la moyenne à long terme des vecteurs de paramètres est soustraite ou non. Les mesures $\mu_{G_{[\beta_{MN}]}}$ et $\mu_{Gc_{[\beta_{MN}]}}$ peuvent aussi être testées simplement, et peuvent être également systématiquement utilisées. Cependant, le manque de justification théorique de cette symétrisation, et la faible amélioration qu'elle apporte en réalité font que l'utilisation de ces deux autres mesures est plus discutable. Néanmoins, si l'approche qui est évaluée est symétrique, il vaut mieux la comparer à une version symétrique.

4.9 Influence de la durée et du contenu phonétique

Après avoir décrit en détails les mesures statistiques du second ordre, nous nous intéressons maintenant à leurs performances en fonction de la durée de l'apprentissage et du test, et en fonction du contenu phonétique du test.

Les expériences présentées dans cette section reprennent en grande partie un article publié à *EUROSPEECH 95* [91]. Une copie de cet article se trouve en annexe de ce document. Ces expériences résultent d'une collaboration avec le Laboratoire d'Informatique d'Avignon (L.I.A.). Il s'agit d'adopter, avec les méthodes statistiques du second ordre, une démarche analytique, et non plus uniquement globale (cf. section 1.5).

A propos des démarches analytiques en reconnaissance du locuteur, on peut d'ailleurs lire [20], [21], [36], [85], [19].

4.9.1 Introduction

Nous étudions dans cette section les performances des méthodes statistiques du second-ordre en fonction du contenu phonétique des tests. Plusieurs auteurs ont récemment adopté ce type de démarche : Eatock [36], qui utilise des dictionnaires de quantification vectorielle, et qui conclut que les consonnes nasales et les voyelles donnent les meilleures performances sur une base en langue anglaise ; Le Floch [85], qui utilise les modèles AR-vectoriels, et qui conclut que les voyelles, les diphtongues et les consonnes nasales donnent les meilleures performances, également sur une base de langue anglaise.

La particularité de ce travail est d'une part l'utilisation des méthodes statistiques du second-ordre, d'autre part le fait que l'apprentissage et le test ne sont pas tout à fait de même nature. En effet, l'apprentissage comporte 15 secondes de parole en Français, phonétiquement équilibrée, tandis que le test est constitué uniquement de segments de parole de même appartenance phonétique.

4.9.2 Mesures utilisées

Nous nous sommes restreints pour cette étude à trois mesures statistiques du second-ordre particulières. Nous avons utilisé les trois mesures μ_G, μ_{Gc} et μ_{Sc}, symétrisées avec les coefficients β_{MN} et β_{NM}.

4.9.3 Base de données et analyse du signal

Pour ces expériences, nous avons utilisé une base de données enregistrée au Laboratoire d'Informatique d'Avignon (L.I.A.), car nous avions besoin de disposer d'une quantité suffisante de parole pour chaque locuteur. Le corpus est constitué de phrases lues en Français, phonétiquement équilibrées [31], dont la transcription phonétique peut être trouvée dans la base BDSONS. Les phrases sont affichées à l'écran. L'enregistrement commence et se termine automatiquement, grâce à un détecteur de parole. Chaque phrase est enregistrée à l'aide d'un micro SHURE SM10A, échantillonnée à 16 kHz, puis codée sur 16 bits, grâce à une carte OROS AU22. Les enregistrements ont lieu dans un couloir du laboratoire, il y a donc un bruit de fond non-négligeable. La base comprend 67 locuteurs, pour la plupart des étudiants, ayant enregistré approximativement 3 minutes de parole chacun.

L'analyse de parole réalisée est celle qui est présentée dans le chapitre 3, avec une longueur de fenêtre de 504 échantillons (31,5 ms).

4.9.4 Expériences sur la durée

La première partie des expériences réalisées concerne l'étude de l'influence de la durée sur les performances des méthodes étudiées. Nous avons choisi plusieurs durées pour l'apprentissage (15, 10, 6, 3 et 2 secondes), et plusieurs durées pour le test (10, 6, 3, 2 et 1 secondes). Pour chaque locuteur, les phrases sont concaténées de façon aléatoire. Les silences de début et de fin de phrase ne sont pas retirés, mais n'excèdent généralement pas 0,1 seconde. Pour plus de détails sur cette expérience, on peut lire [91], dont une copie se trouve en annexe H de ce document. Les résultats détaillés y sont également reportés. Nous donnons FIG. 4.1 une représentation graphique des résultats.

Ces tests nous ont permis en particulier de choisir les durées d'apprentissage et de test pour la deuxième partie des expériences. Nous avons choisi un apprentissage de 15 secondes, ce qui assure une bonne couverture phonétique. La durée de test est prise égale à 1 seconde, pour permettre d'avoir un nombre de tests suffisant (ce qui ne serait pas le cas si on prenait des tests plus longs), et aussi pour permettre de différencier les scores (qui seraient trop bons et proches avec des tests plus longs).

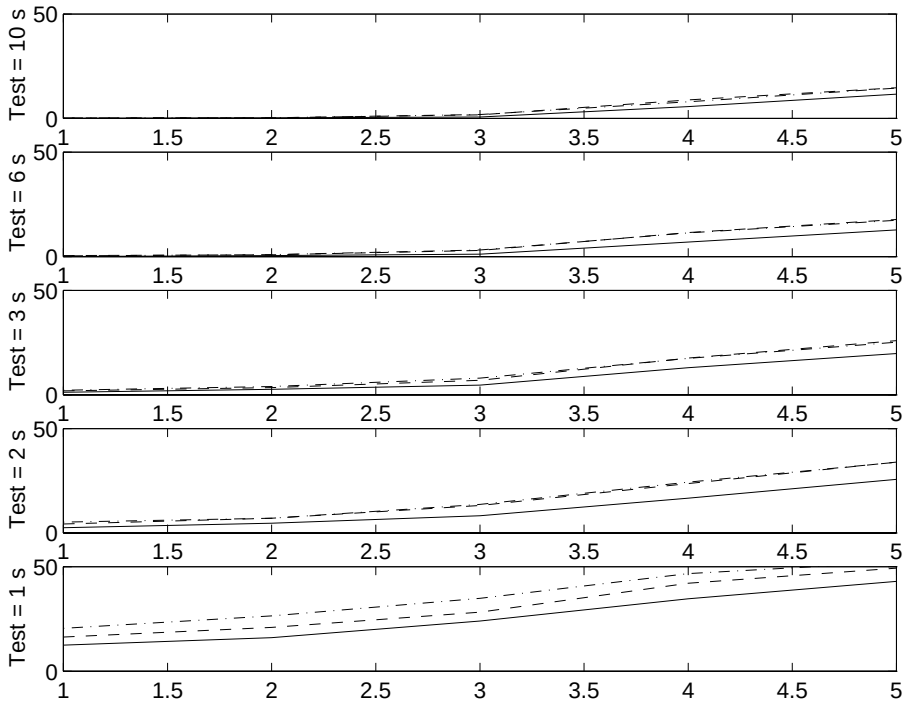


FIG. 4.1: *Influence de la durée sur les performances de trois mesures symétrisées en β_{MN} : la mesure μ_G (en traits pleins), la mesure μ_{Gc} (en traits $-\cdot-$) et la mesure μ_{Sc} (en traits $--$). **Les durées d'apprentissage vont de 15 secondes (graduation 1) à 2 secondes (graduation 5).** Les courbes sont données en pourcentages d'erreurs d'identification pour différentes durées de test.*

4.9.5 Expériences sur le contenu phonétique

4.9.5.a Segmentation

Pour étudier l'influence du contenu phonétique sur quelques mesures statistiques du second-ordre, nous utilisons un système automatique pour segmenter la parole en phonèmes spécifiques ou en classes phonétiques. Le système de localisation automatique repose sur un décodeur acoustico-phonétique ascendant, indépendant du locuteur, dont le principe et les détails sont expliqués dans [102] et [49]. Pour chaque phrase, le décodeur propose un treillis d'hypothèses. Elles sont alors alignées par un algorithme d'alignement gauche-droit. Pour obtenir une grande précision de localisation, nous

avons fixé le seuil de rejet assez haut. Ainsi, 55 % des phrases ont été rejetées car elles ne présentaient pas un degré de fiabilité suffisant. Enfin, il faut noter que l'algorithme donne la localisation pour le noyau d'un phonème, et il se peut que quelques trames appartiennent surtout aux transitions, et dans quelques cas rares aux phonèmes voisins.

4.9.5.b Protocole expérimental

L'apprentissage est constitué de 15 secondes de parole phonétiquement équilibrée, c'est-à-dire qu'aucun événement phonétique spécifique n'est sélectionné pour l'apprentissage. On utilise en fait exactement les mêmes apprentissages que ceux de l'expérience sur les durées avec un apprentissage de 15 secondes.

Pour les tests, des événements phonétiques spécifiques ou des classes phonétiques sont sélectionnés, sur le reste de la parole d'un locuteur, c'est-à-dire celle qui n'est pas déjà utilisée pour l'apprentissage. Pour un phonème donné ou une classe phonétique donnée, toutes les trames étiquetées de façon identique sont alors concaténées ensemble de manière à obtenir autant de tests d'1 seconde que possible.

4.9.5.c Description des classes phonétiques

Nous donnons les résultats sur les phonèmes et les classes phonétiques pour lesquelles nous avons au moins 40 tests.

Une première classe, notée *Tout* regroupe tous les phonèmes, mais n'est pas équivalente à l'expérience sur les durées, car il s'agit de phonèmes concaténés. En particulier, cette classe ne comporte pas de silences, de pauses ou d'événements non-linguistiques.

Enfin, les autres classes sont les suivantes : *Voyelles* (contient les voyelles orales et nasales mais pas les semi-voyelles), *Voyelles Orales*, *Voyelles Nasales*, *Consonnes* (contient aussi les semi-voyelles), *Consonnes non-Nasales* (contient toutes les consonnes exceptées les consonnes nasales), *Consonnes Nasales*, *Occlusives*, *Fricatives* et *Liquides+Semi-Voyelles* (forment ensemble une classe).

4.9.5.d Résultats

Les résultats sur différentes classes phonétiques sont présentés TAB. 4.6. Pour les autres résultats, et notamment ceux sur les différents types de phonèmes, nous renvoyons à [91] (en annexe H du document).

4.9.6 Commentaires

La mesure μ_G donne de meilleurs résultats que les autres, ce qui est quelque peu surprenant. En fait, bien qu'on s'attende à ce que le vecteur moyen au sein d'une classe phonétique soit fortement dépendant de cette classe, et donc soit assez différent du vecteur moyen d'apprentissage, celui-ci semble pourtant garder une certaine consistance à travers les classes phonétiques.

Notons que les résultats pour la classe *Tout* sont meilleurs que ceux de la section sur la durée correspondant au protocole 15 s $\times$ 1 s. Ceci est du au fait que, pour la classe *Tout*, on concatène plusieurs segments de parole identifiés comme appartenant à une classe phonétique, et n'étant donc *a priori* ni du silence, ni des pauses, ce qui n'est pas le cas dans la section sur la durée.

De façon plus détaillée, la classe *Voyelles* donne des performances meilleures que la classe *Consonnes*. La classe *Voyelles Nasales* est meilleure que les classes *Voyelles* et *Voyelles Orales*. La classe *Consonnes non-Nasales*, et plus particulièrement les classes *Occlusives* et *Fricatives*, donnent de moins bons résultats que la classe *Consonnes* qui les regroupent ensemble, tandis que les classes *Liquides + Semi-Voyelles* et *Consonnes Nasales* conduisent à des scores meilleurs.

De toutes les classes phonétiques, les meilleurs résultats sont obtenus avec les classes *Liquides + Semi-Voyelles* et *Voyelles Nasales*, mais les résultats sur les autres classes ne sont jamais très mauvais.

> Les méthodes statistiques du second-ordre semblent donc capturer une information caractéristique du locuteur distribuée à travers les phonèmes.

4.10 Synthèse sur les méthodes statistiques du second-ordre

☞ Nous avons étudié les propriétés et les performances de plusieurs approches simples en reconnaissance du locuteur, nous les avons comparées, et nous avons identifié leurs limites. La large évaluation que nous avons faites sur TIMIT, FTIMIT et NTIMIT illustre clairement l'influence de la qualité et de la durée de la parole sur les performances, et le fait que des méthodes simples peuvent aussi être très performantes lorsque la qualité de la parole est bonne. En particulier, ce type d'approches peut être utilisé dans des applications telles que l'étiquetage automatique par locuteur de données radiophoniques ou télévisuelles, pour lesquelles la qualité du signal est relativement bonne et constante, et pour lesquelles la dérive temporelle est un phénomène plus marginal.

☞ D'autre part, notre travail met en évidence l'extrême précaution avec laquelle on peut tirer les mérites de telle ou telle méthode en dehors de tout point de comparaison rigoureux. Puisqu'il apparaît peu réaliste de comparer une nouvelle méthode à toutes les méthodes classiques de l'état de l'art, il semble préférable de le faire uniquement avec une ou deux approches de référence.

> Les méthodes statistiques du second ordre fondées sur une approche de type maximum de vraisemblance réalisent un bon compromis entre simplicité d'implantation, reproductibilité, faible besoin en stockage et performances, et constituent de ce fait un bon système de référence.

De plus, elles apparaissent, dans leur forme asymétrique, comme de plus simples versions d'approches plus élaborées. Enfin, bien que la symétrisation ne soit pas systématiquement un facteur d'amélioration, on peut malgré tout utiliser les versions symétriques également comme méthodes de référence, surtout si la mesure à évaluer est elle-même symétrique.

☞ Finalement, il apparaît, dans les expériences sur l'influence du contenu phonétique, que l'homogéméité du test soit un facteur permettant d'améliorer les performances, bien que cette amélioration soit faiblement dépendante de la classe phonétique choisie.

☞ Pour terminer ce chapitre sur les méthodes statistiques du second ordre, nous souhaitons revenir sur les deux hypothèses sous jacentes à nos différentes approches. La première hypothèse concerne l'indépendance de deux trames consécutives (hypothèse $\mathcal{H}_2$). Cette hypothèse est bien sûr non réaliste, puisque certains phonèmes couvrent plusieurs trames qui, dans ce cas-là, sont très certainement fortement dépendantes. Cette hypothèse reste cependant nécessaire pour pouvoir utiliser certaines simplifications dans les calculs de probabilités. On peut néanmoins rechercher une méthode qui rendrait plus indépendantes entre elles les trames utilisées. On peut s'attendre à ce que ce type de démarche améliore les performances, si toutefois il permet de conserver suffisamment d'information.

☞ La deuxième hypothèse est la répartition Gaussienne des trames acoustiques (hypothèse $\mathcal{H}_1$). Cette hypothèse n'est pas facilement vérifiable, et il est difficile de savoir dans quelle mesure elle est réaliste. On peut cependant imaginer plusieurs procédés pour rendre les trames davantage Gaussiennes. En particulier, si nous remplaçons les trames acoustiques par les résidus vectoriels de prédiction linéaire, nous espérons obtenir des trames vectorielles dont la répartition probabiliste est plus proche d'une Gaussienne multidimensionnelle. C'est une des idées qui nous a conduit à utiliser les modèles AR-vectoriels en reconnaissance du locuteur. Cette approche est décrite dans le chapitre 5. On peut aussi adopter des démarches de type quantification vectorielle, ou encore décomposition temporelle des trames acoustiques.

Le chapitre suivant propose une première façon de prendre en compte les aspects dynamiques des séquences de vecteurs de paramètres, ce qui n'est pas le cas des méthodes statistiques du second ordre, qui se contentent de modéliser statiquement ces séquences. Ce chapitre propose en effet une étude systématique de différentes mesures et normalisations à base de résiduels de prédiction linéaire vectorielle.

Phonèmes	*Tout* (1334)	*Voyelles* (1247)
μ_G	**9.4**	**2.9**
μ_{Gc}	19.2	7.7
μ_{Sc}	16.5	8.0
Phonèmes	*Voyelles Orales* (1206)	*Voyelles Nasales* (262)
μ_G	**4.0**	**1.9**
μ_{Gc}	8.7	10.7
μ_{Sc}	9.5	6.9
Phonèmes	*Consonnes* (1247)	*Consonnes non-Nasales* (1186)
μ_G	**3.8**	**5.3**
μ_{Gc}	8.9	10.6
μ_{Sc}	8.4	11.0
Phonèmes	*Consonnes Nasales* (390)	*Occlusives* (693)
μ_G	**3.1**	**5.8**
μ_{Gc}	14.1	8.8
μ_{Sc}	4.9	8.7
Phonèmes	*Fricatives* (486)	*Liquides + Semi-Voyelles* (277)
μ_G	**7.8**	**1.1**
μ_{Gc}	16.3	7.6
μ_{Sc}	13.8	7.6

TAB. 4.6: *Pourcentages d'erreur d'identification dans le cas où l'apprentissage est composé de 15 secondes de parole phonétiquement équilibrée, et où le test est composé d'1 seconde de parole homogène appartenant à une même classe phonétique. Le nombre de tests effectués pour chaque classe phonétique est indiqué entre parenthèses.*

Chapitre 5

Modèles AR-vectoriels

Abordons maintenant une autre approche utilisée en reconnaissance automatique du locuteur. Il s'agit des modèles AR-vectoriels.

5.1 Motivations

☞ Certaines publications rapportent que les modèles AR-vectoriels sont de bons modèles de locuteurs, et donnent de bonnes performances en reconnaissance du locuteur. Nous avons donc voulu les comparer avec les méthodes statistiques du second ordre.

☞ D'autre part, on sait que la prise en compte d'une information de type dynamique est très certainement un facteur d'amélioration des performances. Or, les méthodes présentées dans le chapitre 4 ne prennent en compte qu'une information sur le locuteur de type statique. Nous avons donc voulu tester une approche comme les modèles AR-vectoriels, qui sont supposés avoir une bonne capacité à modéliser les caractéristiques dynamiques des locuteurs[1]. Nous avons d'ailleurs voulu mettre cette hypothèse à l'épreuve dans ce chapitre.

☞ Nous voulons également tester de façon systématique différentes combinaisons des matrices de covariance des erreurs résiduelles. En particulier, nous appliquons les mesures de similarité étudiées dans le chapitre 4, ainsi que les différentes procédures de symétrisation.

☞ Finalement, comme nous l'avons écrit à la fin du chapitre 4, nous souhaitons utiliser des trames vectorielles dont la distribution de probabilité

[1] Ce chapitre reprend en grande partie le travail présenté dans [92], dont une copie se trouve en annexe H de ce document.

ressemble davantage à une densité Gaussienne multi-dimensionnelle. Nous espérons que la séquence des erreurs résiduelles de prédiction vectorielle est capable de répondre à cet objectif.

5.2 Introduction

Les modèles auto-régressifs (AR) vectoriels ont fait récemment l'objet de plusieurs études en reconnaissance automatique du locuteur [56], [2], [105], [12], [106], [57], [83], [84]. Bien que le point commun de tous ces travaux soit la modélisation d'un locuteur par un modèle AR-vectoriel, la façon de mesurer la similarité entre de tels modèles diffère d'un article à l'autre.

D'autre part, l'utilisation des modèles AR-vectoriels a souvent été motivée par la conviction que cette approche était un moyen de modéliser les caractéristiques dynamiques du locuteur.

Nous nous sommes donc intéressés d'une part à une évaluation systématique d'un grand nombre de mesures de similarité entre modèles AR-vectoriels, d'autre part à la remise en question de l'hypothèse sur la modélisation des caractéristiques dynamiques du locuteur par les modèles AR-vectoriels.

Pour mettre cette hypothèse à l'épreuve, nous proposons un protocole expérimental qui consiste en la destruction de la structure temporelle des vecteurs de paramètres. En effet, outre un modèle AR-vectoriel d'ordre 2 appris normalement sur des vecteurs de paramètres présentés dans leur ordre naturel, nous proposons de calculer également les coefficients matriciels d'un modèle à partir de vecteurs de paramètres mélangés aléatoirement. Ce processus ayant détruit la structure temporelle des vecteurs de paramètres, on peut s'attendre à une dégradation importante des performances, si toutefois les modèles AR-vectoriels sont effectivement un bon moyen de modéliser dynamiquement un locuteur.

5.3 Définitions et notations

Soit $\{\mathbf{x}_t\}_{1\leq t\leq M}$ une séquence de vecteurs de dimension p. Nous définissons aussi les vecteurs centrés correspondants $\mathbf{x}_t^* = \mathbf{x}_t - \bar{\mathbf{x}}$ où $\bar{\mathbf{x}}$ est le vecteur moyen de la séquence $\{\mathbf{x}_t\}$.

Notons $\boldsymbol{\mathcal{X}}_0$ la matrice de covariance de $\{\mathbf{x}_t\}$:

$$\boldsymbol{\mathcal{X}}_0 = \frac{1}{M}\sum_{t=1}^{M}(\mathbf{x}_t - \bar{\mathbf{x}})\cdot(\mathbf{x}_t - \bar{\mathbf{x}})^T = \frac{1}{M}\sum_{t=1}^{M}\mathbf{x}_t^*\cdot\mathbf{x}_t^{*T} \tag{5.1}$$

Nous définissons également les matrices de covariance décalées $\boldsymbol{\mathcal{X}}_k$ par :

$$\boldsymbol{\mathcal{X}}_k = \frac{1}{M}\sum_{t=k+1}^{M}\mathbf{x}_t^*\cdot\mathbf{x}_{t-k}^{*T} \quad \text{avec} \quad k = 1, ..., q \tag{5.2}$$

et la matrice bloc-Toeplitz $\mathbf{X}_q$:

$$\mathbf{X}_q = \begin{bmatrix} \boldsymbol{\mathcal{X}}_0 & \boldsymbol{\mathcal{X}}_1 & ... & \boldsymbol{\mathcal{X}}_q \\ \boldsymbol{\mathcal{X}}_1^T & \boldsymbol{\mathcal{X}}_0 & ... & \boldsymbol{\mathcal{X}}_{q-1} \\ \vdots & \vdots & & \vdots \\ \boldsymbol{\mathcal{X}}_q^T & \boldsymbol{\mathcal{X}}_{q-1}^T & ... & \boldsymbol{\mathcal{X}}_0 \end{bmatrix} \tag{5.3}$$

Un modèle AR-vectoriel d'ordre q de la séquence $\{\mathbf{x}_t^*\}$ s'écrit classiquement :

$$\boxed{\sum_{i=0}^{q}\mathbf{A}_i\cdot\mathbf{x}_{t-i}^* = \mathbf{e}_t \quad \text{avec} \quad \mathbf{A}_0 = \mathbf{I}_p} \tag{5.4}$$

où $\{\mathbf{A}_i\}$ est un ensemble de $q+1$ coefficients matriciels de prédiction linéaire, et où $\mathbf{e}_t$ est le vecteur d'erreur de prédiction. Les coefficients matriciels $\{\mathbf{A}_1, ..., \mathbf{A}_q\}$ sont obtenus en résolvant l'équation de Yule-Walker vectorielle [166]. Le détail du calcul des coefficients matriciels dans le cas d'un modèle AR-vectoriel d'ordre 2 est donné dans l'Annexe C.

En notant $\mathbf{A} = [\mathbf{A}_0 \; ... \; \mathbf{A}_q]$, la matrice de covariance du résiduel de la séquence $\{\mathbf{x}_t^*\}$ filtrée par $\mathbf{A}$ s'écrit :

$$\mathbf{E}_{\mathbf{X}_q}^{(\mathbf{A})} = \mathbf{A}\mathbf{X}_q\mathbf{A}^T \tag{5.5}$$

De façon similaire, pour une séquence $\{\mathbf{y}_t\}_{1\leq t\leq N}$ avec un modèle $\mathbf{B}$, nous obtenons :

$$\mathbf{E}_{\mathbf{Y}_q}^{(\mathbf{B})} = \mathbf{B}\mathbf{Y}_q\mathbf{B}^T \tag{5.6}$$

Si nous considérons maintenant :

$$\mathbf{E}_{\mathbf{X}_q}^{(\mathbf{B})} = \mathbf{B}\mathbf{X}_q\mathbf{B}^T \tag{5.7}$$

$$\mathbf{E}_{\mathbf{Y}_q}^{(\mathbf{A})} = \mathbf{A}\mathbf{Y}_q\mathbf{A}^T \tag{5.8}$$

ces matrices peuvent être interprétées comme la matrice de covariance du filtrage de $\{\mathbf{x}_t^*\}$ par $\mathbf{B}$, et comme la matrice de covariance du filtrage de $\{\mathbf{y}_t^*\}$ par $\mathbf{A}$. Comme $\mathbf{A}$ est obtenu en minimisant $tr(\mathbf{E}_{\mathbf{X}_q}^{(\mathbf{A})})$ et $\mathbf{B}$ en minimisant $tr(\mathbf{E}_{\mathbf{Y}_q}^{(\mathbf{B})})$ [166], nous avons $tr(\mathbf{E}_{\mathbf{X}_q}^{(\mathbf{B})}) \geq tr(\mathbf{E}_{\mathbf{X}_q}^{(\mathbf{A})})$ et $tr(\mathbf{E}_{\mathbf{Y}_q}^{(\mathbf{A})}) \geq tr(\mathbf{E}_{\mathbf{Y}_q}^{(\mathbf{B})})$.

Notons pour finir, par analogie avec l'équation 4.3, $\Gamma_{\mathbf{X}_q}^{(\mathbf{B}/\mathbf{A})}$ et $\Gamma_{\mathbf{Y}_q/\mathbf{X}_q}^{(\mathbf{A})}$ les deux matrices suivantes :

$$\Gamma_{\mathbf{X}_q}^{(\mathbf{B}/\mathbf{A})} = \left(\mathbf{E}_{\mathbf{X}_q}^{(\mathbf{A})}\right)^{-\frac{1}{2}} \cdot \mathbf{E}_{\mathbf{X}_q}^{(\mathbf{B})} \cdot \left(\mathbf{E}_{\mathbf{X}_q}^{(\mathbf{A})}\right)^{-\frac{1}{2}} \tag{5.9}$$

$$\Gamma_{\mathbf{Y}_q/\mathbf{X}_q}^{(\mathbf{A})} = \left(\mathbf{E}_{\mathbf{X}_q}^{(\mathbf{A})}\right)^{-\frac{1}{2}} \cdot \mathbf{E}_{\mathbf{Y}_q}^{(\mathbf{A})} \cdot \left(\mathbf{E}_{\mathbf{X}_q}^{(\mathbf{A})}\right)^{-\frac{1}{2}} \tag{5.10}$$

où $\mathbf{E}^{\frac{1}{2}}$ est la racine carrée symétrique de $\mathbf{E}$.

La première matrice peut être interprétée comme la matrice de covariance de la séquence $\{\mathbf{x}_t^*\}$ filtrée par $\mathbf{B}$ relativement à celle de $\{\mathbf{x}_t^*\}$ filtrée par $\mathbf{A}$, et la deuxième comme la matrice de covariance de la séquence $\{\mathbf{y}_t^*\}$ filtrée par $\mathbf{A}$ relativement à celle de $\{\mathbf{x}_t^*\}$ filtrée par $\mathbf{A}$.

On trouve Fig. 5.1 une récapitulation des différentes possibilités d'obtenir des matrices de covariances normalisées, à partir desquelles on peut construire des mesures de similarité.

5.4 Modèles de locuteurs

Dans nos expériences, nous avons choisi de caractériser les locuteurs par des modèles AR-vectoriels d'ordre 2. Cet ordre a été retenu pour pouvoir comparer nos résultats à ceux qui sont publiés dans la littérature, et qui portent le plus souvent sur des modèles AR-vectoriels d'ordre 2. Néanmoins, cette approche reste valable pour un autre ordre du modèle. Dans le cas $q = 2$, les coefficients matriciels $\{\mathbf{A}_1, \mathbf{A}_2\}$ sont obtenus en résolvant (cf. Annexe C) :

$$[\mathbf{A}_1\ \mathbf{A}_2] \cdot \begin{bmatrix} \boldsymbol{\mathcal{X}}_0 & \boldsymbol{\mathcal{X}}_1 \\ \boldsymbol{\mathcal{X}}_1^T & \boldsymbol{\mathcal{X}}_0 \end{bmatrix} = -\left[\boldsymbol{\mathcal{X}}_1^T\ \boldsymbol{\mathcal{X}}_2^T\right] \tag{5.11}$$

- Le premier modèle utilisé est un modèle AR-vectoriel d'ordre 2, entrainé sur les trames acoustiques présentées dans leur ordre temporel naturel. Dans ce cas, le modèle du locuteur $\mathcal{X}$ est noté $\{\mathbf{A}, \mathbf{X}_2\}$.

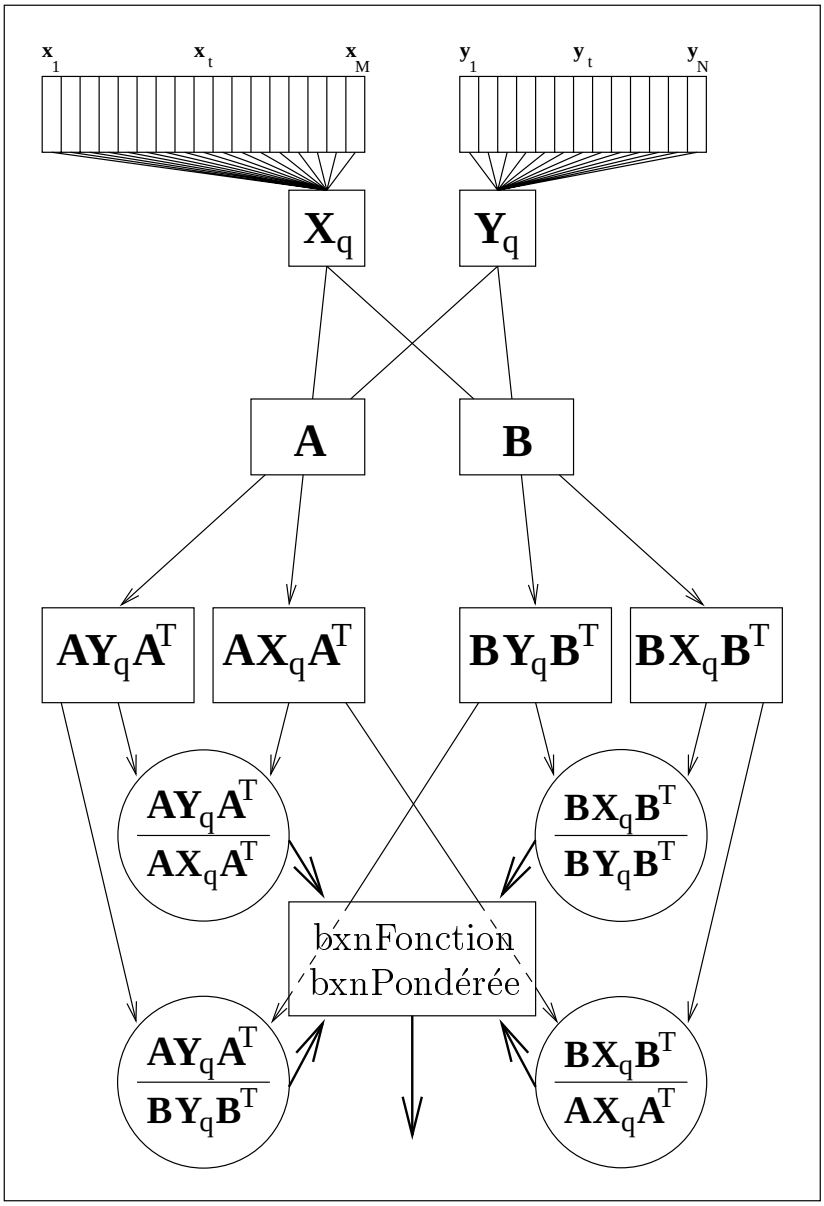


FIG. 5.1: *Schéma explicatif sur la construction des matrices de covariance d'erreurs résiduelles normalisées.*

- Le second modèle est également un modèle AR-vectoriel d'ordre 2, mais entrainé cette fois-ci sur les trames acoustiques mélangées aléatoirement. Si les modèles AR-vectoriels sont effectivement une façon de modéliser les caractéristiques dynamiques du locuteur, ce protocole expérimental devrait dégrader les performances. Pour le locuteur $\mathcal{X}$, ce modèle est noté $\{\mathbf{A}', \mathbf{X}'_2\}$.
- Le troisième modèle est le modèle mono-Gaussien, qui a déjà été étudié en détails dans le chapitre précédent. Ce modèle sert ici de modèle de référence. Le locuteur $\mathcal{X}$ est alors représenté uniquement par $\boldsymbol{\mathcal{X}}_0$ comme nous l'avons déjà vu. En fait, le modèle mono-Gaussien peut s'interpréter comme un modèle AR-vectoriel d'ordre 0, i.e. $\mathbf{A} = [\mathbf{A}_0] = \mathbf{I}_p$ et $\mathbf{X}_0 = [\boldsymbol{\mathcal{X}}_0]$. Ainsi nous notons ce modèle $\{\mathbf{I}, \mathbf{X}_0\}$.

5.5 Mesures de similarité

Considérons maintenant deux locuteurs $\mathcal{X}$ et $\mathcal{Y}$. Nous présentons un formalisme général pour exprimer des mesures de similarité entre leurs modèles AR-vectoriels. Deux familles de mesures de similarité sont étudiées :

$$f^{(\mathbf{B}/\mathbf{A})}_{\mathbf{X}_q}(\mathcal{X}, \mathcal{Y}) = f\left(\Gamma^{(\mathbf{B}/\mathbf{A})}_{\mathbf{X}_q}\right) \tag{5.12}$$

$$f^{(\mathbf{A})}_{\mathbf{Y}_q/\mathbf{X}_q}(\mathcal{X}, \mathcal{Y}) = f\left(\Gamma^{(\mathbf{A})}_{\mathbf{Y}_q/\mathbf{X}_q}\right) \tag{5.13}$$

La première famille peut être interprétée comme une mesure entre deux modèles ($\mathbf{A}$ et $\mathbf{B}$), à travers leur influence sur le même locuteur $\mathcal{X}$, ou plus exactement sur sa matrice bloc-Toeplitz $\mathbf{X}_q$. Cette famille de mesures (à laquelle nous réfèrerons par IV pour Itakura Vectoriel), généralise la mesure d'Itakura au cas vectoriel [71]. Des exemples de cette famille de mesures ont été proposés dans [12] et [57].

La deuxième famille de mesures, quant à elle, peut être vue comme une mesure entre deux locuteurs différents $\mathcal{X}$ et $\mathcal{Y}$, ou plus exactement entre leur matrice bloc-Toeplitz $\mathbf{X}_q$ et $\mathbf{Y}_q$ filtrées à travers le même modèle $\mathbf{A}$. Nous appellerons ces mesures SO pour Second Ordre. Certaines des mesures IS proposées dans [106] et [105] appartiennent à cette seconde famille.

Notons pour finir qu'en prenant le modèle $\{\mathbf{A}, \mathbf{X}\} = \{\mathbf{I}, \mathbf{X}_0\}$ du modèle de locuteur mono-Gaussien, et en appliquant de façon similaire la seconde famille de mesures proposée ci-dessus, nous retrouvons la formulation que

nous avons appliquée à ce modèle au chapitre précédent.

Par analogie avec le chapitre précédent, la fonction f est choisie comme étant une combinaison des différentes quantités canoniques suivantes :

$$\begin{aligned} a(\mathbf{\Gamma}) &= \frac{1}{p}\, tr(\mathbf{\Gamma}) \\ g(\mathbf{\Gamma}) &= [det(\mathbf{\Gamma})]^{\frac{1}{p}} \end{aligned} \tag{5.14}$$

Nous avons déjà vu que a et g étaient positives, et que nous avions $a \geq g$ (cf. équations 4.11). Nous savons également que ces quantités peuvent être calculées très simplement (cf. équation 4.15 et 4.16).

Les fonctions composées $a - \log g - 1$ et $\log(a/g)$ sont respectivement la mesure de maximum de vraisemblance (cf. équation 4.25) et le test de sphéricité (cf. équation 4.37). Nous testons aussi la mesure $a - g$, qui est une combinaison des quantités a et g, et qui possède les propriétés requises pour ce type de mesures.

Nous appliquons finalement les mêmes symétrisations que celles qui ont été étudiées dans le chapitre précédent (cf. section 4.6) :

$$f_{\mathbf{X}_q}^{(\mathbf{B}/\mathbf{A})^\star} = \frac{1}{2}\, f_{\mathbf{X}_q}^{(\mathbf{B}/\mathbf{A})} + \frac{1}{2}\, f_{\mathbf{Y}_q}^{(\mathbf{A}/\mathbf{B})} \tag{5.15}$$

$$f_{\mathbf{X}_q}^{(\mathbf{B}/\mathbf{A})^\diamond} = \frac{M}{M+N}\, f_{\mathbf{X}_q}^{(\mathbf{B}/\mathbf{A})} + \frac{N}{M+N}\, f_{\mathbf{Y}_q}^{(\mathbf{A}/\mathbf{B})} \tag{5.16}$$

$$f_{\mathbf{X}_q}^{(\mathbf{B}/\mathbf{A})^\bullet} = \frac{N}{M+N}\, f_{\mathbf{X}_q}^{(\mathbf{B}/\mathbf{A})} + \frac{M}{M+N}\, f_{\mathbf{Y}_q}^{(\mathbf{A}/\mathbf{B})} \tag{5.17}$$

M est le nombre de trames pour la phrase d'apprentissage, et N celui de la phrase de test. Les mêmes symétrisations sont appliquées à $f_{\mathbf{Y}_q}^{(\mathbf{A}/\mathbf{B})}$, $f_{\mathbf{Y}_q/\mathbf{X}_q}^{(\mathbf{A})}$ et $f_{\mathbf{X}_q/\mathbf{Y}_q}^{(\mathbf{B})}$.

5.6 Expériences et résultats

5.6.1 Description des expériences

5.6.1.a Tâche

La tâche évaluée ici est l'identification du locuteur indépendante du texte en ensemble fermé. Pour chaque mesure, sous une forme symétrique

ou non, nous désignons pour un test donné l'identité du locuteur de la base de référence qui est le plus proche au sens de la mesure testée. Les résultats sont donnés en pourcentages d'erreurs d'identification.

5.6.1.b Bases de données

Les résultats sont reportés pour les bases TIMIT63, FTIMIT63 et NTIMIT63 (cf. chapitre 3).

5.6.1.c Analyse acoustique du signal

Cette analyse acoustique est décrite avec précision dans le chapitre 3. Nous avons choisi des fenêtres de 504 échantillons (31,5 ms).

5.6.1.d Protocole expérimental

Nous avons testé cette fois-ci un seul protocole expérimental, le protocole 5.1 (apprentissage long - test court).

5.6.2 Résultats

Les résultats sont donnés par base de données (TAB. 5.1 et TAB. 5.2 pour TIMIT63, TAB. 5.3 et TAB. 5.4 pour FTIMIT63, TAB 5.5 et TAB 5.6 pour NTIMIT63). Les performances sont données en pourcentage d'erreurs d'identification. Nous donnons les résultats pour chaque mesure canonique et pour chaque mesure combinée dans leurs formes asymétriques et dans leur meilleure forme symétrique. Pour les versions symétriques, un exposant indique de quelle symétrisation il s'agit ($\star$, $\diamond$ or $\bullet$).

function f	a	$\log a$	g	$\log g$
Modèles AR–vectoriels - Trames acoustiques dans leur ordre temporel naturel				
$f_{\mathbf{X}}^{(\mathbf{B/A})} \mid f_{\mathbf{Y}}^{(\mathbf{A/B})}$	16.8 | 8.6	16.8 | 8.6	16.2 | 7.6	16.2 | 7.6
symétrisée	3.5 $^{\bullet}$	4.1 $^{\bullet}$	4.1 $^{\bullet}$	4.1 $^{\bullet}$
$f_{\mathbf{Y/X}}^{(\mathbf{A})} \mid f_{\mathbf{X/Y}}^{(\mathbf{B})}$	75.6 | 51.4	75.6 | 51.4	88.3 | 73.0	88.3 | 73.0
symétrisée	6.0 *	4.8 *	12.4 *	4.8 *
Modèles AR–vectoriels - Trames acoustiques dans un ordre temporel aléatoire				
$f_{\mathbf{X}'}^{(\mathbf{B}'/\mathbf{A}')} \mid f_{\mathbf{Y}'}^{(\mathbf{A}'/\mathbf{B}')}$	2.5 | 56.5	2.5 | 56.5	4.1 | 58.1	4.1 | 58.1
symétrisée	3.5 $^{\diamond}$	3.5 $^{\diamond}$	5.7 $^{\diamond}$	5.7 $^{\diamond}$
$f_{\mathbf{Y}'/\mathbf{X}'}^{(\mathbf{A}')} \mid f_{\mathbf{X}'/\mathbf{Y}'}^{(\mathbf{B}')}$	42.5 | 45.4	42.5 | 45.4	98.1 | 82.9	98.1 | 82.9
symétrisée	4.8 *	2.2 *	46.7 *	12.7 *
Modèle Gaussien				
$f_{\mathbf{Y}_o/\mathbf{X}_o}^{(\mathbf{I})} \mid f_{\mathbf{X}_o/\mathbf{Y}_o}^{(\mathbf{I})}$	37.5 | 47.0	37.5 | 47.0	98.4 | 98.4	98.4 | 98.4
symétrisée	3.8 *	**1.3** *	97.1 *	99.4 *

TAB. 5.1: *Modèles AR-vectoriels. Identification du locuteur indépendante du texte. Base de données TIMIT63. Mesures simples. Les résultats sont donnés en pourcentages d'erreurs d'identification.*

function f	$a - \log g - 1$	$\log(a/g)$	$a - g$
Modèles AR–vectoriels - Trames acoustiques dans leur ordre temporel naturel			
$f_{\mathbf{X}}^{(\mathbf{B/A})} \mid f_{\mathbf{Y}}^{(\mathbf{A/B})}$	19.1 | 10.8	23.8 | 19.4	22.2 | 17.5
symétrisée	3.2 $^{\bullet}$	7.9 $^{\bullet}$	7.3 $^{\bullet}$
$f_{\mathbf{Y/X}}^{(\mathbf{A})} \mid f_{\mathbf{X/Y}}^{(\mathbf{B})}$	15.2 | 34.3	7.6 | 18.7	15.2 | 14.6
symétrisée	5.4 $^{\diamond}$	7.0 $^{\diamond}$	6.0 $^{\diamond}$
Modèles AR–vectoriels - Trames acoustiques dans un ordre temporel aléatoire			
$f_{\mathbf{X}'}^{(\mathbf{B}'/\mathbf{A}')} \mid f_{\mathbf{Y}'}^{(\mathbf{A}'/\mathbf{B}')}$	2.5 | 56.2	4.1 | 55.9	3.5 | 54.6
symétrisée	2.5 $^{\diamond}$	4.1 $^{\diamond}$	4.1 $^{\diamond}$
$f_{\mathbf{Y}'/\mathbf{X}'}^{(\mathbf{A}')} \mid f_{\mathbf{X}'/\mathbf{Y}'}^{(\mathbf{B}')}$	**1.3** | 22.9	**1.0** | 6.7	3.2 | 8.9
symétrisée	2.9 $^{\diamond}$	**1.0** $^{\diamond}$	**1.6** $^{\diamond}$
Modèle Gaussien			
$f_{\mathbf{Y}_o/\mathbf{X}_o}^{(\mathbf{I})} \mid f_{\mathbf{X}_o/\mathbf{Y}_o}^{(\mathbf{I})}$	**0.6** | 7.9	**0.6** | 3.2	2.9 | 6.4
symétrisée	**1.0** $^{\diamond}$	**0.6** $^{\diamond}$	**1.0** $^{\diamond}$

TAB. 5.2: *Modèles AR-vectoriels. Identification du locuteur indépendante du texte. Base de données TIMIT63. Mesures composées. Les résultats sont donnés en pourcentages d'erreurs d'identification.*

function f	a	$\log a$	g	$\log g$
Modèles AR–vectoriels - Trames acoustiques dans leur ordre temporel naturel				
$f^{(\mathbf{B/A})}_{\mathbf{X}} \mid f^{(\mathbf{A/B})}_{\mathbf{Y}}$	38.7 | 30.2	38.7 | 30.2	37.1 | 29.5	37.1 | 29.5
symétrisée	24.8 $^{\bullet}$	25.1 $^{\bullet}$	24.8 $^{\bullet}$	24.4 $^{\bullet}$
$f^{(\mathbf{A})}_{\mathbf{Y/X}} \mid f^{(\mathbf{B})}_{\mathbf{X/Y}}$	93.3 | 86.0	93.3 | 86.0	96.5 | 94.6	96.5 | 94.6
symétrisée	23.5 *	21.3 *	32.4 *	25.4 *
Modèles AR–vectoriels - Trames acoustiques dans un ordre temporel aléatoire				
$f^{(\mathbf{B'/A'})}_{\mathbf{X'}} \mid f^{(\mathbf{A'/B'})}_{\mathbf{Y'}}$	35.9 | 82.2	35.9 | 82.2	36.8 | 81.3	36.8 | 81.3
symétrisée	39.1 $^{\diamond}$	39.1 $^{\diamond}$	40.0 $^{\diamond}$	40.0 $^{\diamond}$
$f^{(\mathbf{A'})}_{\mathbf{Y'/X'}} \mid f^{(\mathbf{B'})}_{\mathbf{X'/Y'}}$	78.7 | 71.4	78.7 | 71.4	98.4 | 93.7	98.4 | 93.7
symétrisée	21.9 *	**14.6** *	69.8 *	52.4 *
Modèle Gaussien				
$f^{(\mathbf{I})}_{\mathbf{Y}_o/\mathbf{X}_o} \mid f^{(\mathbf{I})}_{\mathbf{X}_o/\mathbf{Y}_o}$	77.1 | 71.8	77.1 | 71.8	98.4 | 98.4	98.4 | 98.4
symétrisée	15.6 *	**11.8** *	97.8 *	98.4 *

TAB. 5.3: *Modèles AR-vectoriels. Identification du locuteur indépendante du texte. Base de données FTIMIT63. Mesures simples. Les résultats sont donnés en pourcentages d'erreurs d'identification.*

function f	$a - \log g - 1$	$\log(a/g)$	$a - g$
Modèles AR–vectoriels - Trames acoustiques dans leur ordre temporel naturel			
$f^{(\mathbf{B/A})}_{\mathbf{X}} \mid f^{(\mathbf{A/B})}_{\mathbf{Y}}$	42.5 | 35.2	51.1 | 50.8	49.5 | 49.5
symétrisée	26.3 $^{\bullet}$	35.6 $^{\bullet}$	33.3 $^{\bullet}$
$f^{(\mathbf{A})}_{\mathbf{Y/X}} \mid f^{(\mathbf{B})}_{\mathbf{X/Y}}$	44.1 | 69.8	41.6 | 39.1	49.2 | 39.1
symétrisée	24.4 $^{\diamond}$	34.6 $^{\diamond}$	33.0 $^{\diamond}$
Modèles AR–vectoriels - Trames acoustiques dans un ordre temporel aléatoire			
$f^{(\mathbf{B'/A'})}_{\mathbf{X'}} \mid f^{(\mathbf{A'/B'})}_{\mathbf{Y'}}$	32.4 | 83.5	34.6 | 82.2	34.3 | 81.6
symétrisée	34.3 $^{\diamond}$	33.3 $^{\diamond}$	33.3 $^{\diamond}$
$f^{(\mathbf{A'})}_{\mathbf{Y'/X'}} \mid f^{(\mathbf{B'})}_{\mathbf{X'/Y'}}$	15.9 | 43.8	**13.3** | 21.6	20.3 | 27.3
symétrisée	**14.0** $^{\diamond}$	**13.3** $^{\diamond}$	**14.3** $^{\diamond}$
Modèle Gaussien			
$f^{(\mathbf{I})}_{\mathbf{Y}_o/\mathbf{X}_o} \mid f^{(\mathbf{I})}_{\mathbf{X}_o/\mathbf{Y}_o}$	**14.6** | 27.3	**12.7** | 17.1	20.3 | 21.3
symétrisée	**12.7** $^{\diamond}$	**12.4** $^{\diamond}$	**14.3** $^{\diamond}$

TAB. 5.4: *Modèles AR-vectoriels. Identification du locuteur indépendante du texte. Base de données FTIMIT63. Mesures composées. Les résultats sont donnés en pourcentages d'erreurs d'identification.*

function f	a	$\log a$	g	$\log g$
Modèles AR–vectoriels - Trames acoustiques dans leur ordre temporel naturel				
$f_{\mathbf{X}}^{(\mathbf{B/A})} \mid f_{\mathbf{Y}}^{(\mathbf{A/B})}$	71.8 | 54.6	71.8 | 54.6	67.3 | 54.3	67.3 | 54.3
symétrisée	51.8 $^{\bullet}$	52.1 $^{\bullet}$	50.5 $^{\bullet}$	50.2 $^{\bullet}$
$f_{\mathbf{Y/X}}^{(\mathbf{A})} \mid f_{\mathbf{X/Y}}^{(\mathbf{B})}$	96.8 | 92.4	96.8 | 92.4	97.1 | 95.6	97.1 | 95.6
symétrisée	61.9 *	56.5 *	68.3 *	53.0 *
Modèles AR–vectoriels - Trames acoustiques dans un ordre temporel aléatoire				
$f_{\mathbf{X}'}^{(\mathbf{B}'/\mathbf{A}')} \mid f_{\mathbf{Y}'}^{(\mathbf{A}'/\mathbf{B}')}$	64.4 | 92.1	64.1 | 92.1	65.4 | 91.8	65.4 | 91.8
symétrisée	65.4 $^{\diamond}$	65.1 $^{\diamond}$	67.9 $^{\diamond}$	68.3 $^{\diamond}$
$f_{\mathbf{Y}'/\mathbf{X}'}^{(\mathbf{A}')} \mid f_{\mathbf{X}'/\mathbf{Y}'}^{(\mathbf{B}')}$	94.0 | 94.3	94.0 | 94.3	98.4 | 97.5	98.4 | 97.5
symétrisée	61.9 *	52.4 *	88.3 *	72.4 *
Modèle Gaussien				
$f_{\mathbf{Y}_o/\mathbf{X}_o}^{(\mathbf{I})} \mid f_{\mathbf{X}_o/\mathbf{Y}_o}^{(\mathbf{I})}$	93.0 | 94.6	93.0 | 94.6	98.4 | 98.4	98.4 | 98.4
symétrisée	58.1 *	**49.8** *	97.8 *	98.4 *

TAB. 5.5: *Modèles AR-vectoriels. Identification du locuteur indépendante du texte. Base de données NTIMIT63. Mesures simples. Les résultats sont donnés en pourcentages d'erreurs d'identification.*

function f	$a - \log g - 1$	$\log(a/g)$	$a - g$
Modèles AR–vectoriels - Trames acoustiques dans leur ordre temporel naturel			
$f_{\mathbf{X}}^{(\mathbf{B/A})} \mid f_{\mathbf{Y}}^{(\mathbf{A/B})}$	78.1 | 58.4	83.8 | 69.5	82.9 | 67.9
symétrisée	57.5 $^{\bullet}$	66.0 $^{\bullet}$	65.1 $^{\bullet}$
$f_{\mathbf{Y/X}}^{(\mathbf{A})} \mid f_{\mathbf{X/Y}}^{(\mathbf{B})}$	67.3 | 88.9	66.0 | 78.7	75.2 | 76.8
symétrisée	59.7 $^{\diamond}$	63.2 $^{\diamond}$	66.4 $^{\diamond}$
Modèles AR–vectoriels - Trames acoustiques dans un ordre temporel aléatoire			
$f_{\mathbf{X}'}^{(\mathbf{B}'/\mathbf{A}')} \mid f_{\mathbf{Y}'}^{(\mathbf{A}'/\mathbf{B}')}$	61.9 | 92.4	64.8 | 93.3	64.4 | 93.0
symétrisée	62.2 $^{\diamond}$	64.4 $^{\diamond}$	64.1 $^{\diamond}$
$f_{\mathbf{Y}'/\mathbf{X}'}^{(\mathbf{A}')} \mid f_{\mathbf{X}'/\mathbf{Y}'}^{(\mathbf{B}')}$	**47.0** | 86.4	**46.0** | 63.2	56.8 | 77.1
symétrisée	50.2 $^{\diamond}$	**44.1** $^{\diamond}$	**48.6** $^{\diamond}$
Modèle Gaussien			
$f_{\mathbf{Y}_o/\mathbf{X}_o}^{(\mathbf{I})} \mid f_{\mathbf{X}_o/\mathbf{Y}_o}^{(\mathbf{I})}$	**44.1** | 75.9	**42.5** | 59.7	56.2 | 73.3
symétrisée	**47.6** $^{\diamond}$	**44.1** $^{\diamond}$	**49.2** $^{\diamond}$

TAB. 5.6: *Modèles AR-vectoriels. Identification du locuteur indépendante du texte. Base de données NTIMIT63. Mesures composées. Les résultats sont donnés en pourcentages d'erreurs d'identification.*

5.7 Discussion

Plusieurs observations peuvent être faites à partir de ces résultats :

- La symétrisation est le plus souvent un facteur d'amélioration. Cependant, la symétrisation la plus appropriée est difficile à prévoir. Elle dépend du type de mesure asymétrique, et aussi du fait que les trames vectorielles ont été présentées dans leur ordre temporel naturel, ou bien dans un ordre aléatoire.
- Pour chaque base de données (TIMIT63, FTIMIT63 et NTIMIT63), nous avons souligné les 10 meilleures mesures (ou 11 meilleures dans un cas). Ce sont quasiment les mêmes pour les trois bases de données. La meilleure performance est toujours obtenue avec le modèle du locuteur mono-Gaussien.
- Lorsque les trames acoustiques sont présentées dans leur ordre naturel, les mesures canoniques de type IV donnent en général de meilleures performances que les mesures canoniques de type SO . La tendance est inversée pour les mesures composées.
- Lorsque les trames acoustiques sont présentées dans un ordre aléatoire, les mesures composées symétriques de type SO ont des performances meilleures que toutes les autres mesures effectuées sur les modèles AR-vectoriels, et ceci en dépit de la destruction des caractéristiques spectrales dynamiques.

5.8 Synthèse sur les modèles AR-vectoriels

☞ Les modèles AR-vectoriels ne nous ont pas permis d'obtenir de meilleures performances que le modèle mono-Gaussien. Cette conclusion est en contradiction avec les résultats publiés dans [83], mais cette différence peut être due à un prétraitement et une analyse acoustique du signal différents. Cependant, cette analyse n'est pas décrite dans l'article précédent.

☞ D'autre part, nous obtenons globalement de meilleures performances avec les modèles AR-vectoriels en identification du locuteur lorsque ces derniers sont entrainés sur les trames acoustiques présentées dans un ordre temporel aléatoire.

> Il ne nous a pas été possible de mettre en évidence le rôle joué par les caractéristiques dynamiques dans les performances des modèles AR-vectoriels en identification du locuteur.

Nos résultats suggèrent plutôt que les modèles AR-vectoriels extraient indirectement des caractéristiques du locuteur de nature statique.

☞ Enfin, nous avons vu que ces modèles AR-vectoriels pouvaient être interprétés en terme de filtrage dépendant du locuteur des trames acoustiques vectorielles.

Nous venons de voir que les modèles AR-vectoriels n'apportaient pas une solution satisfaisante à la prise en compte des caractéristiques dynamiques du locuteur. Nous proposons donc dans les chapitres suivants une autre solution.

Chapitre 6

Filtrage vectoriel de trajectoires spectrales

Les modèles AR-vectoriels ne nous ont pas fournit une réponse satisfaisante au problème de la prise en compte des caractéristiques dynamiques du locuteur. Cependant, l'approche par filtrage vectoriel des séquences de vecteurs de paramètres nous semble intéressante. Dans ce chapitre, nous nous proposons de formaliser cette approche dans le cas le plus général possible. Le chapitre suivant fournit alors un autre exemple de filtrage vectoriel.

Le formalisme que nous établissons est suffisamment général pour englober à la fois les modèles AR-vectoriels [56], l'analyse cepstrale [115], les paramètres Δ et $\Delta\Delta$ [46], les paramètres RASTA [65], la transformée en cosinus de trajectoires spectrales [13], ...

6.1 Principe

Le principe du filtrage vectoriel de trajectoires spectrales est de remplacer le vecteur $\mathbf{x}_t$ par un vecteur $\mathbf{f}_t$ dont chaque composante est obtenue par l'application d'une fonction sur les coordonnées du vecteur $\mathbf{x}_t$ et de son contexte. Dans le cas où cette fonction est linéaire, l'opération s'apparente à un produit de convolution, qui peut s'interpréter comme l'application d'un masque temps-fréquence sur une séquence de vecteurs de paramètres (cf. Fig. 6.1).

Comme on peut le voir sur cette figure, le filtrage est appliqué à la fois dans la dimension temporelle et dans la dimension fréquentielle. Il s'agit donc bien de masques temps-fréquence.

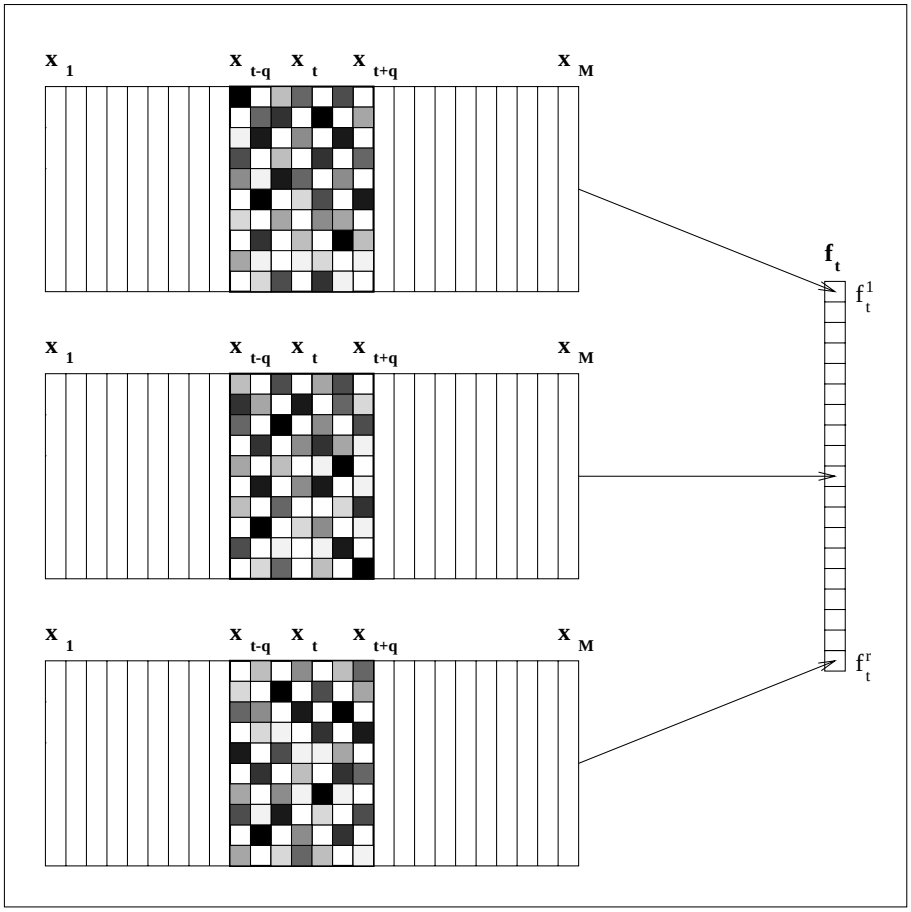


FIG. 6.1: *Principe du filtrage vectoriel de trajectoires spectrales.*

Cette approche généralise un certain nombre d'autres approches de filtrage, présentées récemment dans la littérature, mais qui ne portaient que sur l'une des dimensions temporelle [152], [107], [114] ou fréquentielle [107], [33]. L'avantage de cette approche consiste à combiner ces deux dimensions dans un même filtrage.

Le travail de Milner [104] [103] s'inscrit dans la même direction que le nôtre, à quelques différences près. Tout d'abord, il adopte les coefficients cepstraux comme base de son travail, tandis que nous travaillons directement sur les coefficients de banc de filtres. D'autre part, il se limite à l'application de fonctions dans la direction temporelle, n'abordant pas le filtrage dans la dimension fréquentielle (ou quéfrentielle puisqu'il utilise des coefficients cepstraux). Ce travail fait apparaître néanmoins le même soucis d'unification et de généralisation d'approches déjà existantes, et propose également dans ce nouveau cadre quelques transformations nouvelles.

La FIG. 6.2 illustre plusieurs filtrages vectoriels particulier : le cas d'un paramètre Δ, qui n'opère que dans la dimension temporelle; le cas d'un paramètre cepstral, qui n'opère que dans la dimension fréquentielle, le cas d'une transformée en cosinus de trajectoires spectrales qui n'opère que dans la dimension temporelle. Il faut bien comprendre que chaque masque fournit une seule coordonnée du nouveau vecteur $\mathbf{f}_t$. Ainsi, dans l'exemple de l'analyse cepstrale, chaque coefficient cepstral est obtenu par l'application d'une fonction sinusoïdale différente. Si on se reporte de nouveau à la FIG. 6.1, on trouve trois exemples de masques agissant dans les deux dimensions temporelle et fréquentielle.

Ce chapitre présente un formalisme pour ce type de filtrage, et donne quelques exemples. Le chapitre 7 propose une façon particulière de choisir les masques temps-fréquence.

6.2 Définitions et notations

6.2.1 Cas général

Soit $\mathcal{X}$ un locuteur (un des locuteur de la base de référence par exemple). Le long de la phrase qu'il a prononcée, on extrait une séquence de vecteurs de paramètres $\{\mathbf{x}_t\}_{1\leq t\leq M}$ de dimension p. Nous définissons également la séquence de vecteurs centrés correspondante, $\{\mathbf{x}_t^* = \mathbf{x}_t - \bar{\mathbf{x}}\}$, où $\bar{\mathbf{x}}$ est le vecteur moyen de la séquence $\{\mathbf{x}_t\}$.

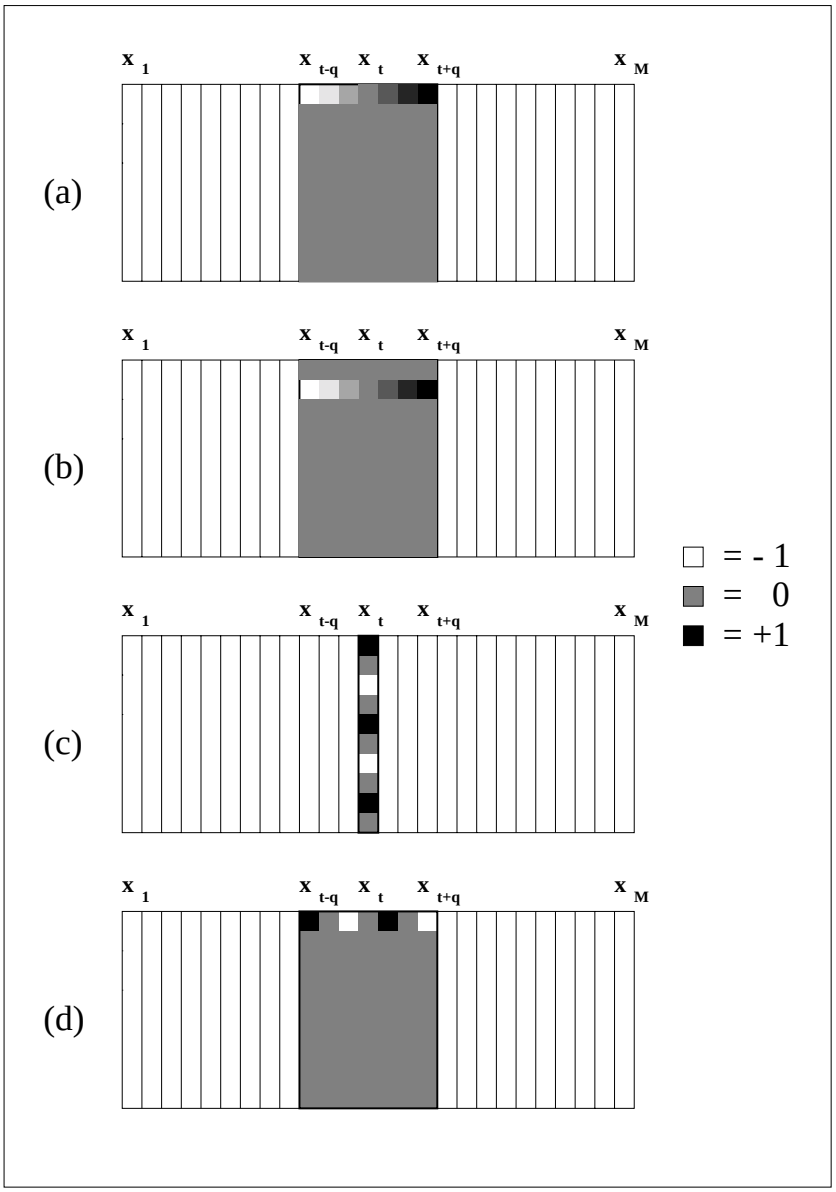


FIG. 6.2: *Quelques exemples de filtrage vectoriel de trajectoires spectrales : (a) paramètre Δ sur la première coordonnée ; (b) paramètre Δ sur la deuxième coordonnée ; (c) coefficient cepstral ; (d) transformée en cosinus de 7 trames consécutives sur la première coordonnée.*

Nous rappelons pour mémoire la définition de la matrice de covariance $\boldsymbol{\mathcal{X}}_0$:

$$\boldsymbol{\mathcal{X}}_0 = \frac{1}{M}\sum_{t=1}^{M}(\mathbf{x}_t - \bar{\mathbf{x}})\cdot(\mathbf{x}_t - \bar{\mathbf{x}})^T = \frac{1}{M}\sum_{t=1}^{M}\mathbf{x}_t^*\cdot\mathbf{x}_t^{*T} \tag{6.1}$$

ainsi que celle des matrices de covariance décalées $\boldsymbol{\mathcal{X}}_k$:

$$\boldsymbol{\mathcal{X}}_k = \frac{1}{M}\sum_{t=k+1}^{M}(\mathbf{x}_t - \bar{\mathbf{x}})\cdot(\mathbf{x}_{t-k} - \bar{\mathbf{x}})^T = \frac{1}{M}\sum_{t=k+1}^{M}\mathbf{x}_t^*\cdot\mathbf{x}_{t-k}^{*T} \tag{6.2}$$

La matrice de covariance ainsi que les matrices de covariance décalées ont pour dimension $p \times p$.

Nous rappelons également la définition de la matrice bloc-Toeplitz d'ordre $2q+1$ correspondante :

$$\mathbf{X}_{2q+1} = \begin{bmatrix} \boldsymbol{\mathcal{X}}_0 & \boldsymbol{\mathcal{X}}_1 & \dots & \boldsymbol{\mathcal{X}}_{2q} \\ \boldsymbol{\mathcal{X}}_1^T & \boldsymbol{\mathcal{X}}_0 & \dots & \boldsymbol{\mathcal{X}}_{2q-1} \\ \vdots & \vdots & & \vdots \\ \boldsymbol{\mathcal{X}}_{2q}^T & \boldsymbol{\mathcal{X}}_{2q-1}^T & \dots & \boldsymbol{\mathcal{X}}_0 \end{bmatrix} \tag{6.3}$$

Elle a pour dimensions $(2q+1)\,p \times (2q+1)\,p$.

Notons $\mathbf{X}_{t-q}^{t+q}$ la séquence des vecteurs $\mathbf{x}_t^*$ entre les temps $t-q$ et $t+q$:

$$\mathbf{X}_{t-q}^{t+q} = \begin{bmatrix} \mathbf{x}_{t+q}^* \\ \vdots \\ \mathbf{x}_t^* \\ \vdots \\ \mathbf{x}_{t-q}^* \end{bmatrix} \tag{6.4}$$

Par convention, un vecteur avec un indice négatif ou nul est pris égal au vecteur nul. De même, $\mathbf{x}_t^* = 0$ pour $t > M$. La dimension du vecteur $\mathbf{X}_{t-q}^{t+q}$ est égale à $(2q+1)\,p$.

Soit alors $\mathcal{H}$ un filtrage pouvant s'appliquer à $\mathbf{X}_{t-q}^{t+q}$:

$$\begin{array}{rcl} \mathcal{H} : (\mathcal{R}^p)^{2q+1} & \longrightarrow & \mathcal{R}^r \\ \mathbf{X}_{t-q}^{t+q} & \longmapsto & \mathbf{f}_t = \mathcal{H}(\mathbf{X}_{t-q}^{t+q}) \end{array} \tag{6.5}$$

$\mathbf{f}_t$ est un vecteur de dimension r.

Dans certains cas, r est un multiple de p. C'est le cas par exemple lorsque le vecteur $\mathbf{f}_t$ est constitué du vecteur $\mathbf{x}_t$ initial, d'une approximation de sa dérivée $\Delta\mathbf{x}_t$, et d'une approximation de sa dérivée seconde $\Delta\Delta\mathbf{x}_t$. Dans ce cas, $r = 3p$.

Nous définissons aussi la séquence des vecteurs $\mathbf{f}_t$ pour $1 \leq t \leq M$ par :

$$\mathcal{H}(X) = [\mathbf{f}_1 \ ... \ \mathbf{f}_M] = \left[\mathcal{H}(\mathbf{X}_{1-q}^{1+q}) \ ... \ \mathcal{H}(\mathbf{X}_{M-q}^{M+q})\right] \tag{6.6}$$

6.2.2 Cas d'un filtrage linéaire

Nous nous intéressons dans la suite uniquement à des filtrages linéaires. Le filtrage $\mathcal{H}$ peut alors s'écrire sous une forme matricielle :

$$\mathbf{H} = [\ \mathbf{H}_{-q} \mid ... \mid \mathbf{H}_0 \mid ... \mid \mathbf{H}_q\ ] \tag{6.7}$$

$\mathbf{H}$ est de dimension $r \times (2q+1)\, p$.

Ainsi nous avons :

$$\begin{array}{rcl} \mathbf{f}_t & = & \mathbf{H} \cdot \mathbf{X}_{t-q}^{t+q} \\ & = & [\ \mathbf{H}_{-q} \mid ... \mid \mathbf{H}_0 \mid ... \mid \mathbf{H}_q\ ] \cdot \begin{bmatrix} \mathbf{x}_{t+q}^* \\ \vdots \\ \mathbf{x}_t^* \\ \vdots \\ \mathbf{x}_{t-q}^* \end{bmatrix} \\ & = & \displaystyle\sum_{k=-q}^{+q} \mathbf{H}_k \cdot \mathbf{x}_{t-k}^* \end{array} \tag{6.8}$$

Remarquons que chaque coefficient matriciel $\mathbf{H}_k$ a pour dimension $r \times p$.

Nous avons également dans ce cas :

$$\mathcal{H}(X) = [\mathbf{f}_1 \ ... \ \mathbf{f}_M] = \mathbf{H} \cdot \left[\mathbf{X}_{1-q}^{1+q} \ ... \ \mathbf{X}_{M-q}^{M+q}\right] \tag{6.9}$$

6.2.3 Interprétation de la matrice de filtrage H

Une interprétation de la matrice de filtrage $\mathbf{H}$ en terme de masque temps-fréquence est donnée FIG. 6.3. Cette interprétation permet de faire le lien entre le principe de filtrage énoncé au début de ce chapitre, et le formalisme mathématique que nous venons de présenter.

6.2.4 Matrice de covariance

Nous pouvons calculer la covariance de la quantité $\mathcal{H}(X)$, qui est en fait la matrice de covariance de la séquence des vecteurs $\{\mathbf{f}_t\}$:

$$\begin{aligned} C_{\mathcal{H}(X)} &= \frac{1}{M}\,\mathcal{H}(X)\cdot\mathcal{H}(X)^T = \frac{1}{M}\sum_{t=1}^{M}\mathbf{f}_t\cdot\mathbf{f}_t^T \\ &= \mathbf{H}\cdot\left(\frac{1}{M}\sum_{t=1}^{M}\mathbf{X}_{t-q}^{t+q}\cdot\mathbf{X}_{t-q}^{t+q\,T}\right)\cdot\mathbf{H}^T \\ &= \mathbf{H}\cdot\mathbf{X}_{2q+1}\cdot\mathbf{H}^T \end{aligned} \tag{6.10}$$

Nous voyons dans cette formule que la matrice de covariance des vecteurs filtrés peut s'exprimer directement en fonction de la matrice de filtrage $\mathbf{H}$ et de la matrice bloc-Toeplitz $\mathbf{X}_{2q+1}$, cette dernière étant la matrice de covariance de la séquence des vecteurs $\{\mathbf{X}_{t-q}^{t+q}\}$.

6.2.5 Filtrage dépendant ou indépendant du locuteur

Remarquons pour finir que le filtrage $\mathcal{H}$ peut également être dépendant du locuteur, ce qui est le cas dans le chapitre précédent. Dans ce cas, nous

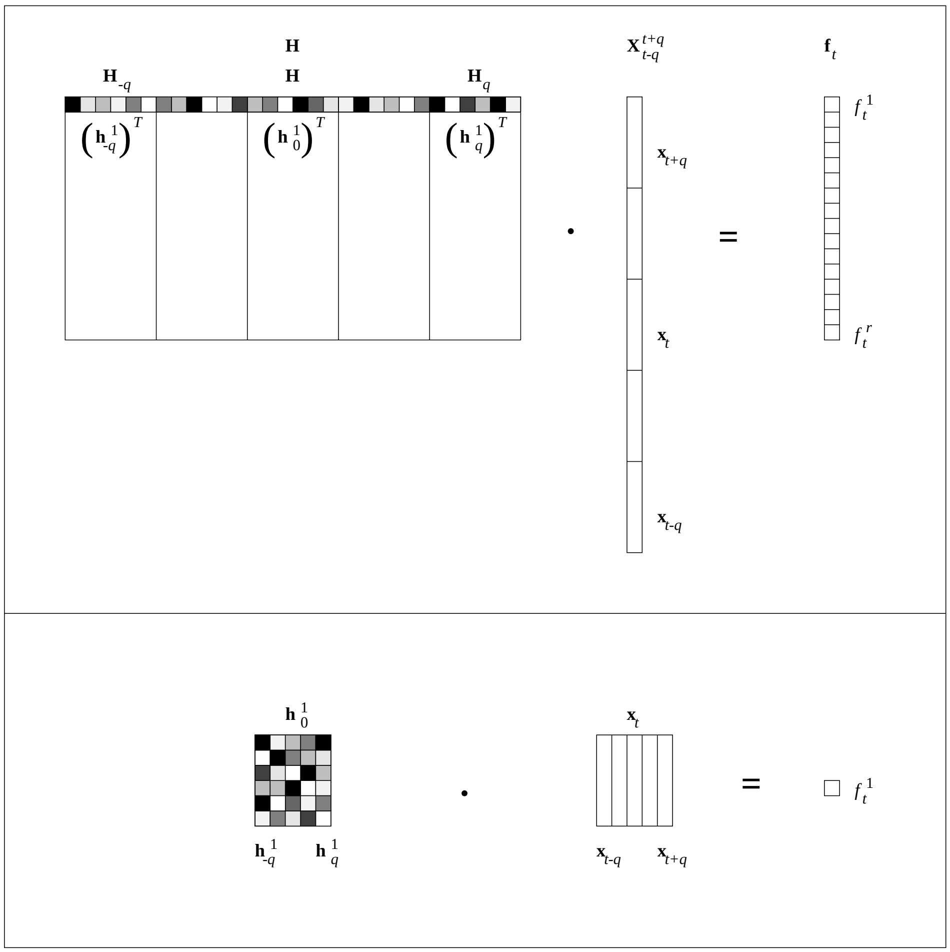


FIG. 6.3: *Interprétation de la matrice de filtrage comme un masque temps-fréquence.*

pouvons noter :

$$\begin{aligned} C_{\mathcal{H}_X(X)} &= \frac{1}{M}\,\mathcal{H}_X(X)\cdot\mathcal{H}_X(X)^T \\ &= \mathbf{H}_X\cdot\mathbf{X}_{2q+1}\cdot\mathbf{H}_X^T \end{aligned} \tag{6.11}$$

6.3 Quelques exemples

6.3.1 Modèle mono-Gaussien

Nous sommes dans le cas où $q = 0$, où $r = p$, et où le filtrage est en fait l'identité, i.e. $\mathbf{H} = \mathbf{I}_p$. On a alors :

$$\mathbf{f}_t = \mathbf{H}\cdot\mathbf{X}_t^t \;=\; \mathbf{H}\cdot\mathbf{x}_t^* \;=\; \mathbf{x}_t^*$$

6.3.2 Modèle AR-vectoriel d'ordre 2

Le modèle AR-vectoriel d'ordre 2 est un exemple de filtrage dépendant du locuteur. Cette fois-ci, $q = 2$, $r = p$, et le filtrage est constitué des coefficients matriciels du modèle AR-vectoriel, i.e. $\mathbf{H}_X = [\,\mathbf{0}_p \mid \mathbf{0}_p \mid \mathbf{I}_p \mid \mathbf{A}_1 \mid \mathbf{A}_2\,]$. On a donc :

$$\mathbf{f}_t = \mathbf{H}_X\cdot\mathbf{X}_{t-2}^{t+2} \;=\; \mathbf{x}_t^* + \mathbf{A}_1\cdot\mathbf{x}_{t-1}^* + \mathbf{A}_2\cdot\mathbf{x}_{t-2}^* \;=\; \mathbf{e}_t$$

Ce formalisme se généralise naturellement à un modèle AR-vectoriel d'ordre quelconque.

6.3.3 Analyse cepstrale

Nous sommes dans le cas d'un filtrage indépendant du locuteur. $q = 0$ car chaque fonction sinusoïdale s'applique uniquement au vecteur $\mathbf{x}_t^*$. $r = k$, où k est le nombre des composantes cepstrales que l'on garde. Le filtre $\mathbf{H}$ est composé des différentes fonctions sinusoïdales correspondant aux différentes composantes cepstrales que l'on souhaite conserver après filtrage :

$$\mathbf{H} = \begin{bmatrix} \mathbf{c}_1^T \\ \vdots \\ \mathbf{c}_k^T \end{bmatrix}$$

On a alors :

$$\mathbf{f}_t = \mathbf{H}\cdot\mathbf{X}_t^t = \begin{bmatrix} \mathbf{c}_1^T \\ \vdots \\ \mathbf{c}_k^T \end{bmatrix}\cdot\mathbf{x}_t^* = \mathbf{c}_t$$

6.3.4 Paramètres Δ et $\Delta\Delta$

Nous sommes dans le cas d'un filtrage indépendant du locuteur. $q = 1, 2, 3, ...$, selon le nombre de vecteurs de paramètres consécutifs que nous conservons pour approximer la dérivée et la dérivée seconde d'un vecteur donné. $r = 3p$ si nous fabriquons un nouveau vecteur comportant à la fois le vecteur initial, un vecteur de dimension p approximant la dérivée, et un vecteur de dimension p approximant la dérivée seconde. On trouve un exemple de filtre $\mathbf{H}$ pour les paramètres Δ et $\Delta\Delta$ dans le TAB. 6.1.

6.3.5 Tableau récapitulatif

On trouve une récapitulation de ces différents exemples TAB. 6.1.

	Modèle mono-Gaussien	Modèle AR-vectoriel d'ordre 2	Analyse cepstrale	Paramètres Δ et $\Delta\Delta$
	Indépendant du locuteur	Dépendant du locuteur	Indépendant du locuteur	Indépendant du locuteur
q	0	2	0	1, 2, ...
r	p	p	k	$3p$
$\mathbf{H}$	$\mathbf{I}_p$	$[\mathbf{0}_p \vert \mathbf{0}_p \vert \mathbf{I}_p \vert \mathbf{A}_1 \vert \mathbf{A}_2]$	$\begin{bmatrix} \mathbf{c}_1^T \\ \vdots \\ \mathbf{c}_k^T \end{bmatrix}$	$\begin{bmatrix} \mathbf{0}_p & \mathbf{0}_p & \mathbf{I}_p & \mathbf{0}_p & \mathbf{0}_p \\ -2\mathbf{I}_p & -\mathbf{I}_p & \mathbf{0}_p & \mathbf{I}_p & 2\mathbf{I}_p \\ -\mathbf{I}_p & \mathbf{0}_p & 2\mathbf{I}_p & \mathbf{0}_p & -\mathbf{I}_p \end{bmatrix}$
$\mathbf{f}_t$	$\mathbf{x}_t$	$\mathbf{e}_t$	$\mathbf{c}_t$	$\begin{bmatrix} \mathbf{x}_t \\ \boldsymbol{\Delta}\mathbf{x}_t \\ \boldsymbol{\Delta}\boldsymbol{\Delta}\mathbf{x}_t \end{bmatrix}$

TAB. 6.1: *Quelques exemples de filtrages vectoriels de trajectoires spectrales : tableau récapitulatif.*

6.4 Mesures de similarité

6.4.1 Cas d'un filtrage indépendant du locuteur

Soit maintenant $\mathcal{Y}$ un deuxième locuteur. Nous définissons de même que pour $\mathcal{X}$ la matrice de covariance des vecteurs filtrés :

$$\begin{aligned} C_{\mathcal{H}(Y)} &= \frac{1}{N}\,\mathcal{H}(Y)\cdot\mathcal{H}(Y)^T \\ &= \mathbf{H}\cdot\left(\frac{1}{N}\sum_{t=1}^{N}\mathbf{Y}_{t-q}^{t+q}\cdot\mathbf{Y}_{t-q}^{t+q\,T}\right)\cdot\mathbf{H}^T \qquad (6.12) \\ &= \mathbf{H}\cdot\mathbf{Y}_{2q+1}\cdot\mathbf{H}^T \end{aligned}$$

Nous définissons alors des mesures de similarité entre les locuteurs $\mathcal{X}$ et $\mathcal{Y}$ de la même façon que dans les deux chapitres précédents, c'est-à-dire à partir des deux matrices de covariance $C_{\mathcal{H}(X)}$ et $C_{\mathcal{H}(Y)}$.

Commençons par introduire la matrice $\mathbf{\Gamma}$ par analogie avec les équations 4.3 et 5.10 :

$$\mathbf{\Gamma} = C_{\mathcal{H}(X)}^{-\frac{1}{2}}\cdot C_{\mathcal{H}(Y)}\cdot C_{\mathcal{H}(X)}^{-\frac{1}{2}} \qquad (6.13)$$

Les mesures de similarité entre le locuteur $\mathcal{X}$ et le locuteur $\mathcal{Y}$ sont alors définies à partir de a et g, respectivement la moyenne arithmétique et la moyenne géométrique des valeurs propres de la matrice $\mathbf{\Gamma}$, comme nous l'avons fait dans les deux chapitres précédents (cf. équations 4.10 et 5.14) :

- Mesures simples : a et g.
- Mesures composées : $a-g$, $log(a)-log(g)$ et $a-log(g)-1$.

Enfin, nous appliquons sur ces mesures les mêmes symétrisations que précédemment (cf. équations 4.46, 4.51 et 4.52).

Propriété importante : La trace et le déterminant de la matrice Γ sont invariants si la matrice de filtrage $\mathbf{H}$ est inversible.

Cette propriété est démontrée en annexe D. Elle entraine l'invariance des mesures a et g par filtrage linéaire inversible, et donc l'invariance de toutes les autres mesures du second ordre proposées dans ce chapitre.

6.4.2 Cas d'un filtrage dépendant du locuteur

Dans le cas d'un filtrage dépendant du locuteur, si nous mesurons la similarité entre le locuteur de test $\mathcal{Y}$ et le locuteur de référence $\mathcal{X}$, nous calculons alors la matrice $C_{\mathcal{H}_X(Y)}$:

$$C_{\mathcal{H}_X(Y)} = \mathbf{H}_X \cdot \mathbf{Y}_{2q+1} \cdot \mathbf{H}_X^T \tag{6.14}$$

Les mesures de similarité se calculent alors sur les valeurs propres de la matrice :

$$\mathbf{\Gamma} = C_{\mathcal{H}_X(X)}^{-\frac{1}{2}} \cdot C_{\mathcal{H}_X(Y)} \cdot C_{\mathcal{H}_X(X)}^{-\frac{1}{2}} \tag{6.15}$$

6.5 Choix du filtrage

Maintenant que ce formalisme a été défini, nous allons étudier un filtrage particulier, qui entre dans le cadre de ce formalisme.

Une première question qui se pose est le choix d'un filtrage dépendant du locuteur, ou indépendant du locuteur. Nous avons vu avec l'utilisation des modèles AR-vectoriels un exemple de filtrage dépendant du locuteur. Notons au passage que nous pourrions fabriquer un modèle AR-vectoriel appris sur de la parole multi-locuteur, et utiliser ensuite ce modèle comme filtrage indépendant du locuteur. Cette approche n'a pas été testée dans le cadre de cette thèse, mais est présentée comme l'une des nombreuses perspectives de ce travail.

Le chapitre 7 présente une autre approche pour le choix du filtrage. Cette approche fournit un filtrage indépendant du locuteur, et repose sur une analyse en composantes principales.

Chapitre 7

Filtrage à base de composantes principales temps-fréquence

Le chapitre 6 présente un formalisme mathématique pour le filtrage de trajectoires spectrales, filtrage qui s'effectue à la fois dans la dimension temporelle et dans la dimension fréquentielle.

Nous présentons dans ce chapitre un filtrage particulier, qui repose sur l'extraction des composantes principales de la parole multi-locuteur.

7.1 Principe

7.1.1 Matrice bloc-Toeplitz pour de la parole multi-locuteur

Nous avons vu que la matrice $\mathbf{X}_{2q+1}$ était en fait la matrice de covariance de la séquence de vecteurs $\{\mathbf{X}_{t-q}^{t+q}\}_{1\leq t\leq M}$.

Au lieu de calculer cette matrice pour un locuteur donné $\mathcal{X}$, nous pouvons la calculer pour plusieurs locuteurs. Ainsi, si $\mathcal{X}_1, ..., \mathcal{X}_S$ sont les S locuteurs de la base de référence, et si $\{\mathbf{x}_t^1\}_{1\leq t\leq M_1}, ..., \{\mathbf{x}_t^S\}_{1\leq t\leq M_S}$ sont les S séquences de vecteurs extraites pour chaque locuteur de cette base, nous définissons alors la matrice bloc-Toeplitz pour la parole multi-locuteurs de cette base de référence par :

$$\mathbf{G}_{2q+1} = \begin{bmatrix} \boldsymbol{\mathcal{G}}_0 & \boldsymbol{\mathcal{G}}_1 & \cdots & \boldsymbol{\mathcal{G}}_{2q} \\ \boldsymbol{\mathcal{G}}_1^T & \boldsymbol{\mathcal{G}}_0 & \cdots & \boldsymbol{\mathcal{G}}_{2q-1} \\ \vdots & \vdots & & \vdots \\ \boldsymbol{\mathcal{G}}_{2q}^T & \boldsymbol{\mathcal{G}}_{2q-1}^T & \cdots & \boldsymbol{\mathcal{G}}_0 \end{bmatrix} \tag{7.1}$$

où la matrice $\boldsymbol{\mathcal{G}}_k$ est la matrice de covariance décalée d'ordre k calculée sur

la séquence $\{\mathbf{x}_t^1\}_{1\leq t\leq M_1}, ..., \{\mathbf{x}_t^S\}_{1\leq t\leq M_S}$.

Notons que la matrice $\mathbf{G}_{2q+1}$ a pour dimension $(2q+1)p \times (2q+1)p$, et que chaque matrice $\boldsymbol{\mathcal{G}}_k$ a pour dimension $p \times p$.

7.1.2 Composantes principales

Une fois que nous avons calculé la matrice bloc-Toeplitz $\mathbf{G}_{2q+1}$ sur de la parole multi-locuteurs, nous extrayons de cette matrice ses composantes principales [76], [146]. Cela revient en fait à calculer les valeurs propres et les vecteurs propres de cette matrice. Le vecteur propre associé à la valeur propre la plus grande est alors la direction de projection qui conserve le maximum de variance (i.e. la distance entre les points projetés la plus grande), le vecteur propre associé à la deuxième valeur propre est la direction de projection qui conserve le maximum de variance de manière décorrélée (ce qui revient à l'orthogonalité d'un point de vue géométrique) à la première, et ainsi de suite. Nous avons alors la décomposition suivante :

$$\mathbf{G}_{2q+1} = \mathbf{V}_{2q+1} \cdot \mathbf{M}_{2q+1} \cdot \mathbf{V}_{2q+1}^T \tag{7.2}$$

avec :

$$\mathbf{V}_{2q+1} = (\mathbf{v}_1, ..., \mathbf{v}_{2q+1}) \tag{7.3}$$

$$\mathbf{M}_{2q+1} = diag(\mu_1, ..., \mu_{2q+1}), \quad \mu_1 \geq ... \geq \mu_{2q+1} \tag{7.4}$$

La matrice $\mathbf{V}_{2q+1}$ et la matrice $\mathbf{M}_{2q+1}$ sont toutes les deux de dimension $(2q+1)p \times (2q+1)p$. Chaque vecteur $\mathbf{v}_i, 1 \leq i \leq 2q+1$, a pour dimension $(2q+1)p$.

Si nous prenons la matrice $\mathbf{G}_{2q+1}$ ou bien la matrice $\mathbf{V}_{2q+1}$ comme matrice de filtrage, nous ne changeons rien aux mesures statistiques du second ordre, puisque ces deux matrices sont inversibles (cf. annexe D).

En revanche, en choisissant un sous-ensemble de vecteurs propres et en projetant sur le sous-espace vectoriel engendré par ce sous-ensemble, nous privilégions certaines informations des matrices initiales. Le problème est alors de déterminer les composantes qui vont permettre de mieux discriminer les locuteurs entre eux.

7.1.3 Choix des composantes

Les composantes principales de la matrice $\mathbf{G}_{2q+1}$ ont été obtenues en maximisant la variance sur chaque direction de projection. Nous obtenons

donc les composantes principales de la parole multi-locuteur en terme de variance décroissante. Les premières composantes étant celles qui maximisent la variance de la projection du nuage de points initial, elles sont donc celles qui contiennent le plus d'information. Les dernières composantes correspondent davantage à du bruit, et contiennent généralement peu d'information utile.

Il existe plusieurs stratégies pour choisir les composantes à conserver, et celles à éliminer. Une stratégie très souvent utilisée pour le choix des paramètres est celle du knock-out. Cette technique est expliquée notamment dans [145]. Elle consiste à supprimer chaque paramètre un par un, et à recalculer un score de reconnaissance en conservant tous les autres paramètres. On a ainsi pour chaque paramètre retiré une mesure de la dégradation des performances (et parfois de leur amélioration). On élimine alors le paramètre dont la suppression dégrade le moins les performances comme étant le moins utile, ou celui dont la suppression améliore le plus les performances si il en existe. Puis on réitère le processus pour en supprimer un second. Et ainsi de suite. On aboutit finalement à un jeu de paramètres plus réduit, qui donne des performances comparables au jeu initial, ou même meilleures.

L'un des principaux inconvénients de cette stratégie est qu'elle nécessite un temps de calcul très long. Nous avons choisi un compromis moins cher en temps de calcul. Nous nous sommes contentés de la première itération de la procédure de knock-out, ce qui nous fournit des fonctions de sensibilité des résultats en fonction des différentes composantes. Puis nous avons adopté une stratégie de choix des composantes reposant sur l'allure de ces fonctions de sensibilité. Nous revenons plus en détails sur le choix des composantes dans la section sur les résultats expérimentaux.

Finalement, la matrice de filtrage $\mathbf{H}$ est construite directement à l'aide des composantes principales. Si on choisit par exemple de conserver les composantes $(\mathbf{v}_2, ..., \mathbf{v}_7, \mathbf{v}_{10}, ..., \mathbf{v}_{14})$, la matrice de filtrage sera :

$$\mathbf{H} = \begin{bmatrix} \mathbf{v}_2^T \\ \vdots \\ \mathbf{v}_7^T \\ \mathbf{v}_{10}^T \\ \vdots \\ \mathbf{v}_{14}^T \end{bmatrix} \tag{7.5}$$

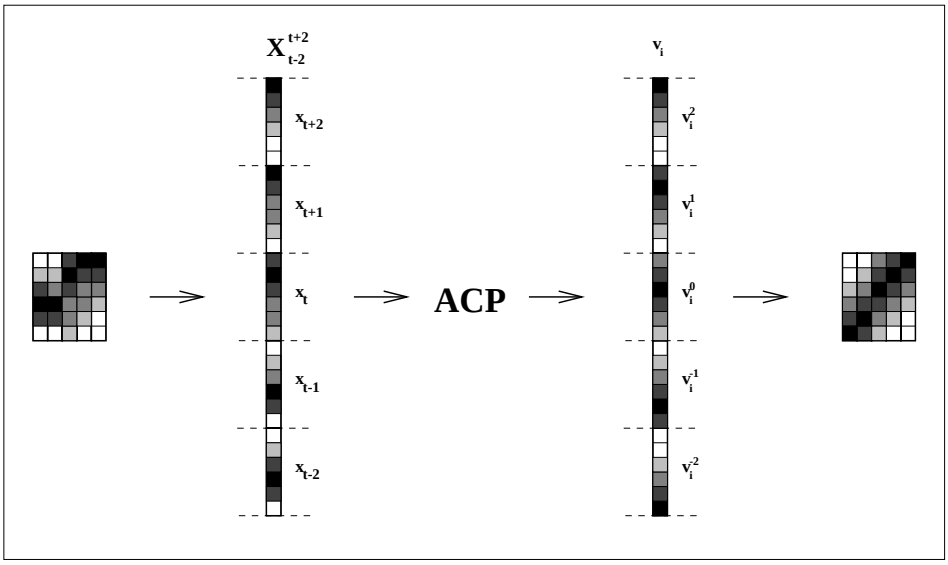


FIG. 7.1: *Interprétation d'une composante principale comme un masque temps-fréquence*

On applique alors ce filtrage indépendant du locuteur aux matrices d'apprentissage $\mathbf{X}_{2q+1}$ et à la matrice de test $\mathbf{Y}_{2q+1}$.

7.1.4 Interprétation d'une composante principale comme un masque temps-fréquence

Chaque vecteur $\mathbf{v}_i, 1 \leq i \leq 2q+1$, a pour dimension $(2q+1)p$. Il peut être découpé en $2q+1$ "tranches" de dimension p, chaque tranche correspondant à un indice temporel différent. En remettant ces tranches côte à côte en tenant compte de leur ordre temporel, on obtient un masque temps-fréquence à appliquer sur une séquence de vecteurs initiaux de même taille. Cette interprétation est illustrée FIG. 7.1.

7.2 Expériences et résultats

7.2.1 Description des expériences

7.2.1.a Tâche

La tâche évaluée ici est toujours l'identification du locuteur indépendante du texte en ensemble fermé. Pour une mesure donnée, sous une forme

symétrique ou non, nous désignons pour un test donné l'identité du locuteur de la base de référence qui est le plus proche au sens de la mesure testée. Les résultats sont donnés en pourcentages d'erreurs d'identification.

7.2.1.b Bases de données

Les résultats sont reportés pour les bases TIMIT63, FTIMIT63 et NTIMIT63 (cf. chapitre 3).

7.2.1.c Constitution de la matrice $\mathbf{G}_{2q+1}$

Plusieurs valeurs de q ont été testées. La valeur 0 correspond à un futur et un passé de longueur nulle, c'est-à-dire le filtrage s'applique uniquement à la trame courante.

Nous avons également testé la valeur $q = 1$, ce qui correspond à un filtrage s'appliquant à une séquence de trois trames $(\mathbf{x}_{t-1}, \mathbf{x}_t, \mathbf{x}_{t+1})$, ainsi que la valeur $q = 2$, ce qui correspond à un filtrage s'appliquant à une séquence de cinq trames $(\mathbf{x}_{t-2}, \mathbf{x}_{t-1}, \mathbf{x}_t, \mathbf{x}_{t+1}, \mathbf{x}_{t+2})$.

Les matrices de covariance décalées $\boldsymbol{\mathcal{G}}_k$ ont été calculées sur une grande séquence de vecteurs constituée des vecteurs de paramètres extraits des 5 phrases "sx" des 63 locuteurs des bases TIMIT63, FTIMIT63 ou NTIMIT63.

7.2.1.d Choix des composantes principales conservées

Des projections sur différentes combinaisons de composantes principales ont été testées. Cela est précisé au niveau des résultats.

7.2.1.e Analyse acoustique du signal

Cette analyse acoustique est décrite avec précision dans le chapitre 3. Nous avons choisi des fenêtres de 504 échantillons (31,5 ms).

7.2.1.f Protocole expérimental

Nous avons testé uniquement le protocole 5.1 (apprentissage long - test court).

7.2.2 Visualisation et interprétation de quelques composantes principales

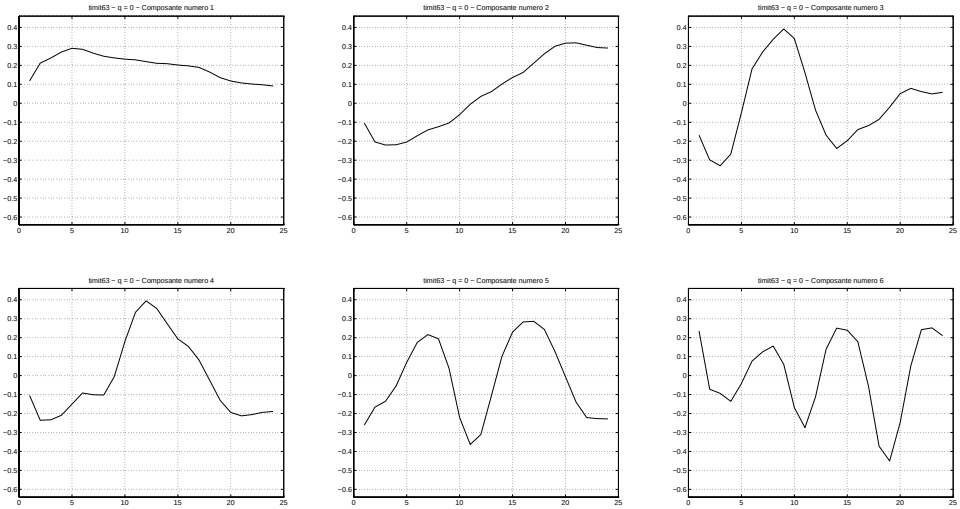
Afin de donner une idée précise de ce que sont les composantes principales pour les différentes bases de données et pour les différentes valeurs de q, nous en donnons ici une représentation visuelle. Toutes les composantes ne sont pas représentées dans ce chapitre, nous donnons juste quelques exemples. On trouvera des représentations visuelles supplémentaires en annexe E pour la base TIMIT63, en annexe F pour la base FTIMIT63, et en annexe G pour la base NTIMIT63.

Pour $q = 0$, la représentation est bi-dimensionnelle. L'axe des abcisses correspond au numéros des coordonnées de la composante principale, l'axe des ordonnées aux valeurs de ces coordonnées.

Pour $q = 1$ et $q = 2$, la représentation est tri-dimensionnelle. Nous donnons alors conjointement une visualisation 2D et 3D. Pour la visualisation 2D, l'axe des abcisses correspond à l'indice temporel, de 1 à 3 pour $q = 1$, et de 1 à 5 pour $q = 2$. L'axe des ordonnées correspond aux numéros de la coordonnée. La troisième dimension est rendue par le niveau de gris, le noir correspondant aux valeurs maximales et le blanc aux valeurs minimales. Il n'y a pas de normalisation le long des composantes, c'est-à-dire que les échelles de niveaux de gris ne sont pas les mêmes d'une composante à l'autre. Ceci nous permet de garder une bonne dynamique de représentation pour chaque composante. L'axe des cotes de la visualisation 3D donne les valeurs numériques des coordonnées, et permet donc de se faire une idée quantitative. Notons pour finir que nous avons effectué un lissage sous forme d'interpolation polynomiale, afin de faciliter la visualisation des composantes. Les variations du niveau de gris sont ainsi continues, au lieu d'être discrètes.

7.2.2.a Timit63

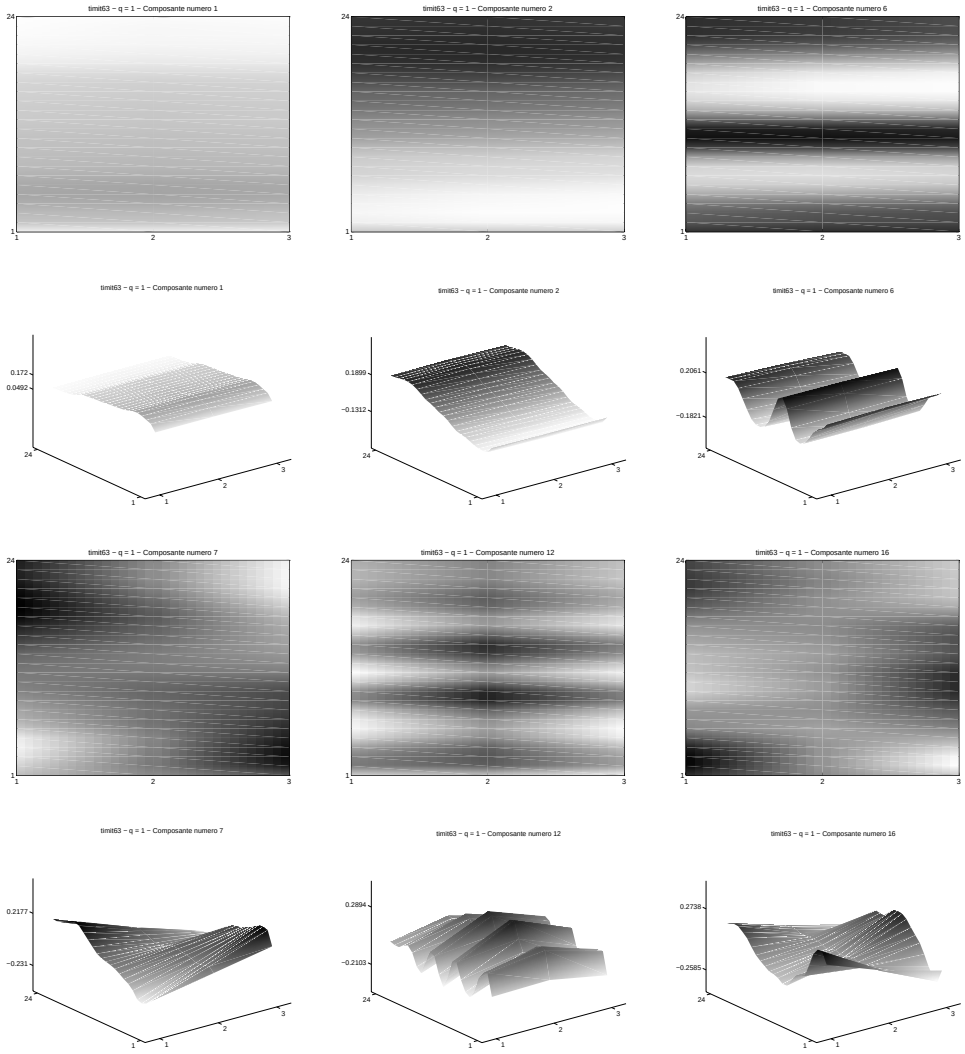
$q = 0$ La dimension de la matrice $\mathbf{G}_1$ est alors 24×24. Dans ce cas, les filtres sont simplement des filtres fréquentiels.



La première composante principale est plus ou moins une moyenne de tous les coefficients fréquentiels. La seconde composante ressemble à une dérivée première fréquentielle, c'est-à-dire elle réalise une différence entre les hautes fréquences et les basses fréquences, en donnant un peu plus de poids aux hautes fréquences. Attention toutefois au fait que l'échelle fréquentielle est une échelle Mel et non pas une échelle linéaire.

Il est plus difficile d'interpréter pareillement les autres composantes, mais nous pouvons néanmoins remarquer que les composantes sont de plus en plus oscillantes, et ont une structure qui ressemble à des sinusoïdes, ce qui les apparente fortement à des composantes cepstrales. Mais cette tendance est moins marquée pour les dernières composantes, et on observe en outre des effets de bord importants.

$q = 1$ La dimension de la matrice $\mathbf{G}_3$ est alors 72×72. Cette fois-ci, les filtres sont réellement des filtres temps-fréquence, et ils s'appliquent à des séquences de 3 trames.



Une fois encore, le premier filtre réalise plus ou moins une moyenne de tous les canaux fréquentiels, ainsi qu'une moyenne temporelle. Le second filtre est quasiment invariant dans le temps, il coupe un peu les basses fréquences et laisse passer les hautes fréquences.

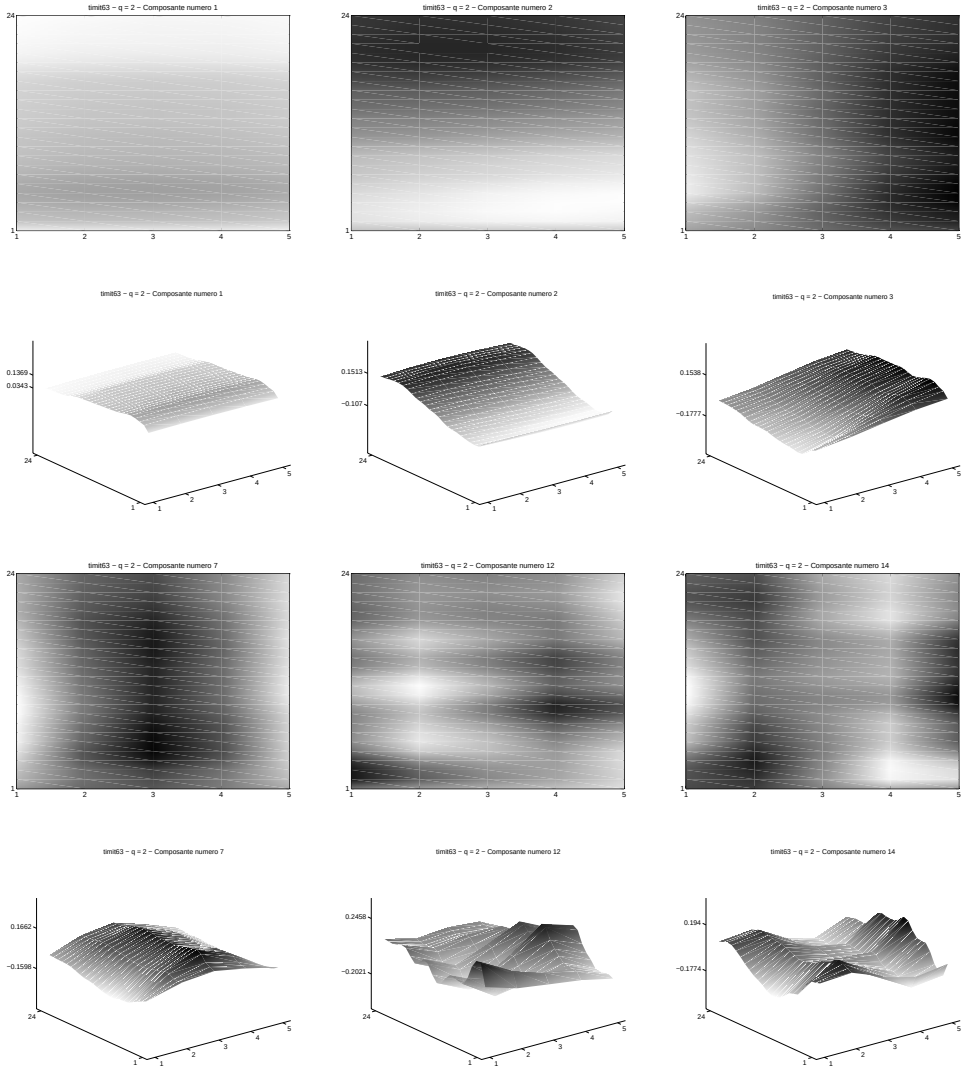
Nous avons également représenté la sixième composante, qui est un autre exemple de filtrage invariant dans le temps, et qui présente une forme sinusoïdale dans sa dimension fréquentielle. Il s'agit donc d'une moyenne temporelle d'un coefficient cepstrale le long de trois trames consécutives. Le nombre d'oscillations nous donne à peu près l'indice du coefficient cepstral, ici c_4.

La composante 12 est oscillante dans sa dimension fréquentielle, et ressemble à une approximation de dérivée seconde dans sa dimension temporelle. Cela ressemble fortement à l'approximation de la dérivée seconde d'un coefficient cepstral, ici probablement c_8.

Remarquons que la structure temporelle des filtres est très souvent ou bien symétrique (ce qui correspond par exemple à un moyennage temporel comme pour les composantes 1 ou 6, ou à une dérivée seconde des séquences comme l'illustre la composante 12), ou bien anti-symétrique (ce qui correspond par exemple à une dérivée première comme l'illustrent les composantes 7 ou 16). Cette tendance se retrouve aussi pour $q = 2$, et pour les deux autres bases de données. Notons au passage qu'avec $q = 1$, il n'est pas possible d'approximer des dérivées d'ordre supérieur à 2.

Pour finir, les composantes 7 ou 16 offrent des exemples de filtres temps-fréquence qui ne sont ni invariants en temps, ni invariants en fréquence, et n'offrent pas de structures oscillantes. Seul un filtrage bi-dimensionnel permet l'extraction de telles composantes.

$q = 2$ La matrice $\mathbf{G}_5$ a pour dimension 120×120. Les filtres temps-fréquence s'appliquent alors à des séquences de 5 trames.



On retrouve les mêmes tendances à l'ordre 5. La première composante est un exemple de moyennage à la fois temporel et fréquentiel. Le deuxième filtre est toujours un filtre fréquentiel, qui coupe un peu les basses fréquences et laisse passer les hautes. Le troisième filtre, quant à lui, est quasiment invariant en fréquence, et réalise une approximation de la dérivée première dans la dimension temporelle. Il s'agit donc d'un Δ sur les coefficients de

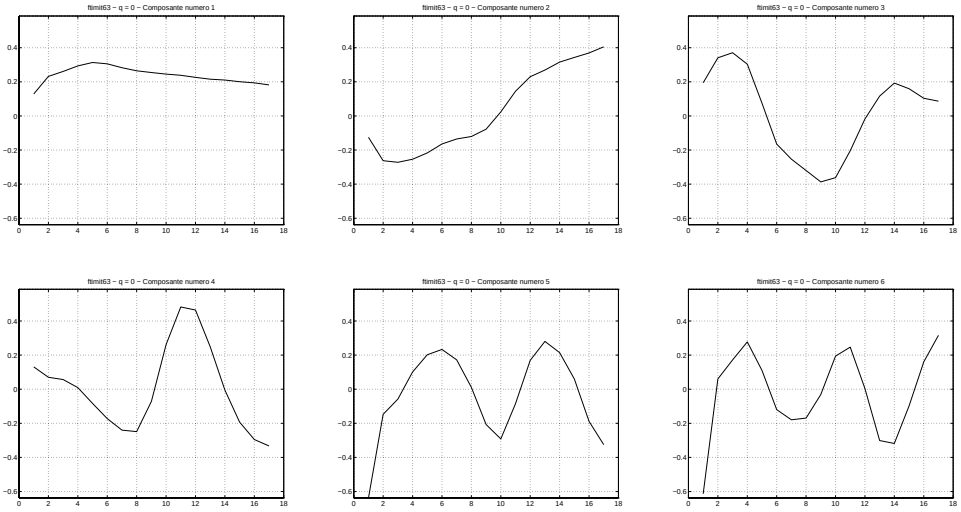
banc de filtres.

Les composantes 7, 12 ou 14 ne sont ni invariantes en temps, ni invariantes en fréquence. Elles ne présentent pas non plus de structure oscillante.

Enfin, nous pouvons faire la même remarque sur l'interprétation temporelle des filtres. Les formes temporellement symétriques correspondent à des dérivées paires des paramètres (moyennage pour la composante 1 ou 2, dérivée seconde pour la composante 7, il n'y a pas d'exemple de dérivée quatrième dans les 24 premières composantes), et les formes temporellement anti-symétriques à des dérivées impaires (dérivée première pour la composante 3, dérivée troisième pour les composantes 12 ou 14). Notons pour finir qu'avec $q = 2$, il n'est pas possible d'approximer des dérivées d'ordre supérieur à 4.

7.2.2.b Ftimit63

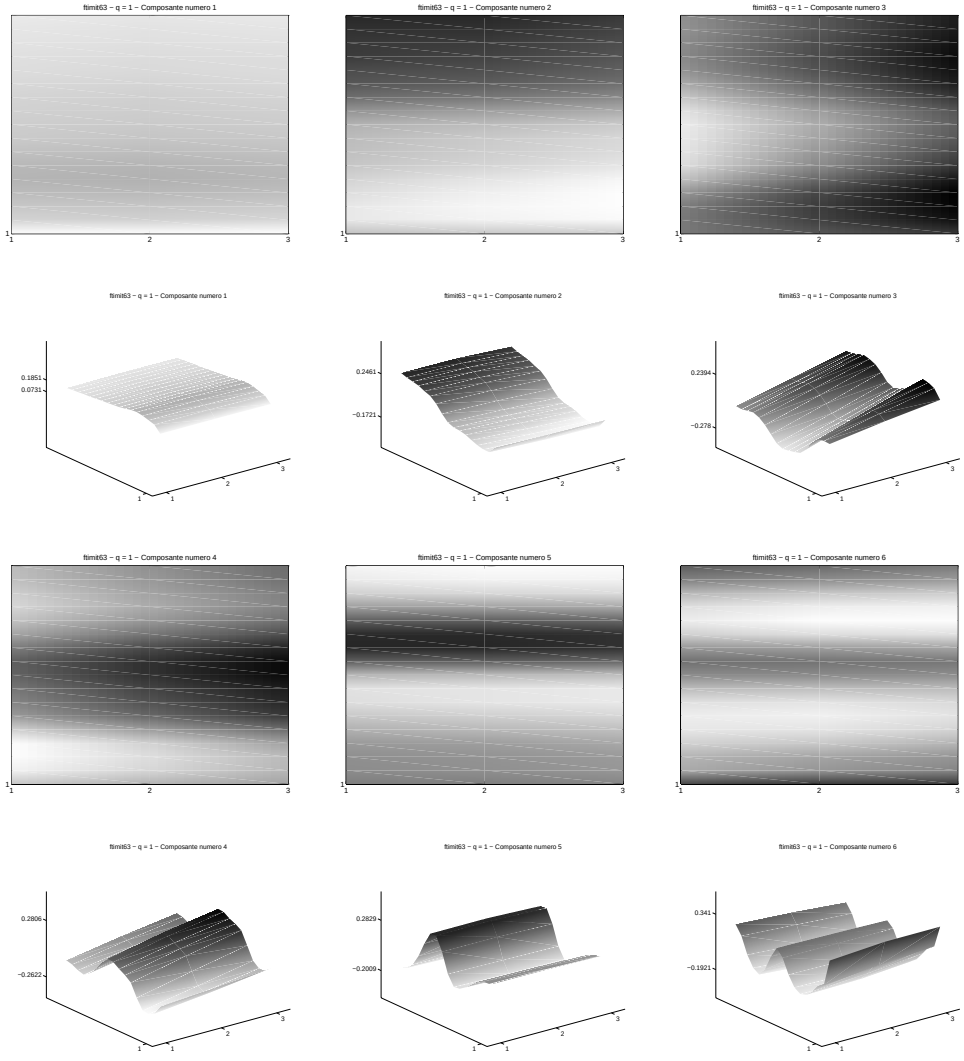
$q = 0$ La dimension de la matrice $\mathbf{G}_1$ est 17×17. Les filtres sont purement fréquentiels.



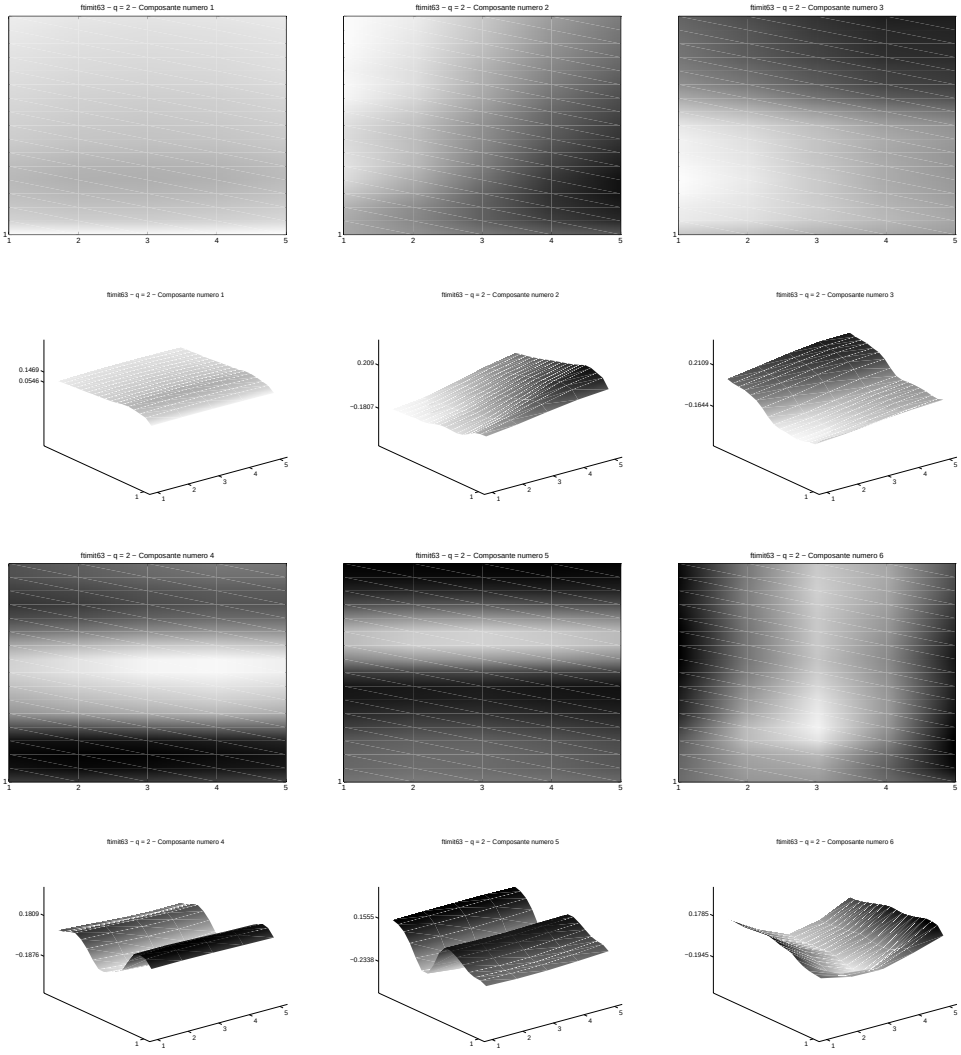
Il est intéressant de comparer la forme des composantes principales sur FTIMIT63 avec celles que nous avons pour TIMIT63. Les formes des composantes sont très similaires, et les pics fréquentiels sont à peu près aux mêmes endroits. On observe néanmoins un certain nombre de différences. Par exemple, sur la composante 3, on a une inversion du signe de la courbe.

Dans l'ensemble, les composantes obtenues sur FTIMIT63 sont des versions tronquées de celles obtenues sur TIMIT63. On remarque en particulier cette même structure de plus en plus oscillante, qui apparente certaines composantes à des coefficients cepstraux.

$q = 1$ La dimension de la matrice $\mathbf{G}_3$ est cette fois-ci 51×51. Les filtres sont des filtres temps-fréquence, qui s'appliquent à des séquences de 3 trames.

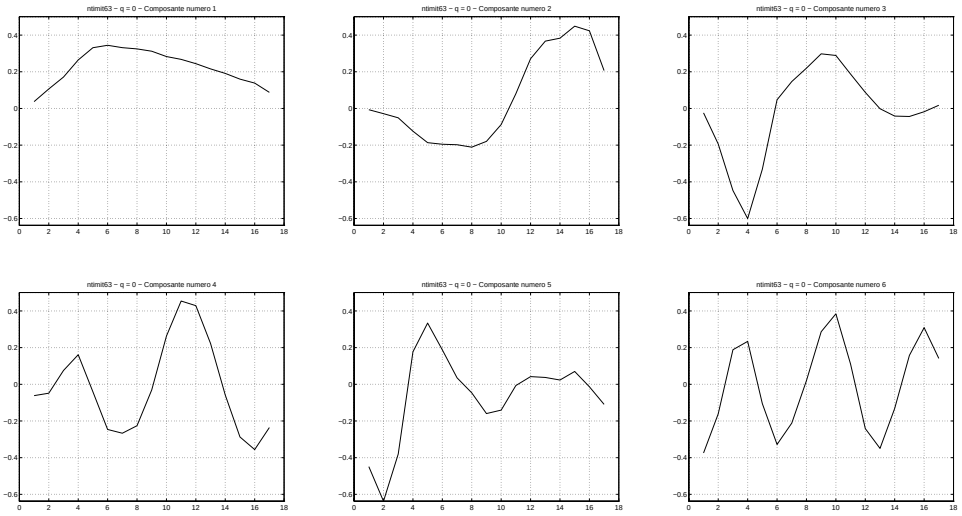


$q = 2$ La matrice $\mathbf{G}_5$ a pour dimension 85×85. Les filtres temps-fréquence s'appliquent alors à des séquences de 5 trames.

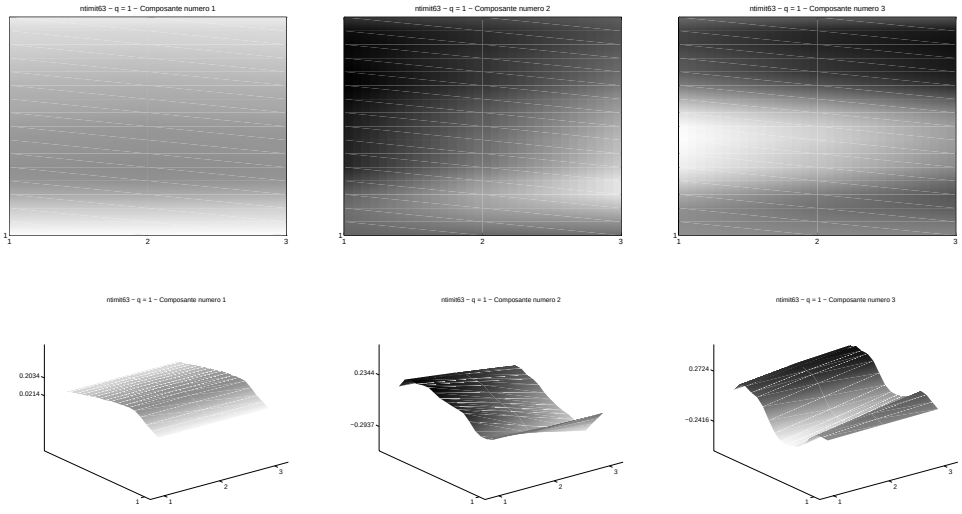


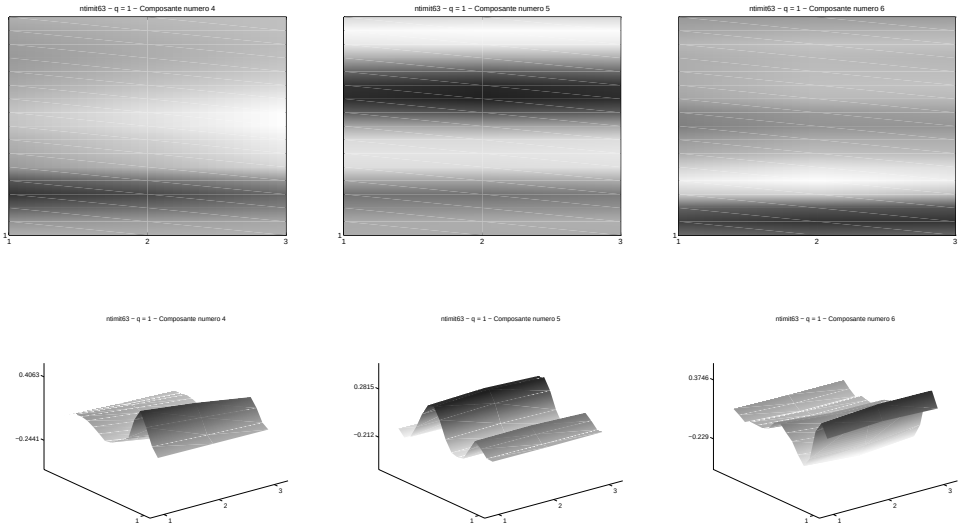
7.2.2.c Ntimit63

$q = 0$ La dimension de la matrice $\mathbf{G}_1$ est, comme pour FTIMIT63, 17×17. Les filtres sont de simples filtres fréquentiels.

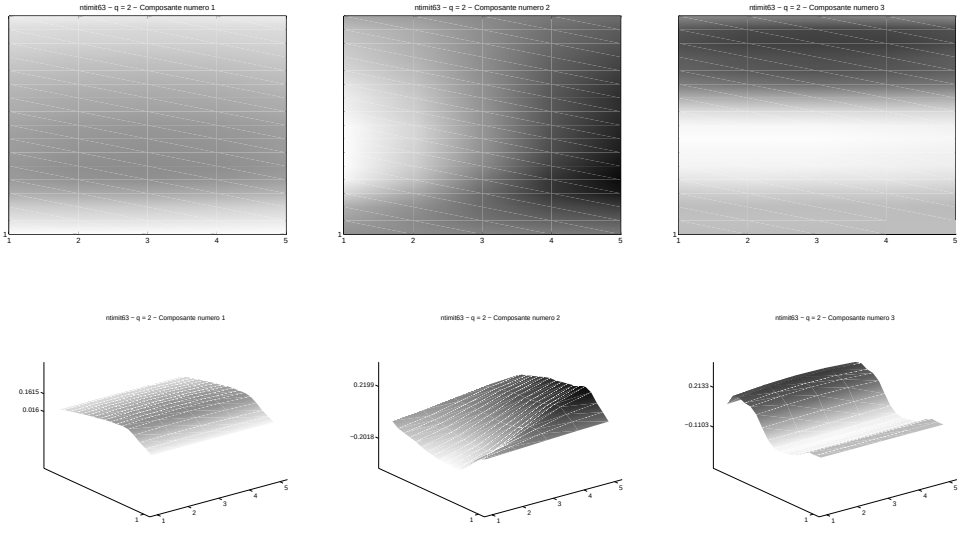


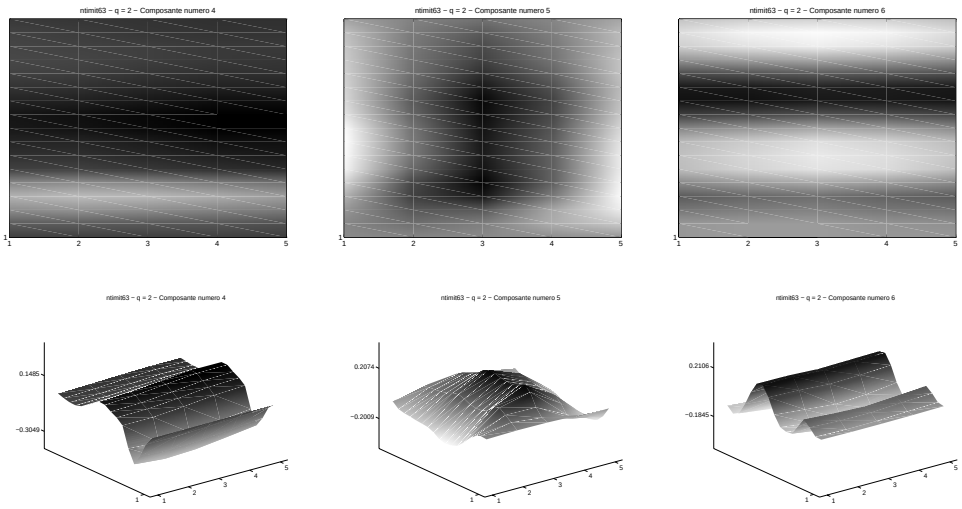
$q = 1$ La dimension de la matrice $\mathbf{G}_3$ est 51×51. Les filtres sont des filtres temps-fréquence s'appliquant à des séquences de 3 trames.





$q = 2$ La matrice $\mathbf{G}_5$ a pour dimension 85×85. Les filtres temps-fréquence s'appliquent alors à des séquences de 5 trames.





7.2.3 Fonctions de sensibilité

Nous avons réalisé la première itération d'une procédure de knock-out afin de nous faire une idée sur la sensibilité de notre système d'identification du locuteur en fonction des différentes composantes principales.

7.2.3.a Base TIMIT63

On trouve Fig. 7.2 une courbe donnant les différents pourcentages d'erreurs d'identification lorsqu'on retire une à une les 24 composantes principales. Cette courbe est obtenue avec la mesure de sphéricité $(log(a) - log(g))$ et pour $q = 0$.

On trouve Fig. 7.3 une courbe donnant les différents pourcentages d'erreurs d'identification lorsqu'on retire une à une les 72 composantes principales. Cette courbe est obtenue avec la mesure de sphéricité $(log(a) - log(g))$ et pour $q = 1$.

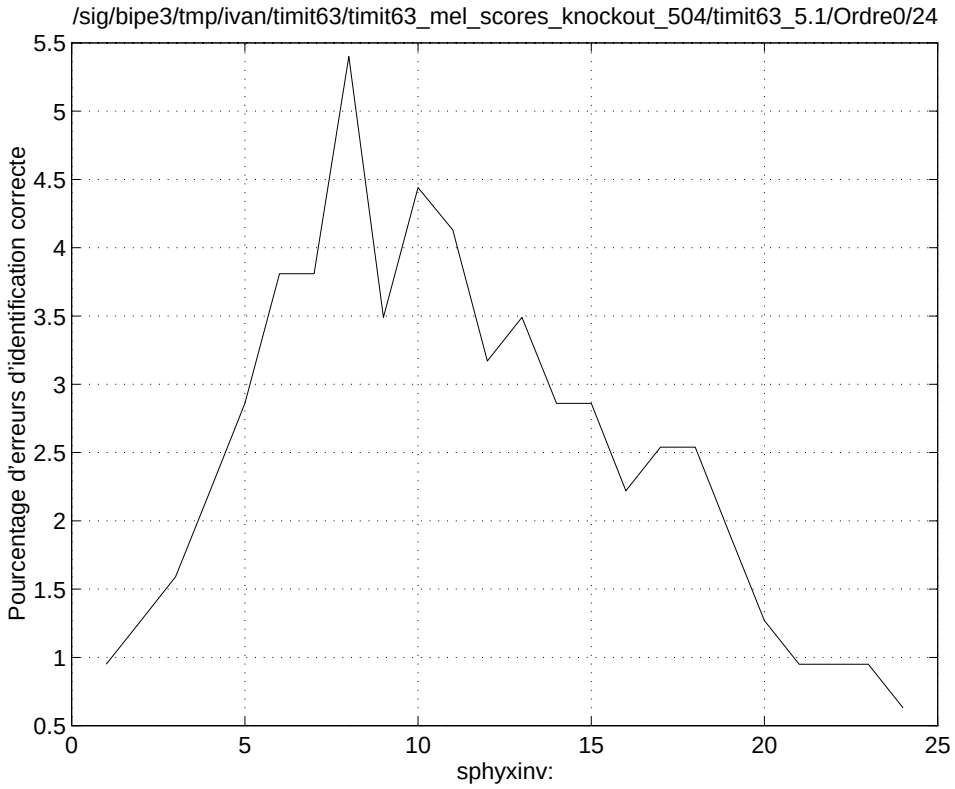


FIG. 7.2: *Première itération de la procédure de knock-out sur TIMIT63 ($q = 0$). Mesure de sphéricité non symétrisée. Les résultats sont donnés en pourcentage d'erreurs d'identification.*

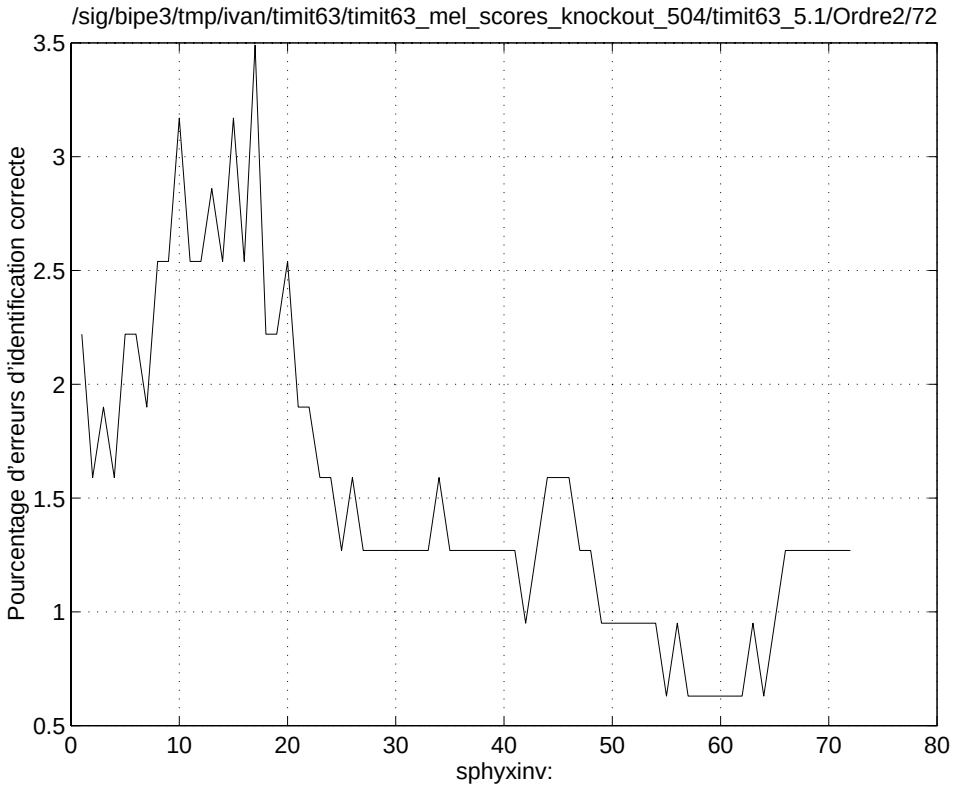


FIG. 7.3: *Première itération de la procédure de knock-out sur TIMIT63 ($q = 1$). Mesure de sphéricité non symétrisée. Les résultats sont donnés en pourcentage d'erreurs d'identification.*

7.2.3.b Base FTIMIT63

On trouve Fig. 7.4 une courbe donnant les différents pourcentages d'erreurs d'identification lorsqu'on retire une à une les 17 composantes principales. Cette courbe est obtenue avec la mesure de sphéricité $(log(a) - log(g))$ et pour $q = 0$.

On trouve Fig. 7.5 une courbe donnant les différents pourcentages d'erreurs d'identification lorsqu'on retire une à une les 51 composantes principales. Cette courbe est obtenue avec la mesure de sphéricité $(log(a) - log(g))$ et pour $q = 1$.

7.2.3.c Base NTIMIT63

On trouve Fig. 7.6 une courbe donnant les différents pourcentages d'erreurs d'identification lorsqu'on retire une à une les 17 composantes principales. Cette courbe est obtenue avec la mesure de sphéricité $(log(a) - log(g))$ et pour $q = 0$.

On trouve Fig. 7.7 une courbe donnant les différents pourcentages d'erreurs d'identification lorsqu'on retire une à une les 51 composantes principales. Cette courbe est obtenue avec la mesure de sphéricité $(log(a) - log(g))$ et pour $q = 1$.

7.2.3.d Stratégie adoptée

Dans leur ensemble, les courbes représentant les fonctions de sensibilité ne présentent pas de discontinuités importantes, à l'exception de la courbe correspondant a NTIMIT63 pour $q = 0$. Cette dernière courbe étant la seule à présenter une discontinuité de cette nature, nous avons choisi finalement de tester différents ensembles de composantes principales consécutives. Ceci nous permet de tester à moindre frais de nombreuses combinaisons différentes, sans toutefois tester toutes les combinaisons possibles.

7.2.4 Résultats principaux

Nous présentons les meilleurs résultats obtenus en testant différentes combinaisons de composantes principales consécutives.

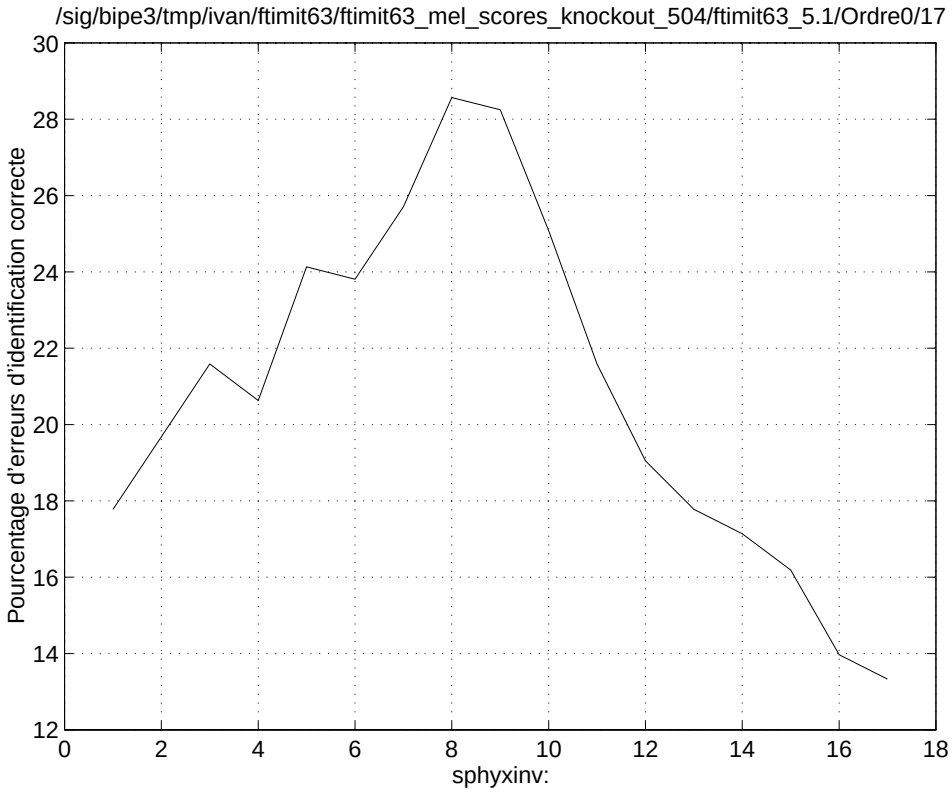


FIG. 7.4: *Première itération de la procédure de knock-out sur FTIMIT63 ($q = 0$). Mesure de sphéricité non symétrisée. Les résultats sont donnés en pourcentage d'erreurs d'identification.*

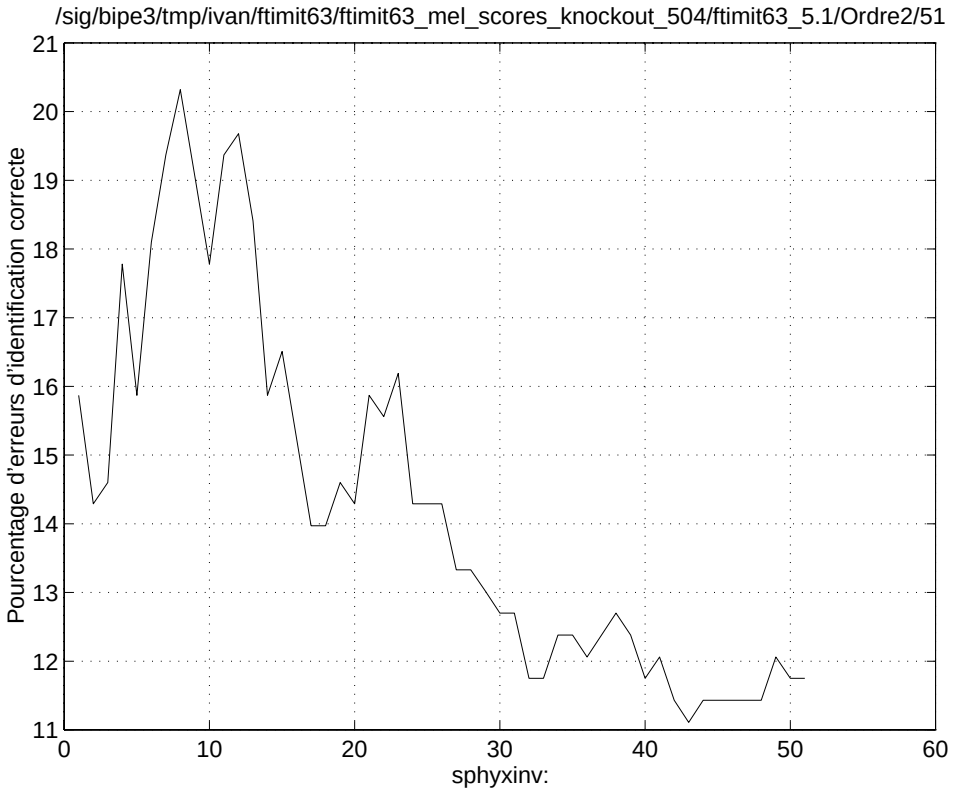


FIG. 7.5: *Première itération de la procédure de knock-out sur FTIMIT63 ($q = 1$). Mesure de sphéricité non symétrisée. Les résultats sont donnés en pourcentage d'erreurs d'identification.*

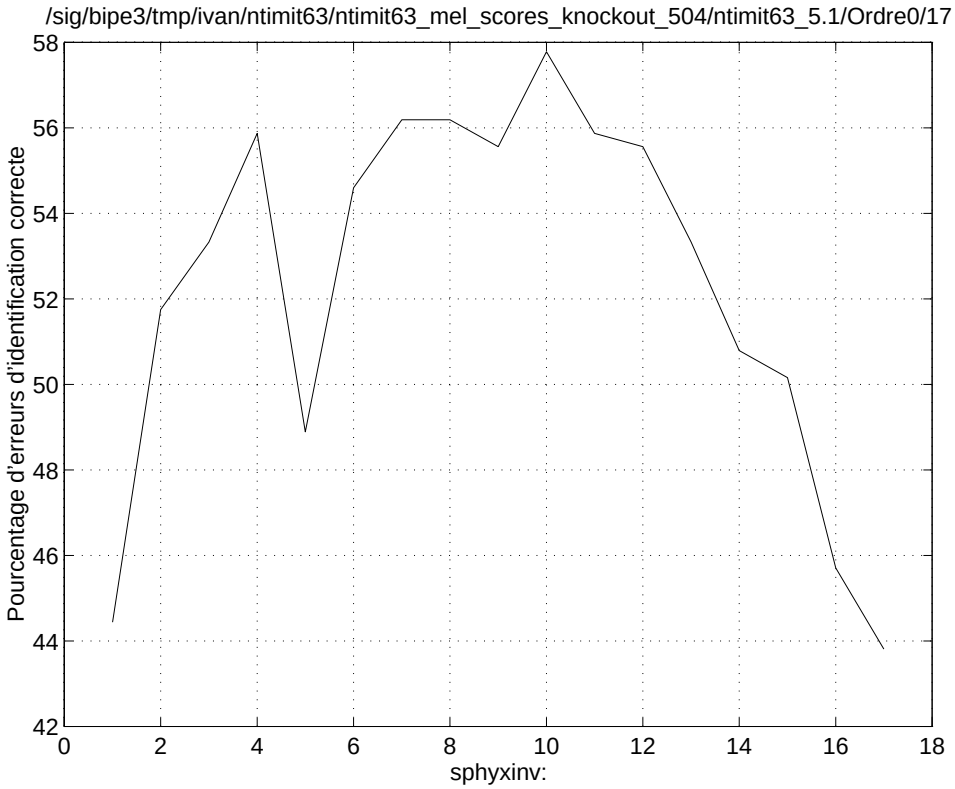


FIG. 7.6: *Première itération de la procédure de knock-out sur NTIMIT63 ($q = 0$). Mesure de sphéricité non symétrisée. Les résultats sont donnés en pourcentage d'erreurs d'identification.*

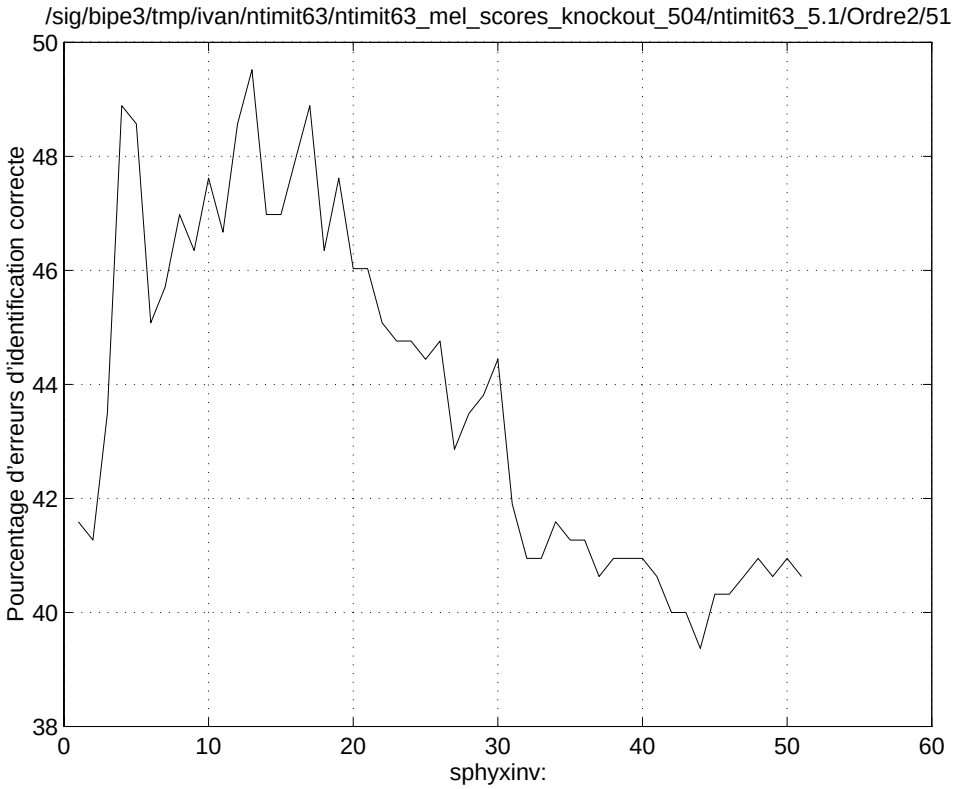


FIG. 7.7: *Première itération de la procédure de knock-out sur NTIMIT63 ($q = 1$). Mesure de sphéricité non symétrisée. Les résultats sont donnés en pourcentage d'erreurs d'identification.*

Nous trouvons TAB. 7.1 des résultats sur la base TIMIT63, TAB. 7.2 des résultats sur la base FTIMIT63, et TAB. 7.3 des résultats sur la base NTIMIT63. Nous donnons à chaque fois le meilleur résultat obtenu. Celui-ci correspond souvent à la mesure de sphéricité, parfois sous sa forme asymétrique, parfois sous une forme symétrique. Le score de référence correspond au score obtenu en appliquant les mesures directement aux matrices bloc-Toeplitz de l'ordre correspondant, ce qui revient en fait à conserver toutes les composantes principales (à cause de l'invariance par filtrage linéaire inversible). Enfin, nous indiquons à chaque fois les composantes qui ont été conservées.

Valeurs de q	Mesure	Composantes	Meilleur score	Score de référence
$q = 0$	μ_{Sc}	1 à 23	0.63 %	0.63 %
$q = 1$	μ_{Sc}	1 à 32	0.63 %	1.27 %
$q = 2$	μ_{Sc}	1 à 60	0.95 %	2.86 %

TAB. 7.1: *Quelques combinaisons de composantes successives sur la base TIMIT63. Les résultats sont donnés en pourcentages d'erreurs d'identification.*

Valeurs de q	Mesure	Composantes	Meilleur score	Score de référence
$q = 0$	$\mu_{Sc\,[\alpha_{MN}]}$	1 à 17	12.06 %	12.06 %
$q = 1$	μ_{Sc}	1 à 42	11.43 %	11.75 %
$q = 2$	μ_{Sc}	1 à 52	13.33 %	14.60 %

TAB. 7.2: *Quelques combinaisons de composantes successives sur la base FTIMIT63. Les résultats sont donnés en pourcentages d'erreurs d'identification.*

Valeurs de q	Mesure	Composantes	Meilleur score	Score de référence
$q = 0$	μ_{Sc}	1 à 17	40.32 %	40.32 %
$q = 1$	μ_{Sc}	1 à 41	37.46 %	40.95 %
$q = 2$	μ_{Sc}	1 à 64	40.00 %	43.49 %

TAB. 7.3: *Quelques combinaisons de composantes successives sur la base NTIMIT63. Les résultats sont donnés en pourcentages d'erreurs d'identification.*

- Nous pouvons remarquer, pour commencer, que la sélection des composantes principales permet toujours d'obtenir de meilleurs résultats que le score de référence correspondant au même ordre, pour $q = 1$ et $q = 2$.

- Pour $q = 0$, les résultats sont moins concluants. Le score de référence, dans ce cas, n'est jamais dépassé. D'ailleurs, pour FTIMIT63 ou NTIMIT63, il n'y a aucune combinaison autre que l'ensemble des 17 composantes pour obtenir ce score de référence.
- Sur la base FTIMIT63, nous obtenons le meilleur score d'identification avec $q = 1$, score qui est meilleur que tous les scores de référence. Nous obtenons en effet **11.43 %** d'erreurs d'identification.
- Mais le résultat le plus intéressant est obtenu sur la base NTIMIT63. Cette fois encore, nous obtenons le meilleur score avec $q = 1$, et ce score est meilleur que tous les scores de référence. Nous obtenons **37.46 %** d'erreurs d'identification, soit une réduction du taux d'erreur d'environ 7 %.

7.3 Conclusions

Nos expériences ne nous ont pas permis d'isoler des composantes qui représenteraient essentiellement la parole, et d'autres qui représenteraient essentiellement l'identité du locuteur. Il semblerait, au vu de nos expériences, que ces deux contributions soient étroitement mélangées.

Nous obtenons néanmoins des résultats très prometteurs, puisque nous avons obtenu, sur les bases FTIMIT63 et NTIMIT63, nos meilleurs scores d'identification. En particulier, nous obtenons les meilleurs résultats sur de la parole téléphonique. Cela semble donc valider cette approche de filtrage vectoriel, et montrer plus particulièrement son intérêt en cas de parole téléphonique.

Nous pouvons même espérer améliorer encore les résultats en testant d'autres filtrages vectoriels, car nous en avons choisi un parmi tant d'autres possibles, dans le cadre du formalisme proposé au chapitre précédent.

Cependant, le critère optimisé ne dégage pas nécessairement une paramétrisation pour la reconnaissance du locuteur. Un critère qui dégagerait davantage une paramétrisation spécifique serait de faire une analyse en composantes principales pour chaque locuteur de la base de référence. On réaliserait ainsi une projection sur un espace propre à chaque locuteur, ce qui pourrait améliorer les propriétés discriminantes de la paramétrisation.

Enfin, il faut encore tester cette approche de filtrage vectoriel sur d'autres bases de données, et avec d'autres méthodes d'identification du locuteur (par exemple les mixtures de Gaussiennes) pour pouvoir conclure définitivement à sa grande utilité en reconnaissance du locuteur.

Conclusions-Perspectives

Conclusions et perspectives

Ce chapitre résume les principales conclusions de notre travail de thèse, puis propose quelques perspectives ouvertes par celui-ci.

Conclusions

- ☞ Nous avons testé de façon systématique une famille de mesures de similarité, qui utilisent les statistiques du second ordre des séquences de vecteurs de paramètres, et nous les avons symétrisées. Ces mesures sont simples à implanter et faciles à reproduire. Elles ont été un point de départ à notre travail, en fournissant une méthode de référence. Elles constituent en outre un bon compromis entre simplicité, reproductibilité et performances, ce qui pourrait en faire, selon nous, une famille de méthodes de référence au niveau de l'évaluation en reconnaissance du locuteur.
- ☞ Nous avons ensuite testé un ensemble de mesures de similarité et de normalisations à partir des résiduels de prédiction vectorielle linéaire, ces derniers étant obtenus à l'aide d'un modèle auto-régressif vectoriel d'ordre 2. Un protocole expérimental original nous a permis de tester l'influence de l'information dynamique sur les performances des modèles AR-vectoriels en identification du locuteur. Nous ne sommes pas parvenus à montrer que les performances de ces modèles étaient liées à leur capacité à capturer des caractéristiques dynamiques du locuteur.
- ☞ Nous avons alors généralisé cette approche en proposant le filtrage vectoriel de trajectoires spectrales. La formalisation mathématique de ce filtrage nous a permis de montrer qu'il unifiait de nombreuses

approches différentes (modèles AR-vectoriels, analyse cepstrale, paramètres Δ et $\Delta\Delta$, paramètres RASTA, transformées en cosinus de trajectoires spectrales, ...). L'interprétation du filtrage vectoriel en termes de masques temps-fréquence permet de comprendre physiquement le filtrage choisi. Nous avons enfin proposé un nouveau filtrage vectoriel reposant sur l'extraction des composantes principales spectro-temporelles de parole multi-locuteur. La mise en œuvre de ce filtrage nous a permis d'obtenir une légère amélioration des performances sur une base de données de qualité téléphonique.

Perspectives

De nombreuses perspectives permettent de prolonger ce travail dans diverses directions. Nous les regroupons en différents pôles.

- **Au niveau de la paramétrisation initiale :**
  - Utiliser un détecteur de parole pour supprimer les silences de début et de fin de phrases.
  - Utiliser des techniques d'égalisation pour déconvoluer le signal de parole et l'effet du canal de transmission (pour améliorer les performances sur la parole téléphonique).
  - Tester différentes méthodes pour essayer de "Gaussianiser" les vecteurs de paramètres (quantification vectorielle, décomposition temporelle, ...).
  - Modéliser le signal de parole par des méthodes non-linéaires.
  - Utiliser d'autres transformations temps-fréquence que la transformée de Fourier (Wigner-Ville, ...).
- **Au niveau du filtrage vectoriel :**
  - Tester d'autres filtrages vectoriels dans le cadre du formalisme proposé : un modèle AR-vectoriel d'ordre 2 appris sur de la parole multi-locuteur, une approche à base de composantes principales dépendantes du locuteur, ...
  - Utiliser la formalisation mathématique du filtrage vectoriel pour pouvoir choisir a priori le bon filtrage vectoriel en fonction du critère à optimiser.
  - Tester le filtrage vectoriel à base de composantes principales sur d'autres bases de données (Switchboard, ...) ou/et avec d'autres méthodes de reconnaissance du locuteur (mélanges de Gaussiennes, modèles de Markov cachés, ...).

- **Au niveau de la modélisation statistique :**
  - Augmenter l'ordre de la modélisation statistique. En effet, comme il apparaît que les vecteurs de paramètres ne suivent pas une loi de probabilité rigoureusement Gaussienne, nous pouvons espérer apporter de l'information supplémentaire en utilisant les moments d'ordre 3 et 4 de cette distribution de vecteurs.
  - Utiliser une autre modélisation statistique comme par exemple la modélisation par mélange de Gaussiennes, qui est largement utilisée en reconnaissance du locuteur. Nous pouvons modéliser ainsi les vecteurs obtenus après filtrage vectoriel.
  - Utiliser une autre modélisation non-Gaussienne des vecteurs de paramètres (distribution de Laplace, ...).
- **Au niveau applicatif :**
  - Tester le filtrage vectoriel à base de composantes principales sur une tâche simple de reconnaissance de la parole.
  - Utiliser le filtrage vectoriel en codage de la parole, en synthèse de la parole, ...
  - Utiliser le filtrage vectoriel pour d'autres problèmes de classification en traitement du signal.

De ce qui est fait, rien n'est si beau qu'on puisse s'y reposer, rien n'est si laid qu'on ne le puisse sauver.

Alain, *Propos sur le bonheur.*

Annexes

Annexe A

Échelle d'analyse et taille de la fenêtre

Dans cette annexe, nous présentons plusieurs résultats relatifs à différentes conditions d'analyse. Nous avons réalisé la même analyse que celle qui est décrite dans le chapitre 3, mais en faisant varier deux facteurs.

A.1 Variation de la taille de la fenêtre d'analyse

Le premier facteur que nous avons fait varier est la taille de la fenêtre d'analyse, mais sans changer le décalage entre les trames. Nous avons testé des longueurs de trames de 252 échantillons (15,75 ms) et de 126 échantillons (7,88 ms), et comparé les résultats obtenus pour ces deux nouvelles tailles de fenêtre avec l'analyse initiale qui consistait à prendre des fenêtres de 504 échantillons (31,5 ms).

A.2 Échelle d'analyse linéaire

Nous avons également testé des bancs de filtres dont l'échelle était linéaire. Dans le cas de la base TIMIT63, nous avons découpé la bande de fréquence 0 – 8 kHz en 24 filtres triangulaires linéairement répartis. Pour les bases FTIMIT63 et NTIMIT63, nous avons découpé la bande de fréquence 0 – 8 kHz en 34 filtres linéairement répartis, de façon à en conserver également 17 pour simuler la bande passante téléphonique.

A.3 Expériences et résultats

La tâche évaluée est toujours l'identification du locuteur indépendante du texte en ensemble fermé. Pour chaque mesure, sous une forme symétrique

ou non, nous désignons pour un test donné l'identité du locuteur de la base de référence qui est le plus proche au sens de la mesure testée. Les résultats sont donnés en pourcentages d'erreurs d'identification. Les expériences sont menées sur les bases TIMIT63, FTIMIT63 et NTIMIT63. Nous avons testé un seul protocole expérimental, le protocole 5.1 (apprentissage long - test court). Nous avons réalisés ces expériences sur les matrices de covariance ($q = 0$), sur les matrices bloc-Toeplitz d'ordre 3 ($q = 1$), et sur les matrices bloc-Toeplitz d'ordre 5 ($q = 2$). Les résultats sont donnés TAB. A.1 pour TIMIT63, TAB. A.2 pour FTIMIT63, et TAB. A.3 pour NTIMIT63. Pour chaque base, nous ne faisons figurer pour chaque expérience que le score de la meilleure mesure (symétrisée ou non selon les cas).

Matrices de covariance ($q = 0$)		
Échelle Mel		
126 échantillons	252 échantillons	504 échantillons
0.95 %	0.95 %	0.63 %
Échelle linéaire		
126 échantillons	252 échantillons	504 échantillons
0.63 %	**0.00 %**	**0.00 %**
Matrices bloc-Toeplitz d'ordre 3 ($q = 1$)		
Échelle Mel		
126 échantillons	252 échantillons	504 échantillons
1.59 %	0.95 %	1.27 %
Échelle linéaire		
126 échantillons	252 échantillons	504 échantillons
0.95 %	**0.32 %**	**0.32 %**
Matrices bloc-Toeplitz d'ordre 5 ($q = 2$)		
Échelle Mel		
126 échantillons	252 échantillons	504 échantillons
2.86 %	**1.90 %**	2.86 %
Échelle linéaire		
126 échantillons	252 échantillons	504 échantillons
2.54 %	**1.90 %**	2.22 %

TAB. A.1: *Différentes conditions d'analyses sur la base TIMIT63. Les résultats sont donnés en pourcentages d'erreurs d'identification.*

Matrices de covariance ($q = 0$)		
Échelle Mel		
126 échantillons	252 échantillons	504 échantillons
17.14 %	12.06 %	12.06 %
Échelle linéaire		
126 échantillons	252 échantillons	504 échantillons
12.70 %	**11.75 %**	12.38 %
Matrices bloc-Toeplitz d'ordre 3 ($q = 1$)		
Échelle Mel		
126 échantillons	252 échantillons	504 échantillons
13.33 %	10.48 %	11.75 %
Échelle linéaire		
126 échantillons	252 échantillons	504 échantillons
8.89 %	10.16 %	11.11 %
Matrices bloc-Toeplitz d'ordre 5 ($q = 2$)		
Échelle Mel		
126 échantillons	252 échantillons	504 échantillons
15.56 %	14.92 %	14.60 %
Échelle linéaire		
126 échantillons	252 échantillons	504 échantillons
10.79 %	12.70 %	15.24 %

TAB. A.2: *Différentes conditions d'analyses sur la base FTIMIT63. Les résultats sont donnés en pourcentages d'erreurs d'identification.*

Matrices de covariance ($q = 0$)		
Échelle Mel		
126 échantillons	252 échantillons	504 échantillons
49.21 %	42.22 %	**40.32 %**
Échelle linéaire		
126 échantillons	252 échantillons	504 échantillons
51.11 %	47.94 %	46.98 %
Matrices bloc-Toeplitz d'ordre 3 ($q = 1$)		
Échelle Mel		
126 échantillons	252 échantillons	504 échantillons
46.98 %	**40.32 %**	40.95 %
Échelle linéaire		
126 échantillons	252 échantillons	504 échantillons
42.86 %	44.44 %	42.22 %
Matrices bloc-Toeplitz d'ordre 5 ($q = 2$)		
Échelle Mel		
126 échantillons	252 échantillons	504 échantillons
47.94 %	**42.22 %**	43.49 %
Échelle linéaire		
126 échantillons	252 échantillons	504 échantillons
45.71 %	46.35 %	45.71 %

TAB. A.3: *Différentes conditions d'analyses sur la base NTIMIT63. Les résultats sont donnés en pourcentages d'erreurs d'identification.*

A.4 Discussion

Nous pouvons faire plusieurs remarques sur ces différents résultats :

- Sur la base TIMIT63, les meilleurs résultats sont à chaque fois obtenus avec l'échelle linéaire, et plus particulièrement avec une fenêtre d'analyse de 252 échantillons (15,75 ms). En particulier, nous obtenons 0 % d'erreurs en utilisant les matrices de covariance ($q = 0$). Ce score est meilleur que le score de référence (0.63 %).
- Sur la base FTIMIT63 également, les meilleurs résultats sont obtenus avec l'échelle linéaire. A l'exception des expériences sur les matrices de covariance ($q = 0$) pour lesquelles une fenêtre de 252 échantillons (15,75 ms) semble préférable, une petite fenêtre de 126 échantillons (7,88 ms) donnent les meilleurs résultats. Le meilleur score, 8.89 % d'erreurs, est obtenu sur les matrices bloc-Toeplitz d'ordre 3 ($q = 1$) avec une fenêtre de 126 échantillons. Ce score est meilleur que le score de référence (12.06 %).
- La tendance est complètement inversée sur la base NTIMIT63. En effet, les meilleurs résultats sont toujours obtenus avec une échelle Mel. De plus, excepté pour les expériences sur les matrices de covariance ($q = 0$), où une fenêtre de 252 échantillons (15,75 ms) semble préférable, la fenêtre de 504 échantillons (31,5 ms) donne les meilleurs résultats dans les deux autres cas. A noter que le meilleur résultat, 40.32 %, est obtenu pour $q = 0$ et une fenêtre de 504 échantillons, et pour $q = 1$ et une fenêtre de 252 échantillons. Ce score est égal au score de référence.

A.5 Conclusions

☞ Il semble préférable d'utiliser des échelles d'analyses linéaires sur des données de bonne qualité (TIMIT63) ou même filtrées en fréquence (FTIMIT63). En revanche, dans le cas de parole téléphonique, avec en outre la variabilité du canal (NTIMIT63), il vaut mieux choisir une échelle Mel.

☞ D'autre part, une longueur de fenêtre d'analyse plus petite est plus appropriée pour la parole de bonne qualité (252 échantillons pour TIMIT63) ou même filtrée (126 échantillons pour FTIMIT63), alors qu'il est préférable d'utiliser une longueur de fenêtre plus grande pour la parole téléphonique (504 échantillons pour NTIMIT63).

Annexe B

Résultats supplémentaires

Résultats pour les mesures μ_{D1c} et μ_{D2c}

Nous donnons ici les résultats pour les mesures μ_{D1c} et μ_{D2c}. Nous avons mis aussi dans les tableaux ceux de la mesure μ_{Dc}, déjà donnés précédemment, afin de pouvoir comparer ces trois mesures entre elles. Les résultats sont donnés pour les trois bases : Tab. B.1 pour la base Timit, Tab. B.2 pour la base Ftimit, et Tab. B.3 pour la base Ntimit,

Mesures		μ_{D1c}		μ_{D2c}		μ_{Dc}	
$\mu(\mathcal{X},\mathcal{Y})$	$\mu(\mathcal{Y},\mathcal{X})$	0.3 %	0.3 %	3.8 %	0.6 %	0.5 %	0.2 %
$\mu_{[0.5]}(\mathcal{X},\mathcal{Y})$		0.0 %		0.3 %		0.0 %	
$\mu_{[\alpha_{MN}]}(\mathcal{X},\mathcal{Y})$		0.0 %		0.3 %		0.0 %	
$\mu_{[\beta_{MN}]}(\mathcal{X},\mathcal{Y})$		0.0 %		0.3 %		0.0 %	

Protocole 5.5
apprentissage long (5 phrases $\approx$ 14.4 s) – test long (5 phrases $\approx$ 15.9 s)
$\bar{T} \approx 3000\ cs$, $\bar{\rho} \approx 1.10$, $\bar{\alpha}_{MN} \approx 0.48$, $\bar{\beta}_{MN} \approx 0.49$

Mesures		μ_{D1c}		μ_{D2c}		μ_{Dc}	
$\mu(\mathcal{X},\mathcal{Y})$	$\mu(\mathcal{Y},\mathcal{X})$	33.6 %	12.4 %	32.9 %	18.1 %	26.7 %	7.9 %
$\mu_{[0.5]}(\mathcal{X},\mathcal{Y})$		6.4 %		8.6 %		4.9 %	
$\mu_{[\alpha_{MN}]}(\mathcal{X},\mathcal{Y})$		4.9 %		7.9 %		3.0 %	
$\mu_{[\beta_{MN}]}(\mathcal{X},\mathcal{Y})$		4.6 %		9.2 %		3.0 %	

Protocole 2.5
apprentissage court (2 phrases $\approx$ 5.7 s) – test long (5 phrases $\approx$ 15.9 s)
$\bar{T} \approx 2150\ cs$, $\bar{\rho} \approx 2.79$, $\bar{\alpha}_{MN} \approx 0.26$, $\bar{\beta}_{MN} \approx 0.37$

Mesures		μ_{D1c}		μ_{D2c}		μ_{Dc}	
$\mu(\mathcal{X},\mathcal{Y})$	$\mu(\mathcal{Y},\mathcal{X})$	21.6 %	56.2 %	38.9 %	51.0 %	16.4 %	40.6 %
$\mu_{[0.5]}(\mathcal{X},\mathcal{Y})$		5.1 %		15.8 %		2.7 %	
$\mu_{[\alpha_{MN}]}(\mathcal{X},\mathcal{Y})$		3.3 %		16.5 %		2.4 %	
$\mu_{[\beta_{MN}]}(\mathcal{X},\mathcal{Y})$		7.2 %		23.0 %		5.2 %	

Protocole 5.1
apprentissage long (5 phrases $\approx$ 14.4 s) – test court (1 sentence $\approx$ 3.2 s)
$\bar{T} \approx 1750\ cs$, $\bar{\rho} \approx 0.22$, $\bar{\alpha}_{MN} \approx 0.82$, $\bar{\beta}_{MN} \approx 0.68$

Mesures		μ_{D1c}		μ_{D2c}		μ_{Dc}	
$\mu(\mathcal{X},\mathcal{Y})$	$\mu(\mathcal{Y},\mathcal{X})$	55.4 %	66.4 %	61.3 %	65.6 %	47.1 %	54.6 %
$\mu_{[0.5]}(\mathcal{X},\mathcal{Y})$		19.9 %		35.6 %		15.6 %	
$\mu_{[\alpha_{MN}]}(\mathcal{X},\mathcal{Y})$		19.9 %		35.6 %		15.8 %	
$\mu_{[\beta_{MN}]}(\mathcal{X},\mathcal{Y})$		23.9 %		38.2 %		19.9 %	

Protocole 2.1
apprentissage court (2 phrases $\approx$ 5.7 s) – test court (1 sentence $\approx$ 3.2 s)
$\bar{T} \approx\ 900\ cs$, $\bar{\rho} \approx 0.56$, $\bar{\alpha}_{MN} \approx 0.64$, $\bar{\beta}_{MN} \approx 0.57$

TAB. B.1: *Méthodes statistiques du second ordre du type μ_{Dc}. Identification du locuteur indépendante du texte. Base de données TIMIT. Les résultats sont donnés en pourcentage d'erreur d'identification.*

Mesures		μ_{D1c}		μ_{D2c}		μ_{Dc}	
$\mu(\mathcal{X},\mathcal{Y})$	$\mu(\mathcal{Y},\mathcal{X})$	10.5 %	5.6 %	27.8 %	15.4 %	9.5 %	4.6 %
$\mu_{[0.5]}(\mathcal{X},\mathcal{Y})$		1.6 %		7.0 %		1.7 %	
$\mu_{[\alpha_{MN}]}(\mathcal{X},\mathcal{Y})$		1.4 %		7.0 %		1.4 %	
$\mu_{[\beta_{MN}]}(\mathcal{X},\mathcal{Y})$		1.4 %		7.1 %		1.6 %	

Protocole 5.5
apprentissage long (5 phrases $\approx$ 14.4 s) – test long (5 phrases $\approx$ 15.9 s)
$\bar{T} \approx 3000\ cs$, $\bar{\rho} \approx 1.10$, $\bar{\alpha}_{MN} \approx 0.48$, $\bar{\beta}_{MN} \approx 0.49$

Mesures		μ_{D1c}		μ_{D2c}		μ_{Dc}	
$\mu(\mathcal{X},\mathcal{Y})$	$\mu(\mathcal{Y},\mathcal{X})$	60.5 %	42.4 %	66.8 %	52.5 %	55.9 %	34.4 %
$\mu_{[0.5]}(\mathcal{X},\mathcal{Y})$		30.3 %		38.2 %		23.3 %	
$\mu_{[\alpha_{MN}]}(\mathcal{X},\mathcal{Y})$		28.4 %		35.4 %		22.4 %	
$\mu_{[\beta_{MN}]}(\mathcal{X},\mathcal{Y})$		29.7 %		36.8 %		23.8 %	

Protocole 2.5
apprentissage court (2 phrases $\approx$ 5.7 s) – test long (5 phrases $\approx$ 15.9 s)
$\bar{T} \approx 2150\ cs$, $\bar{\rho} \approx 2.79$, $\bar{\alpha}_{MN} \approx 0.26$, $\bar{\beta}_{MN} \approx 0.37$

Mesures		μ_{D1c}		μ_{D2c}		μ_{Dc}	
$\mu(\mathcal{X},\mathcal{Y})$	$\mu(\mathcal{Y},\mathcal{X})$	53.6 %	73.2 %	74.3 %	77.5 %	51.9 %	66.7 %
$\mu_{[0.5]}(\mathcal{X},\mathcal{Y})$		35.9 %		51.9 %		27.8 %	
$\mu_{[\alpha_{MN}]}(\mathcal{X},\mathcal{Y})$		32.4 %		55.2 %		26.9 %	
$\mu_{[\beta_{MN}]}(\mathcal{X},\mathcal{Y})$		39.2 %		63.1 %		35.6 %	

Protocole 5.1
apprentissage long (5 phrases $\approx$ 14.4 s) – test court (1 sentence $\approx$ 3.2 s)
$\bar{T} \approx 1750\ cs$, $\bar{\rho} \approx 0.22$, $\bar{\alpha}_{MN} \approx 0.82$, $\bar{\beta}_{MN} \approx 0.68$

Mesures		μ_{D1c}		μ_{D2c}		μ_{Dc}	
$\mu(\mathcal{X},\mathcal{Y})$	$\mu(\mathcal{Y},\mathcal{X})$	78.9 %	83.8 %	86.5 %	87.3 %	76.9 %	79.4 %
$\mu_{[0.5]}(\mathcal{X},\mathcal{Y})$		59.8 %		71.7 %		53.5 %	
$\mu_{[\alpha_{MN}]}(\mathcal{X},\mathcal{Y})$		58.9 %		72.4 %		53.2 %	
$\mu_{[\beta_{MN}]}(\mathcal{X},\mathcal{Y})$		61.5 %		74.8 %		56.4 %	

Protocole 2.1
apprentissage court (2 phrases $\approx$ 5.7 s) – test court (1 sentence $\approx$ 3.2 s)
$\bar{T} \approx\ 900\ cs$, $\bar{\rho} \approx 0.56$, $\bar{\alpha}_{MN} \approx 0.64$, $\bar{\beta}_{MN} \approx 0.57$

TAB. B.2: *Méthodes statistiques du second ordre du type μ_{Dc}. Identification du locuteur indépendante du texte. Base de données FTIMIT. Les résultats sont donnés en pourcentage d'erreur d'identification.*

Mesures		μ_{D1c}		μ_{D2c}		μ_{Dc}	
$\mu(\mathcal{X},\mathcal{Y})$	$\mu(\mathcal{Y},\mathcal{X})$	61.5 %	50.0 %	74.4 %	71.3 %	59.0 %	49.0 %
$\mu_{[0.5]}(\mathcal{X},\mathcal{Y})$		38.6 %		51.8 %		32.1 %	
$\mu_{[\alpha_{MN}]}(\mathcal{X},\mathcal{Y})$		39.0 %		51.3 %		31.6 %	
$\mu_{[\beta_{MN}]}(\mathcal{X},\mathcal{Y})$		38.9 %		51.3 %		31.4 %	

Protocole 5.5
apprentissage long (5 phrases $\approx$ 14.4 s) – test long (5 phrases $\approx$ 15.9 s)
$\bar{T} \approx 3000\ cs$, $\bar{\rho} \approx 1.10$, $\bar{\alpha}_{MN} \approx 0.48$, $\bar{\beta}_{MN} \approx 0.49$

Mesures		μ_{D1c}		μ_{D2c}		μ_{Dc}	
$\mu(\mathcal{X},\mathcal{Y})$	$\mu(\mathcal{Y},\mathcal{X})$	88.2 %	78.4 %	92.2 %	90.7 %	87.6 %	77.5 %
$\mu_{[0.5]}(\mathcal{X},\mathcal{Y})$		74.4 %		81.0 %		69.7 %	
$\mu_{[\alpha_{MN}]}(\mathcal{X},\mathcal{Y})$		71.8 %		80.6 %		69.2 %	
$\mu_{[\beta_{MN}]}(\mathcal{X},\mathcal{Y})$		71.8 %		80.2 %		69.2 %	

Protocole 2.5
apprentissage court (2 phrases $\approx$ 5.7 s) – test long (5 phrases $\approx$ 15.9 s)
$\bar{T} \approx 2150\ cs$, $\bar{\rho} \approx 2.79$, $\bar{\alpha}_{MN} \approx 0.26$, $\bar{\beta}_{MN} \approx 0.37$

Mesures		μ_{D1c}		μ_{D2c}		μ_{Dc}	
$\mu(\mathcal{X},\mathcal{Y})$	$\mu(\mathcal{Y},\mathcal{X})$	86.9 %	95.5 %	92.9 %	95.8 %	85.9 %	94.8 %
$\mu_{[0.5]}(\mathcal{X},\mathcal{Y})$		80.0 %		84.1 %		74.8 %	
$\mu_{[\alpha_{MN}]}(\mathcal{X},\mathcal{Y})$		79.0 %		85.1 %		73.6 %	
$\mu_{[\beta_{MN}]}(\mathcal{X},\mathcal{Y})$		81.4 %		88.4 %		79.5 %	

Protocole 5.1
apprentissage long (5 phrases $\approx$ 14.4 s) – test court (1 sentence $\approx$ 3.2 s)
$\bar{T} \approx 1750\ cs$, $\bar{\rho} \approx 0.22$, $\bar{\alpha}_{MN} \approx 0.82$, $\bar{\beta}_{MN} \approx 0.68$

Mesures		μ_{D1c}		μ_{D2c}		μ_{Dc}	
$\mu(\mathcal{X},\mathcal{Y})$	$\mu(\mathcal{Y},\mathcal{X})$	94.1 %	97.1 %	96.6 %	97.9 %	93.6 %	96.9 %
$\mu_{[0.5]}(\mathcal{X},\mathcal{Y})$		88.8 %		90.6 %		85.6 %	
$\mu_{[\alpha_{MN}]}(\mathcal{X},\mathcal{Y})$		88.6 %		90.7 %		85.2 %	
$\mu_{[\beta_{MN}]}(\mathcal{X},\mathcal{Y})$		89.8 %		91.7 %		87.3 %	

Protocole 2.1
apprentissage court (2 phrases $\approx$ 5.7 s) – test court (1 sentence $\approx$ 3.2 s)
$\bar{T} \approx\ 900\ cs$, $\bar{\rho} \approx 0.56$, $\bar{\alpha}_{MN} \approx 0.64$, $\bar{\beta}_{MN} \approx 0.57$

TAB. B.3: *Méthodes statistiques du second ordre du type μ_{Dc}. Identification du locuteur indépendante du texte. Base de données NTIMIT. Les résultats sont donnés en pourcentage d'erreur d'identification.*

Annexe C

Calcul des coefficients matriciels d'un modèle AR-vectoriel

Dans cette annexe, nous détaillons le calcul qui permet d'obtenir les coefficients matriciels d'un modèle AR-vectoriel. Nous traitons uniquement le cas d'un modèle AR-vectoriel d'ordre 2. Le point de départ est l'expression de l'erreur résiduelle de prédiction :

$$\mathbf{e}_t = \mathbf{x}^*_t + \mathbf{A}_1 \cdot \mathbf{x}^*_{t-1} + \mathbf{A}_2 \cdot \mathbf{x}^*_{t-2}$$

Calculons la vraie matrice de covariance de ce vecteur aléatoire :

$$\begin{aligned}
&E[\mathbf{e}_t \cdot \mathbf{e}_t^T] = \\
&E[\mathbf{x}^*_t \cdot \mathbf{x}^{*\,T}_t] + E[\mathbf{x}^*_t \cdot \mathbf{x}^{*\,T}_{t-1}] \cdot \mathbf{A}_1^T + E[\mathbf{x}^*_t \cdot \mathbf{x}^{*\,T}_{t-2}] \cdot \mathbf{A}_2^T \\
&+ \mathbf{A}_1 \cdot E[\mathbf{x}^*_{t-1} \cdot \mathbf{x}^{*\,T}_t] + \mathbf{A}_1 \cdot E[\mathbf{x}^*_{t-1} \cdot \mathbf{x}^{*\,T}_{t-1}] \cdot \mathbf{A}_1^T + \mathbf{A}_1 \cdot E[\mathbf{x}^*_{t-1} \cdot \mathbf{x}^{*\,T}_{t-2}] \cdot \mathbf{A}_2^T \\
&+ \mathbf{A}_2 \cdot E[\mathbf{x}^*_{t-2} \cdot \mathbf{x}^{*\,T}_t] + \mathbf{A}_2 \cdot E[\mathbf{x}^*_{t-2} \cdot \mathbf{x}^{*\,T}_{t-1}] \cdot \mathbf{A}_1^T + \mathbf{A}_2 \cdot E[\mathbf{x}^*_{t-2} \cdot \mathbf{x}^{*\,T}_{t-2}] \cdot \mathbf{A}_2^T
\end{aligned}$$

En notant alors :

$$\begin{aligned}
\boldsymbol{\mathcal{X}}_0 &= E[\mathbf{x}^*_t \cdot \mathbf{x}^{*\,T}_t] \\
\boldsymbol{\mathcal{X}}_1 &= E[\mathbf{x}^*_t \cdot \mathbf{x}^{*\,T}_{t-1}] \\
\boldsymbol{\mathcal{X}}_2 &= E[\mathbf{x}^*_t \cdot \mathbf{x}^{*\,T}_{t-2}]
\end{aligned}$$

l'équation précédente devient plus simplement :

$$\begin{aligned}
&E[\mathbf{e}_t \cdot \mathbf{e}_t^T] = \\
&\boldsymbol{\mathcal{X}}_0 + \boldsymbol{\mathcal{X}}_1 \cdot \mathbf{A}_1^T + \boldsymbol{\mathcal{X}}_2 \cdot \mathbf{A}_2^T \\
&+ \mathbf{A}_1 \cdot \boldsymbol{\mathcal{X}}_1^T + \mathbf{A}_1 \cdot \boldsymbol{\mathcal{X}}_0 \cdot \mathbf{A}_1^T + \mathbf{A}_1 \cdot \boldsymbol{\mathcal{X}}_1 \cdot \mathbf{A}_2^T \\
&+ \mathbf{A}_2 \cdot \boldsymbol{\mathcal{X}}_2^T + \mathbf{A}_2 \cdot \boldsymbol{\mathcal{X}}_1^T \cdot \mathbf{A}_1^T + \mathbf{A}_2 \cdot \boldsymbol{\mathcal{X}}_0 \cdot \mathbf{A}_2^T
\end{aligned}$$

Pour obtenir alors les coefficients du modèle, on choisit de minimiser la trace de la matrice de covariance de l'erreur résiduelle de prédiction. Pour minimiser cette trace, on la dérive par rapport à chaque coefficient, et on annule ces dérivées :

$$\begin{aligned}\frac{\partial\ tr\ E[\mathbf{e}_t \cdot \mathbf{e}_t^T]}{\partial \mathbf{A}_1} &= tr\left(\frac{\partial E[\mathbf{e}_t \cdot \mathbf{e}_t^T]}{\partial \mathbf{A}_1}\right) \\ &= tr\left(\boldsymbol{\mathcal{X}}_1 + \boldsymbol{\mathcal{X}}_1^T + \boldsymbol{\mathcal{X}}_0 \cdot \mathbf{A}_1^T + \mathbf{A}_1 \cdot \boldsymbol{\mathcal{X}}_0 + \boldsymbol{\mathcal{X}}_1 \cdot \mathbf{A}_2^T + \mathbf{A}_2 \cdot \boldsymbol{\mathcal{X}}_1^T\right) \\ &= 2\ tr\left(\boldsymbol{\mathcal{X}}_1 + \mathbf{A}_1 \cdot \boldsymbol{\mathcal{X}}_0 + \mathbf{A}_2 \cdot \boldsymbol{\mathcal{X}}_1^T\right) \\ &= 2\ tr\left(\left[\ \mathbf{I}\ \mathbf{A}_1\ \mathbf{A}_2\ \right] \begin{bmatrix} \boldsymbol{\mathcal{X}}_1 \\ \boldsymbol{\mathcal{X}}_0 \\ \boldsymbol{\mathcal{X}}_1^T \end{bmatrix}\right)\end{aligned}$$

$$\begin{aligned}\frac{\partial\ tr\ E[\mathbf{e}_t \cdot \mathbf{e}_t^T]}{\partial \mathbf{A}_2} &= tr\left(\frac{\partial E[\mathbf{e}_t \cdot \mathbf{e}_t^T]}{\partial \mathbf{A}_2}\right) \\ &= tr\left(\boldsymbol{\mathcal{X}}_2 + \boldsymbol{\mathcal{X}}_2^T + \boldsymbol{\mathcal{X}}_0 \cdot \mathbf{A}_2^T + \mathbf{A}_2 \cdot \boldsymbol{\mathcal{X}}_0 + \boldsymbol{\mathcal{X}}_1^T \cdot \mathbf{A}_1^T + \mathbf{A}_1 \cdot \boldsymbol{\mathcal{X}}_1\right) \\ &= 2\ tr\left(\boldsymbol{\mathcal{X}}_2 + \mathbf{A}_1 \cdot \boldsymbol{\mathcal{X}}_1 + \mathbf{A}_2 \cdot \boldsymbol{\mathcal{X}}_0\right) \\ &= 2\ tr\left(\left[\ \mathbf{I}\ \mathbf{A}_1\ \mathbf{A}_2\ \right] \begin{bmatrix} \boldsymbol{\mathcal{X}}_2 \\ \boldsymbol{\mathcal{X}}_1 \\ \boldsymbol{\mathcal{X}}_0 \end{bmatrix}\right)\end{aligned}$$

La condition

$$\left[\ \mathbf{I}\ \mathbf{A}_1\ \mathbf{A}_2\ \right] \begin{bmatrix} \boldsymbol{\mathcal{X}}_1 & \boldsymbol{\mathcal{X}}_2 \\ \boldsymbol{\mathcal{X}}_0 & \boldsymbol{\mathcal{X}}_1 \\ \boldsymbol{\mathcal{X}}_1^T & \boldsymbol{\mathcal{X}}_0 \end{bmatrix} = \mathbf{0}$$

annule alors simultanément les deux traces précédentes. Elle se réécrit plus simplement selon :

$$\left[\ \mathbf{A}_1\ \mathbf{A}_2\ \right] \begin{bmatrix} \boldsymbol{\mathcal{X}}_0 & \boldsymbol{\mathcal{X}}_1 \\ \boldsymbol{\mathcal{X}}_1^T & \boldsymbol{\mathcal{X}}_0 \end{bmatrix} = -\left[\ \boldsymbol{\mathcal{X}}_1\ \boldsymbol{\mathcal{X}}_2\ \right]$$

Dans le cas où l'on ne dispose pas des vraies matrices de covariance, il suffit de les remplacer dans la formule précédente par leurs estimations.

Annexe D

Invariance des mesures par filtrage inversible

Soit $\mathbf{H}$ une matrice de filtrage inversible. On a alors :

$$
\begin{aligned}
& tr\left[(\mathbf{H}\cdot\mathbf{X}_{2q+1}\cdot\mathbf{H}^T)^{-\frac{1}{2}}\cdot(\mathbf{H}\cdot\mathbf{Y}_{2q+1}\cdot\mathbf{H}^T)\cdot(\mathbf{H}\cdot\mathbf{X}_{2q+1}\cdot\mathbf{H}^T)^{-\frac{1}{2}}\right] \\
= \; & tr\left[(\mathbf{H}\cdot\mathbf{Y}_{2q+1}\cdot\mathbf{H}^T)\cdot(\mathbf{H}\cdot\mathbf{X}_{2q+1}\cdot\mathbf{H}^T)^{-\frac{1}{2}}\cdot(\mathbf{H}\cdot\mathbf{X}_{2q+1}\cdot\mathbf{H}^T)^{-\frac{1}{2}}\right] \\
= \; & tr\left[(\mathbf{H}\cdot\mathbf{Y}_{2q+1}\cdot\mathbf{H}^T)\cdot(\mathbf{H}\cdot\mathbf{X}_{2q+1}\cdot\mathbf{H}^T)^{-1}\right] \\
= \; & tr\left[\mathbf{H}\cdot\mathbf{Y}_{2q+1}\cdot(\mathbf{H}^T\cdot\mathbf{H}^{-T})\cdot\mathbf{X}_{2q+1}^{-1}\cdot\mathbf{H}^{-1}\right] \\
= \; & tr\left[\mathbf{H}\cdot\mathbf{Y}_{2q+1}\cdot\mathbf{X}_{2q+1}^{-1}\cdot\mathbf{H}^{-1}\right] \\
= \; & tr\left[\mathbf{Y}_{2q+1}\cdot\mathbf{X}_{2q+1}^{-1}\cdot(\mathbf{H}^{-1}\cdot\mathbf{H})\right] \\
= \; & tr\left[\mathbf{Y}_{2q+1}\cdot\mathbf{X}_{2q+1}^{-1}\right] \\
= \; & tr\left[\mathbf{X}_{2q+1}^{-\frac{1}{2}}\cdot\mathbf{Y}_{2q+1}\cdot\mathbf{X}_{2q+1}^{-\frac{1}{2}}\right]
\end{aligned}
$$

De plus :

$$
\begin{aligned}
& det\left[(\mathbf{H}\cdot\mathbf{X}_{2q+1}\cdot\mathbf{H}^T)^{-\frac{1}{2}}\cdot(\mathbf{H}\cdot\mathbf{Y}_{2q+1}\cdot\mathbf{H}^T)\cdot(\mathbf{H}\cdot\mathbf{X}_{2q+1}\cdot\mathbf{H}^T)^{-\frac{1}{2}}\right] \\
= \;& det\left[(\mathbf{H}\cdot\mathbf{X}_{2q+1}\cdot\mathbf{H}^T)^{-\frac{1}{2}}\right]\cdot det\left[\mathbf{H}\cdot\mathbf{Y}_{2q+1}\cdot\mathbf{H}^T\right]\cdot det\left[(\mathbf{H}\cdot\mathbf{X}_{2q+1}\cdot\mathbf{H}^T)^{-\frac{1}{2}}\right] \\
= \;& \frac{det\left[\mathbf{H}\cdot\mathbf{Y}_{2q+1}\cdot\mathbf{H}^T\right]}{det\left[\mathbf{H}\cdot\mathbf{X}_{2q+1}\cdot\mathbf{H}^T\right]} \\
= \;& \frac{det\left[\mathbf{H}\right]\cdot det\left[\mathbf{Y}_{2q+1}\right]\cdot det\left[\mathbf{H}\right]}{det\left[\mathbf{H}\right]\cdot det\left[\mathbf{X}_{2q+1}\right]\cdot det\left[\mathbf{H}\right]} \\
= \;& \frac{det\left[\mathbf{Y}_{2q+1}\right]}{det\left[\mathbf{X}_{2q+1}\right]} \\
= \;& det\left[\mathbf{Y}_{2q+1}\cdot\mathbf{X}_{2q+1}^{-1}\right] \\
= \;& det\left[\mathbf{X}_{2q+1}^{-\frac{1}{2}}\cdot\mathbf{Y}_{2q+1}\cdot\mathbf{X}_{2q+1}^{-\frac{1}{2}}\right]
\end{aligned}
$$

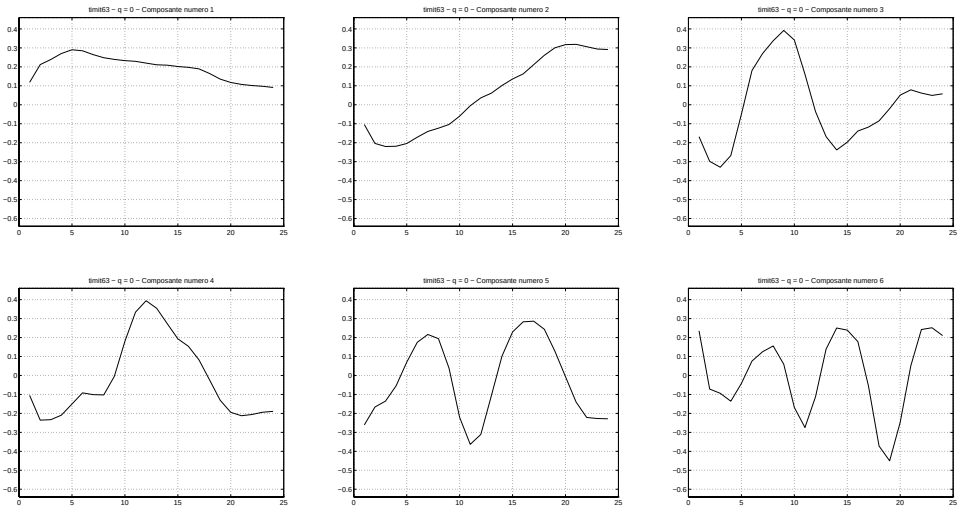
Q.E.D.

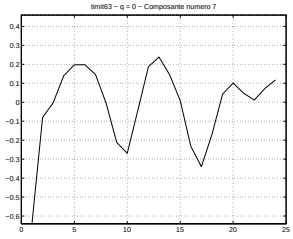
Annexe E

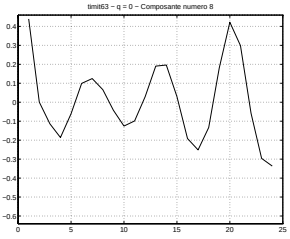
Composantes principales pour la base TIMIT63

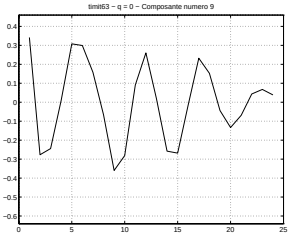
Nous donnons dans cette annexe la représentation graphique de nombreuses composantes principales pour la base TIMIT63. Cela permet notamment de voir quelles sont celles qui ont été choisies dans le chapitre 7, et de comparer leurs formes avec celles qui ont été écartées. Les représentations sont données pour les valeurs de $q = 0$, $q = 1$ et $q = 2$. Enfin, nous ne représentons, pour chaque valeur de q, que les 24 premières composantes.

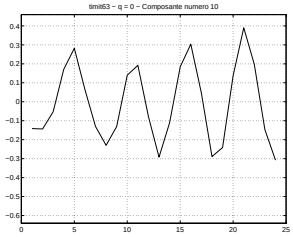
E.1 $q = 0$: 24 composantes principales

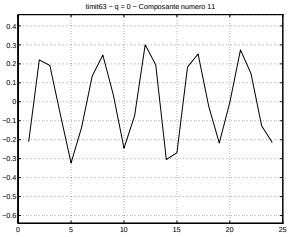


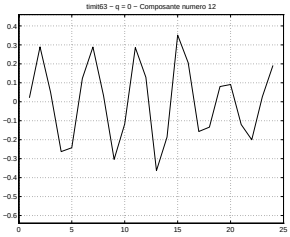

timit63 – q = 0 – Composante numero 7

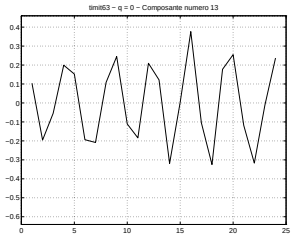

timit63 – q = 0 – Composante numero 8

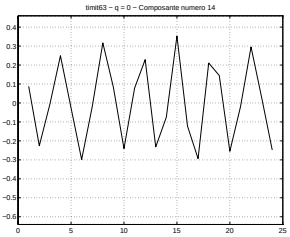

timit63 – q = 0 – Composante numero 9

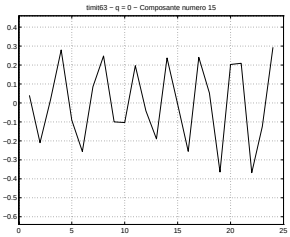

timit63 – q = 0 – Composante numero 10

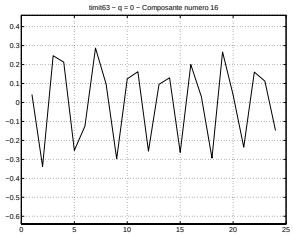

timit63 – q = 0 – Composante numero 11

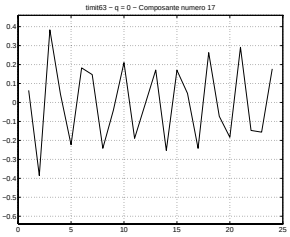

timit63 – q = 0 – Composante numero 12

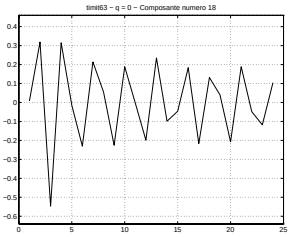

timit63 – q = 0 – Composante numero 13

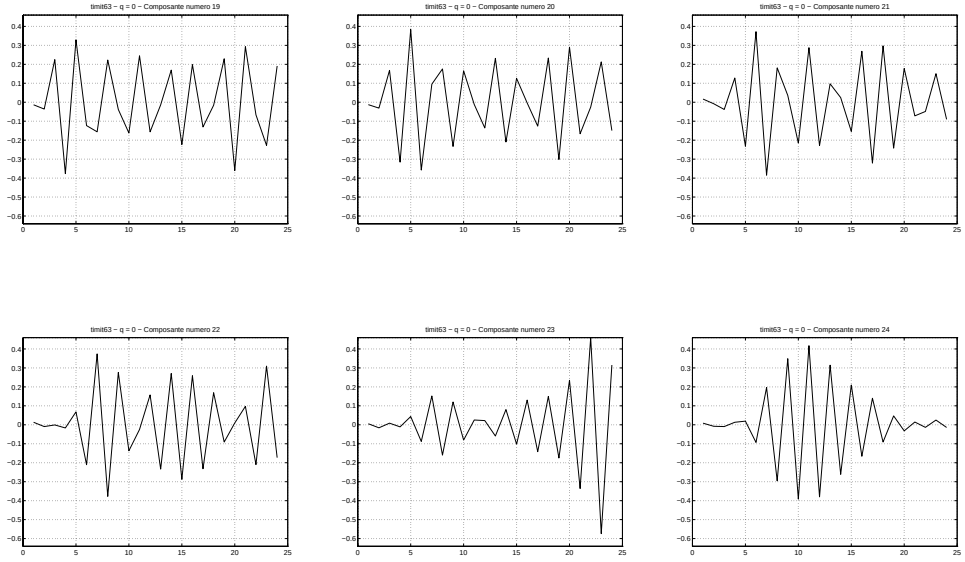

timit63 – q = 0 – Composante numero 14


timit63 – q = 0 – Composante numero 15

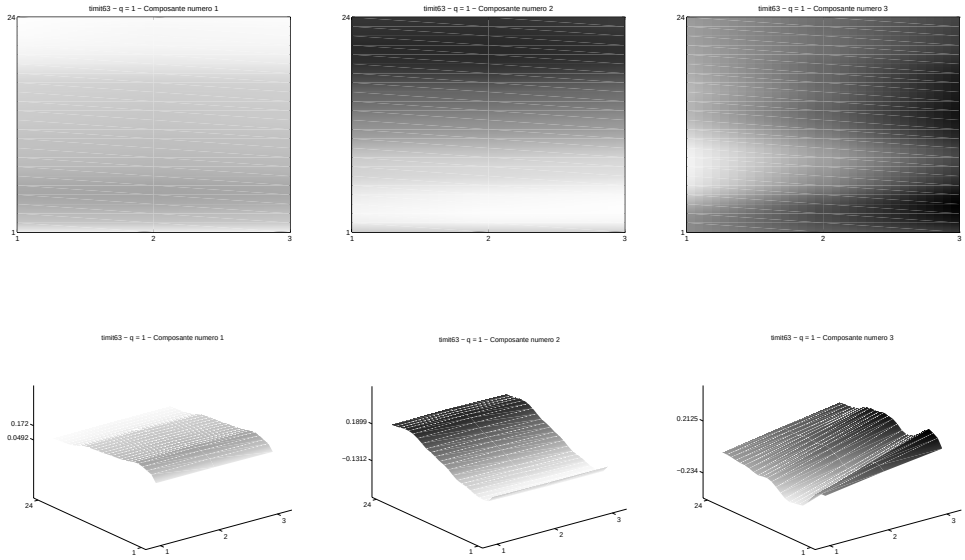

timit63 – q = 0 – Composante numero 16

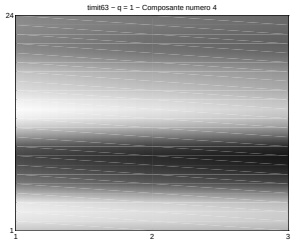

timit63 – q = 0 – Composante numero 17

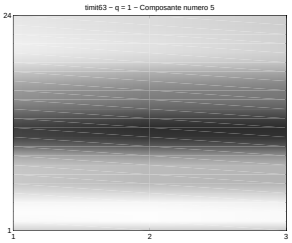

timit63 – q = 0 – Composante numero 18

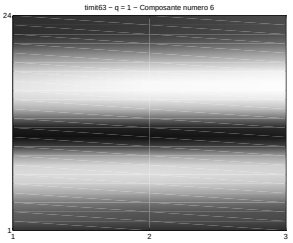


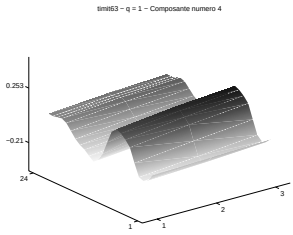
E.2 $q = 1$: 24 premières composantes principales

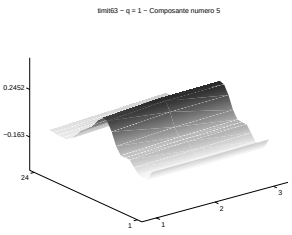


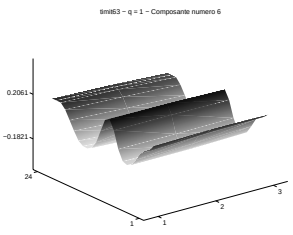

timit63 – q = 1 – Composante numero 4

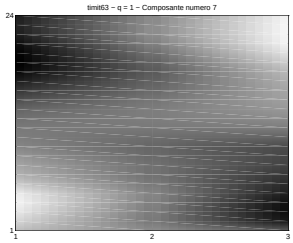

timit63 – q = 1 – Composante numero 5

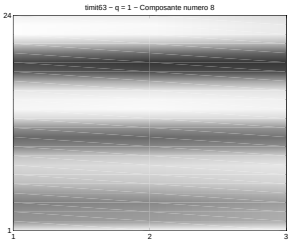

timit63 – q = 1 – Composante numero 6

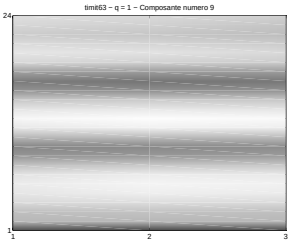

timit63 – q = 1 – Composante numero 4
0.253
−0.21

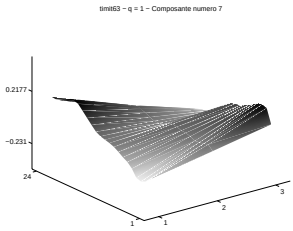

timit63 – q = 1 – Composante numero 5
0.2452
−0.163

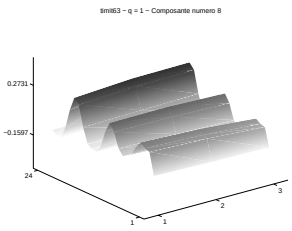

timit63 – q = 1 – Composante numero 6
0.2061
−0.1821

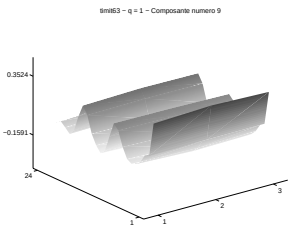

timit63 – q = 1 – Composante numero 7

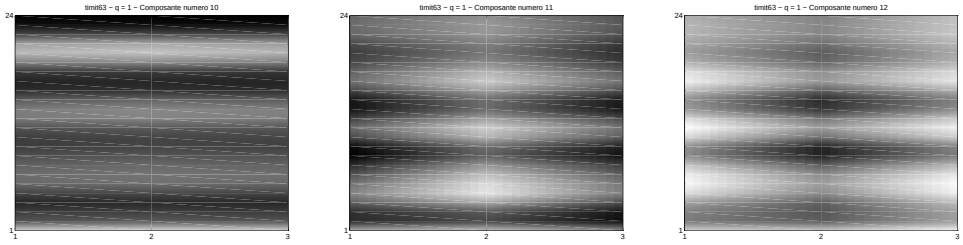

timit63 – q = 1 – Composante numero 8

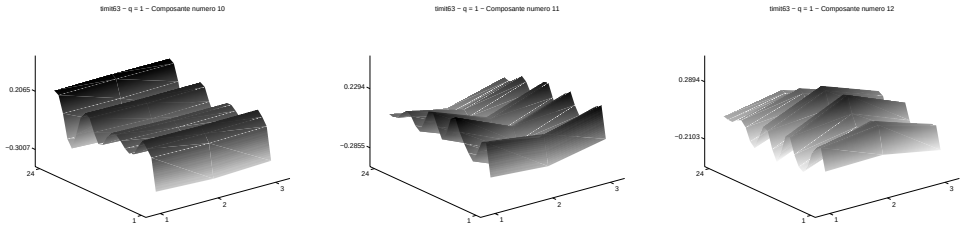

timit63 – q = 1 – Composante numero 9

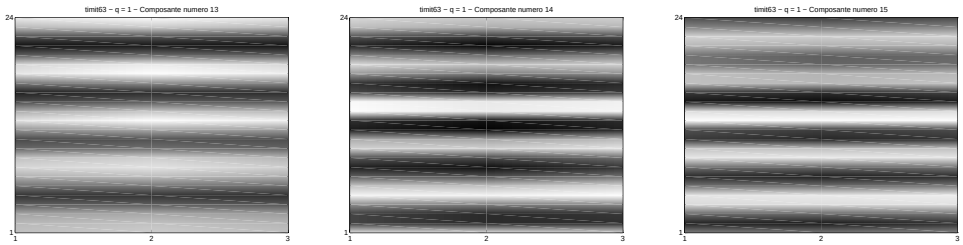

timit63 – q = 1 – Composante numero 7
0.2177
−0.231

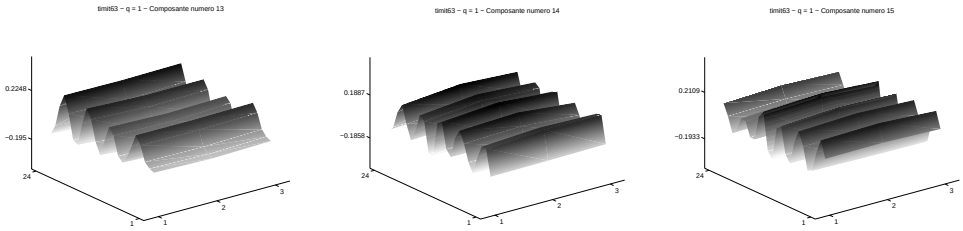

timit63 – q = 1 – Composante numero 8
0.2731
−0.1597

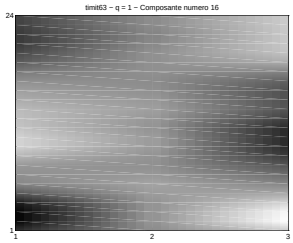

timit63 – q = 1 – Composante numero 9
0.3524
−0.1591

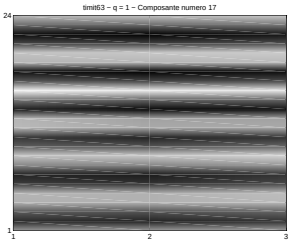

timit63 − q = 1 − Composante numero 10
timit63 − q = 1 − Composante numero 11
timit63 − q = 1 − Composante numero 12

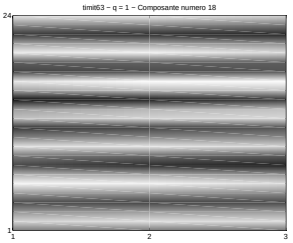

timit63 − q = 1 − Composante numero 10
timit63 − q = 1 − Composante numero 11
timit63 − q = 1 − Composante numero 12

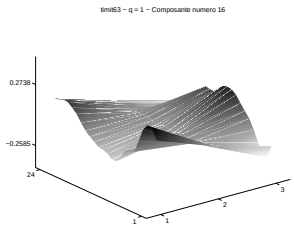

timit63 − q = 1 − Composante numero 13
timit63 − q = 1 − Composante numero 14
timit63 − q = 1 − Composante numero 15

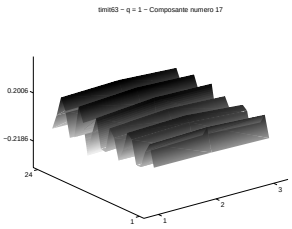

timit63 − q = 1 − Composante numero 13
timit63 − q = 1 − Composante numero 14
timit63 − q = 1 − Composante numero 15

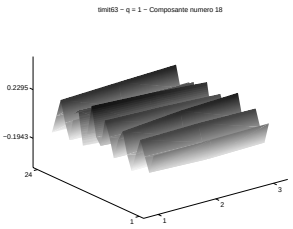

timit63 – q = 1 – Composante numero 16

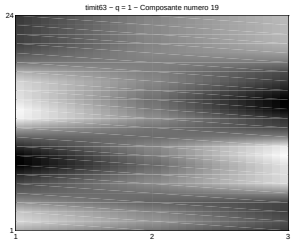

timit63 – q = 1 – Composante numero 17

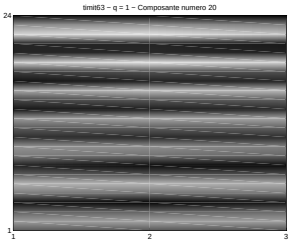

timit63 – q = 1 – Composante numero 18

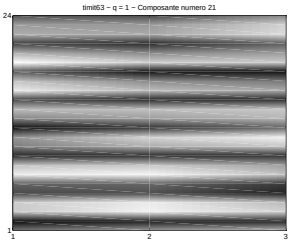

timit63 – q = 1 – Composante numero 16
0.2738
−0.2585

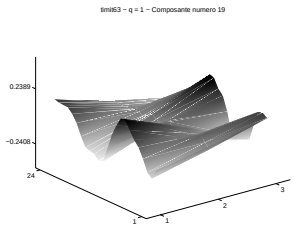

timit63 – q = 1 – Composante numero 17
0.2006
−0.2186

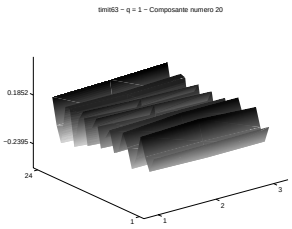

timit63 – q = 1 – Composante numero 18
0.2295
−0.1943

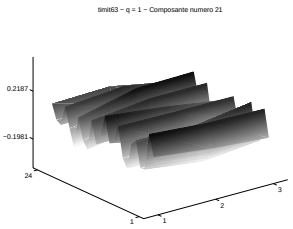

timit63 – q = 1 – Composante numero 19

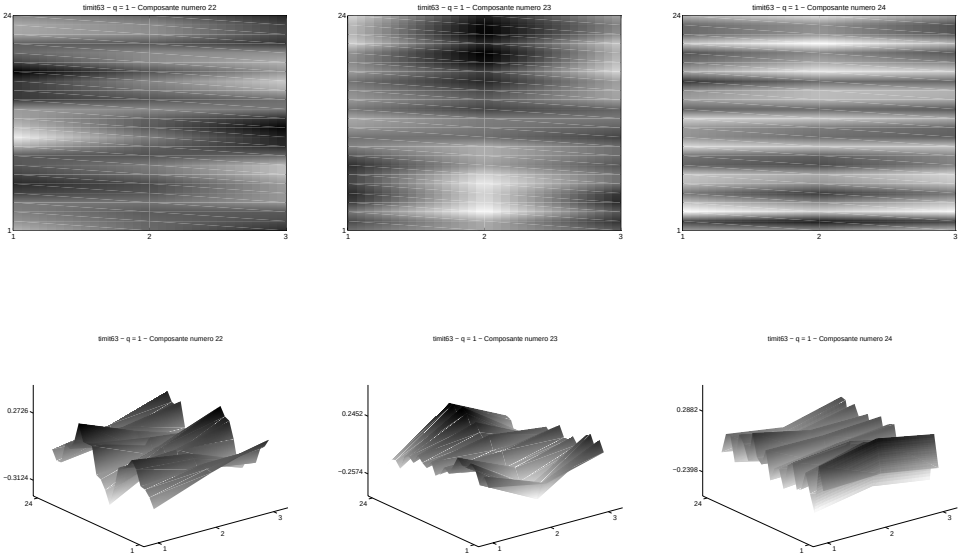

timit63 – q = 1 – Composante numero 20


timit63 – q = 1 – Composante numero 21

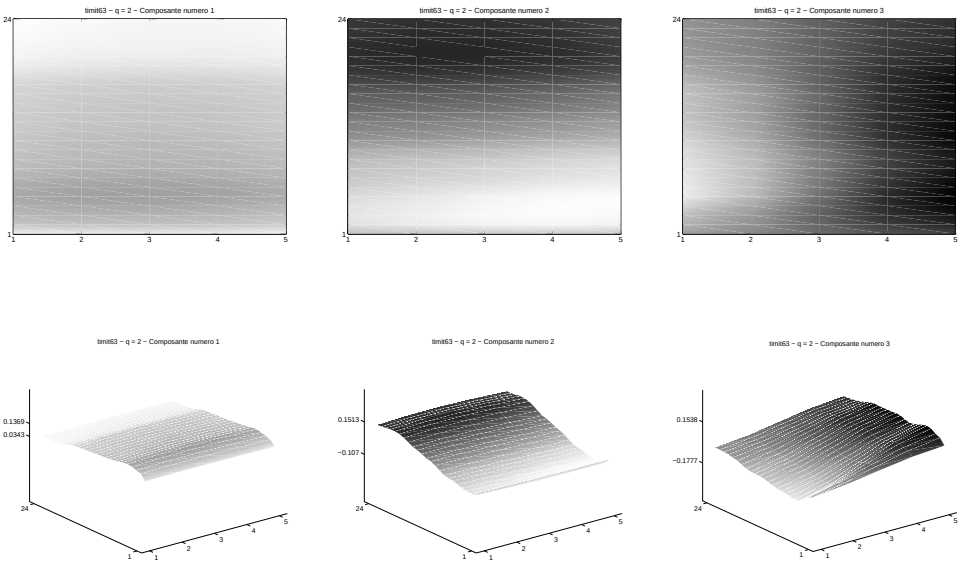

timit63 – q = 1 – Composante numero 19
0.2389
−0.2408

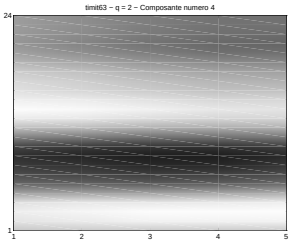

timit63 – q = 1 – Composante numero 20
0.1852
−0.2395

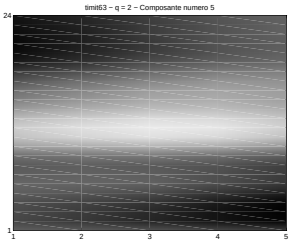

timit63 – q = 1 – Composante numero 21
0.2187
−0.1981

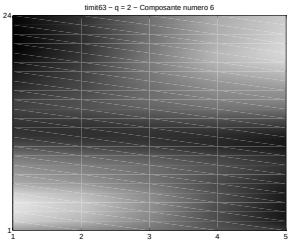


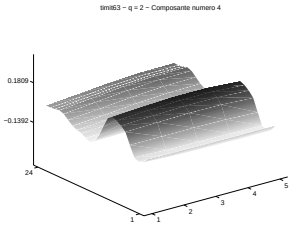
E.3 $q = 2$: 24 premières composantes principales

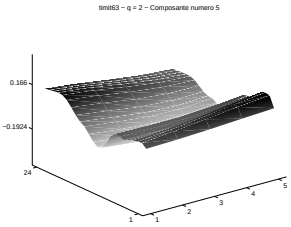


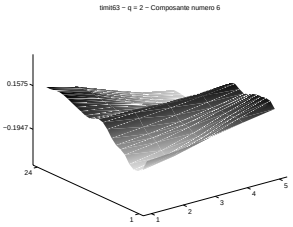

timit63 – q = 2 – Composante numero 4

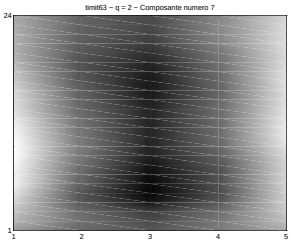

timit63 – q = 2 – Composante numero 5

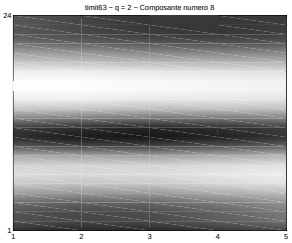

timit63 – q = 2 – Composante numero 6

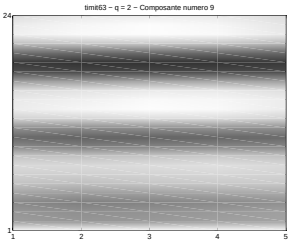

timit63 – q = 2 – Composante numero 4
0.1809
−0.1392

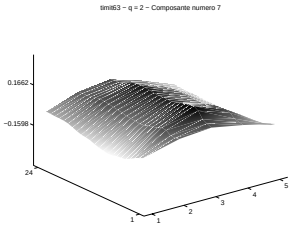

timit63 – q = 2 – Composante numero 5
0.166
−0.1924

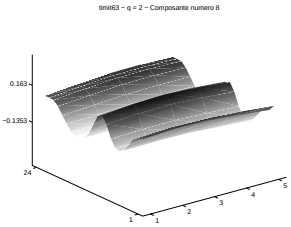

timit63 – q = 2 – Composante numero 6
0.1575
−0.1947

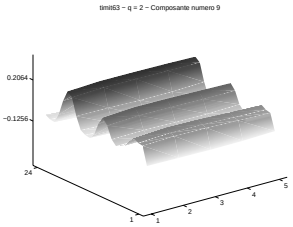

timit63 – q = 2 – Composante numero 7

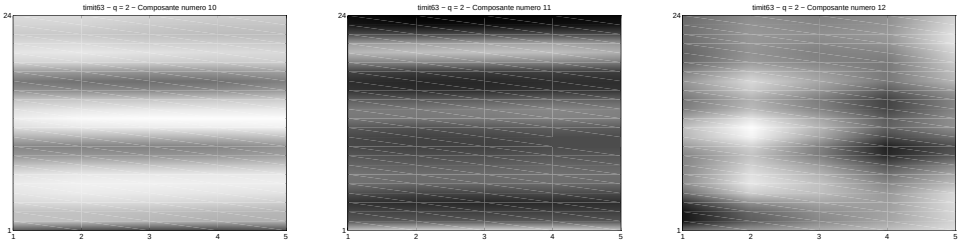

timit63 – q = 2 – Composante numero 8

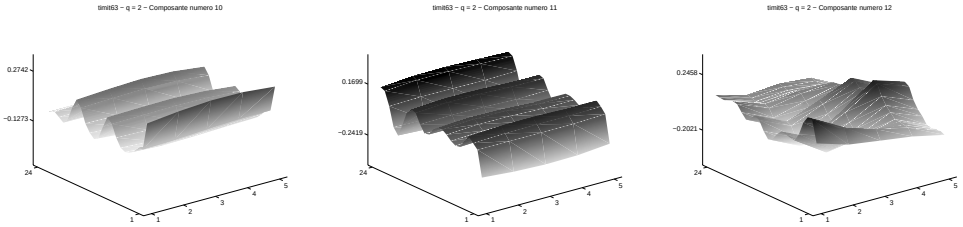

timit63 – q = 2 – Composante numero 9

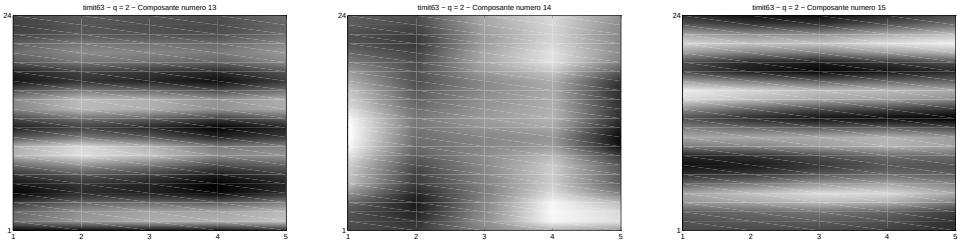

timit63 – q = 2 – Composante numero 7
0.1662
−0.1598

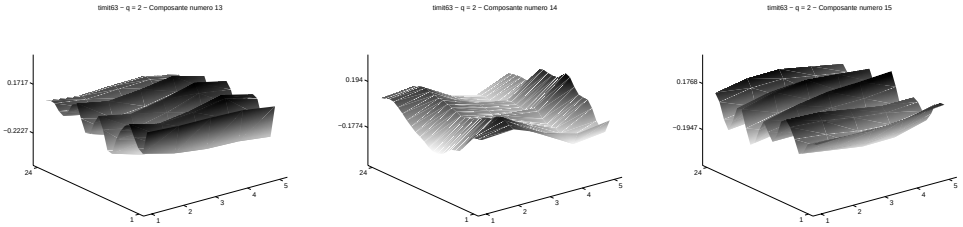

timit63 – q = 2 – Composante numero 8
0.163
−0.1353

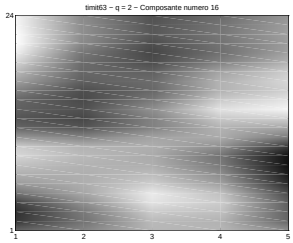

timit63 – q = 2 – Composante numero 9
0.2064
−0.1256

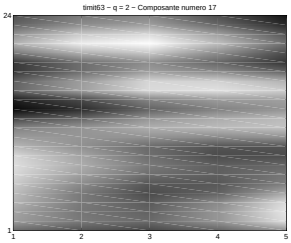

timit63 – q = 2 – Composante numero 10
timit63 – q = 2 – Composante numero 11
timit63 – q = 2 – Composante numero 12

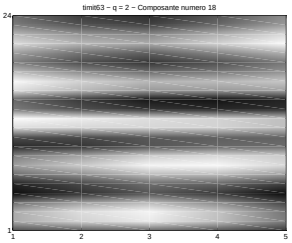

timit63 – q = 2 – Composante numero 10
timit63 – q = 2 – Composante numero 11
timit63 – q = 2 – Composante numero 12

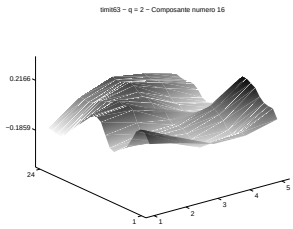

timit63 – q = 2 – Composante numero 13
timit63 – q = 2 – Composante numero 14
timit63 – q = 2 – Composante numero 15

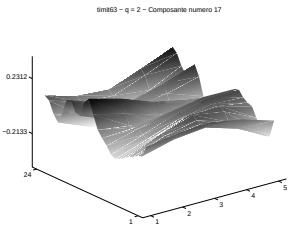

timit63 – q = 2 – Composante numero 13
timit63 – q = 2 – Composante numero 14
timit63 – q = 2 – Composante numero 15

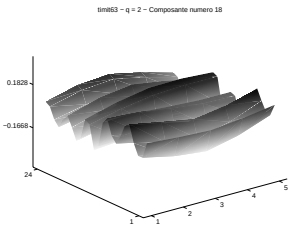

timit63 – q = 2 – Composante numero 16

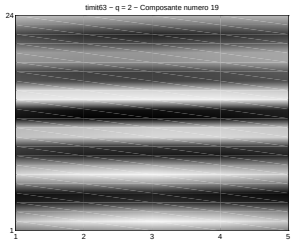

timit63 – q = 2 – Composante numero 17

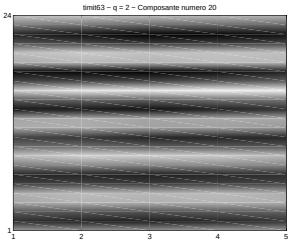

timit63 – q = 2 – Composante numero 18

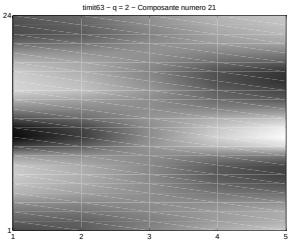

timit63 – q = 2 – Composante numero 16
0.2166
−0.1859

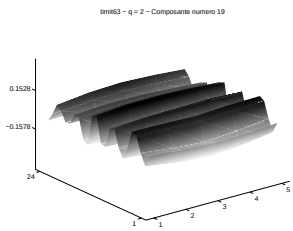

timit63 – q = 2 – Composante numero 17
0.2312
−0.2133

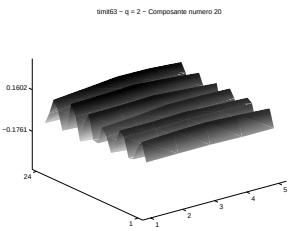

timit63 – q = 2 – Composante numero 18
0.1828
−0.1668

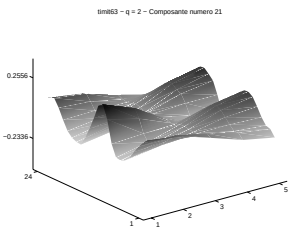

timit63 – q = 2 – Composante numero 19

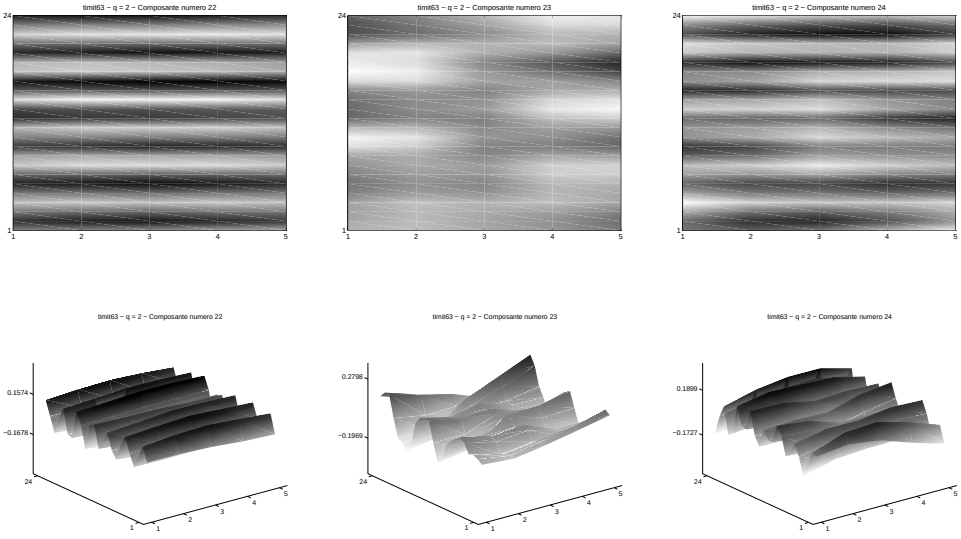

timit63 – q = 2 – Composante numero 20


timit63 – q = 2 – Composante numero 21


timit63 – q = 2 – Composante numero 19
0.1528
−0.1578


timit63 – q = 2 – Composante numero 20
0.1602
−0.1761


timit63 – q = 2 – Composante numero 21
0.2556
−0.2336

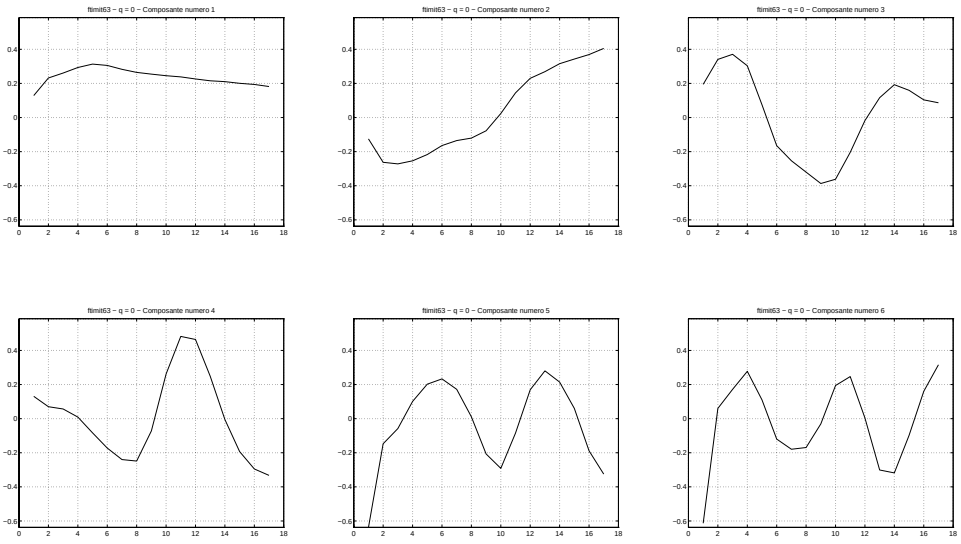

timit63 − q = 2 − Composante numero 22
timit63 − q = 2 − Composante numero 23
timit63 − q = 2 − Composante numero 24
timit63 − q = 2 − Composante numero 22
timit63 − q = 2 − Composante numero 23
timit63 − q = 2 − Composante numero 24

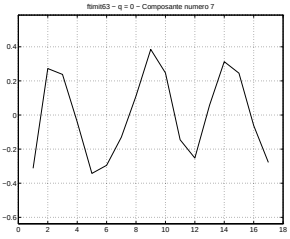
Annexe F

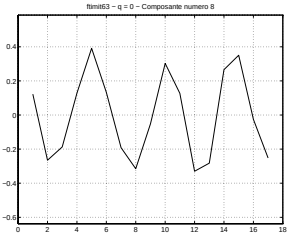
Composantes principales pour la base FTIMIT63

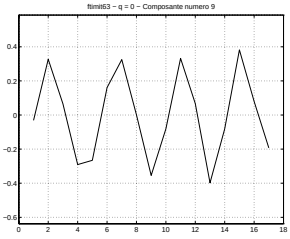
Nous donnons dans cette annexe la représentation graphique de nombreuses composantes principales pour la base FTIMIT63. Cela permet notamment de voir quelles sont celles qui ont été choisies dans le chapitre 7, et de comparer leurs formes avec celles qui ont été écartées. Les représentations sont données pour les valeurs de $q = 0$, $q = 1$ et $q = 2$. Enfin, nous ne représentons, pour chaque valeur de q, que les 17 premières composantes.

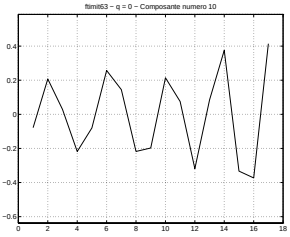
F.1 $q = 0$: 17 composantes principales

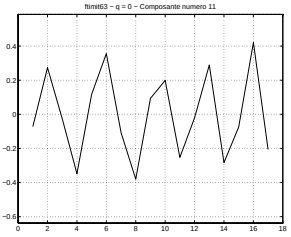


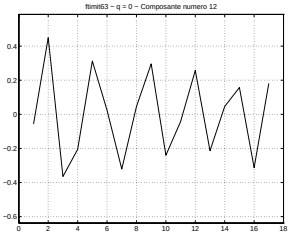

ftimit63 – q = 0 – Composante numero 7

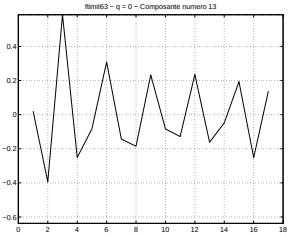

ftimit63 – q = 0 – Composante numero 8

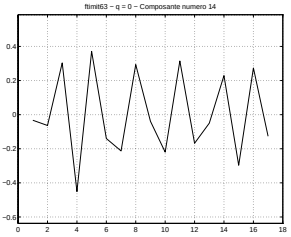

ftimit63 – q = 0 – Composante numero 9

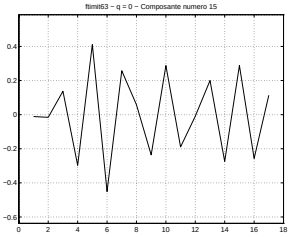

ftimit63 – q = 0 – Composante numero 10

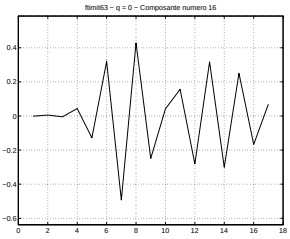

ftimit63 – q = 0 – Composante numero 11

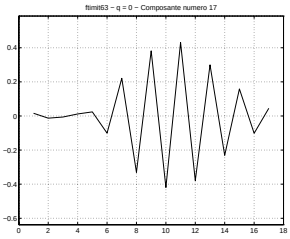

ftimit63 – q = 0 – Composante numero 12


ftimit63 – q = 0 – Composante numero 13

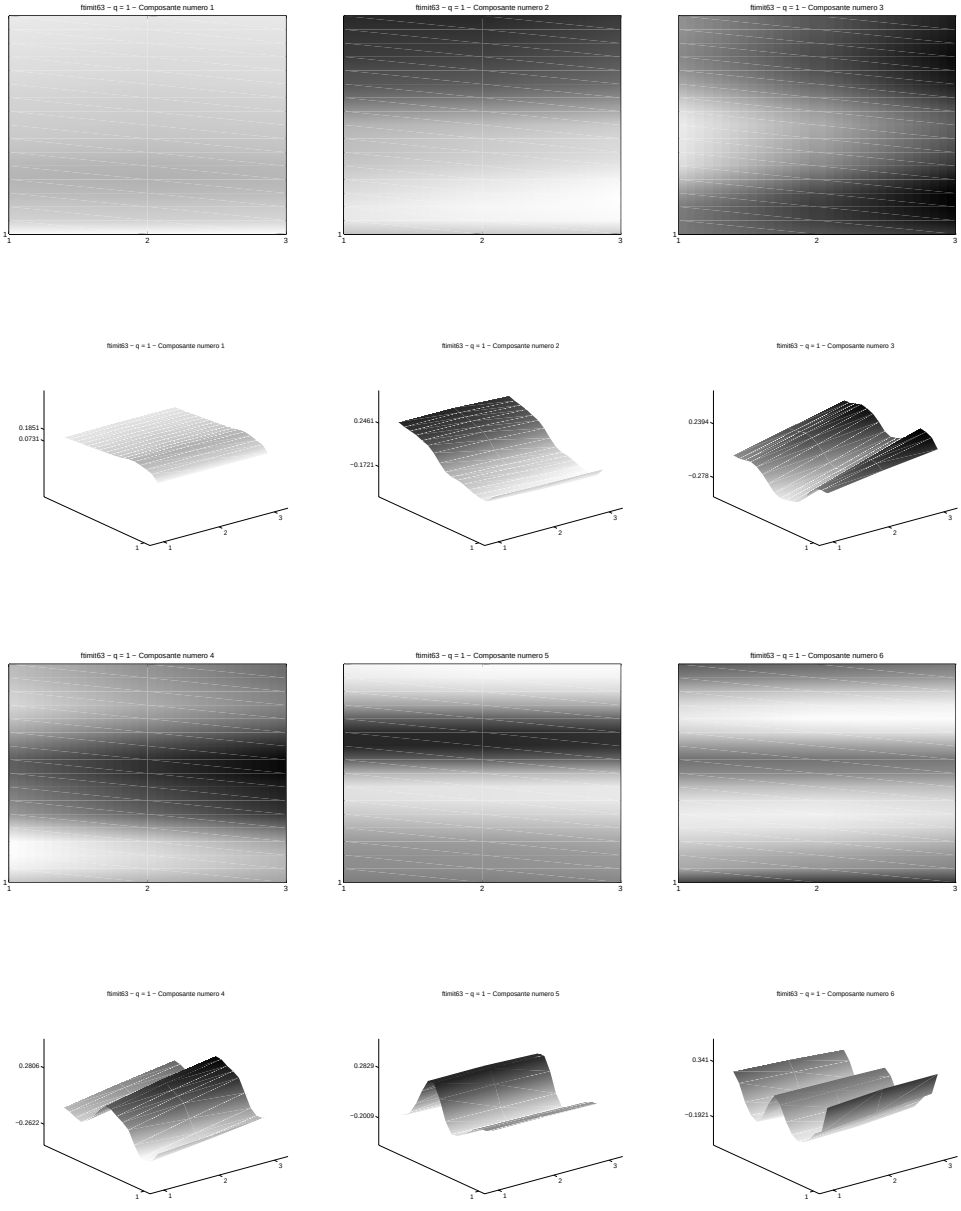

ftimit63 – q = 0 – Composante numero 14

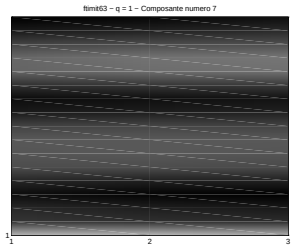

ftimit63 – q = 0 – Composante numero 15

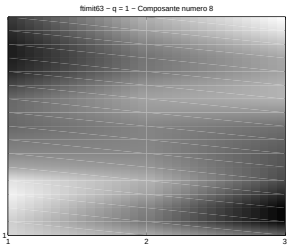

ftimit63 – q = 0 – Composante numero 16

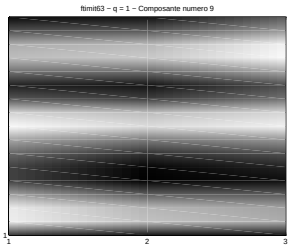

ftimit63 – q = 0 – Composante numero 17

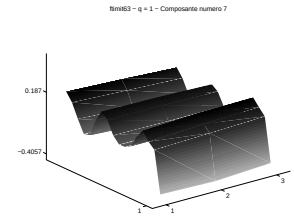
F.2 $q = 1$: 17 premières composantes principales

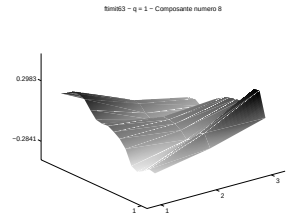


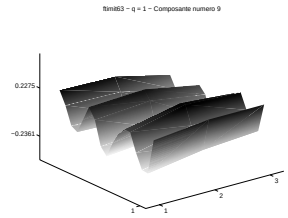

ftimit63 − q = 1 − Composante numero 7

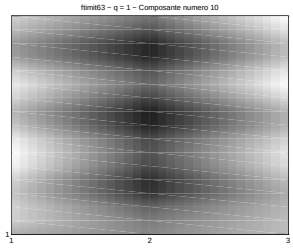

ftimit63 − q = 1 − Composante numero 8

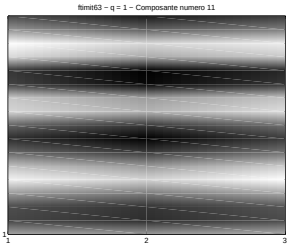

ftimit63 − q = 1 − Composante numero 9

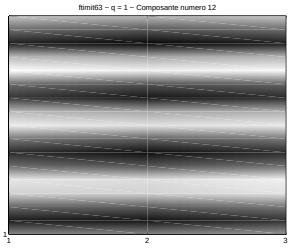

ftimit63 − q = 1 − Composante numero 7
0.187
−0.4057

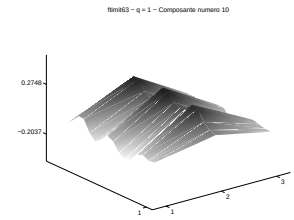

ftimit63 − q = 1 − Composante numero 8
0.2983
−0.2841

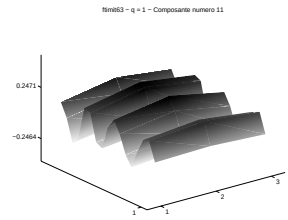

ftimit63 − q = 1 − Composante numero 9
0.2275
−0.2361

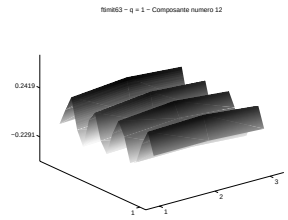

ftimit63 − q = 1 − Composante numero 10

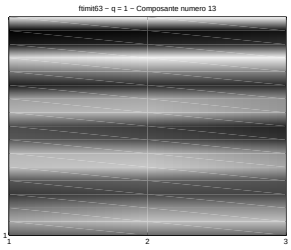

ftimit63 − q = 1 − Composante numero 11

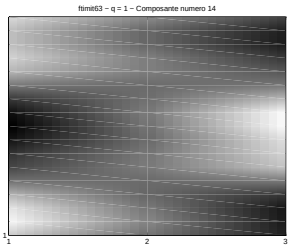

ftimit63 − q = 1 − Composante numero 12

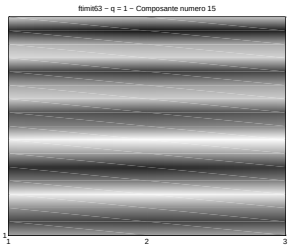

ftimit63 − q = 1 − Composante numero 10
0.2748
−0.2037

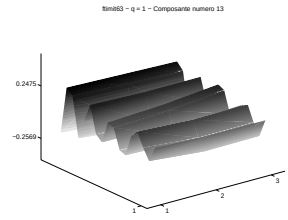

ftimit63 − q = 1 − Composante numero 11
0.2471
−0.2464

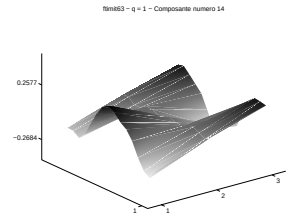

ftimit63 − q = 1 − Composante numero 12
0.2419
−0.2291

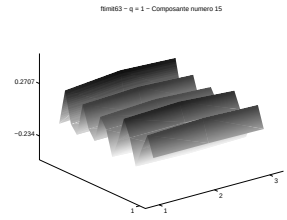

ftimit63 – q = 1 – Composante numero 13

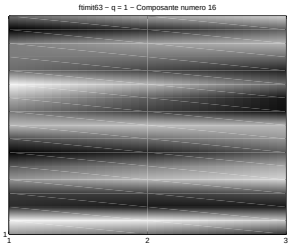

ftimit63 – q = 1 – Composante numero 14

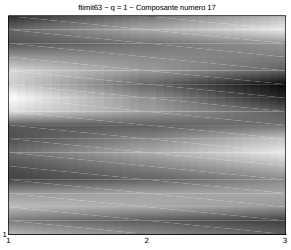

ftimit63 – q = 1 – Composante numero 15

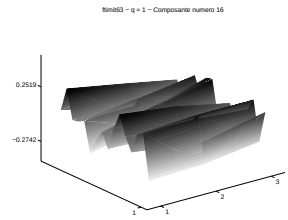

ftimit63 – q = 1 – Composante numero 13
0.2475
−0.2569

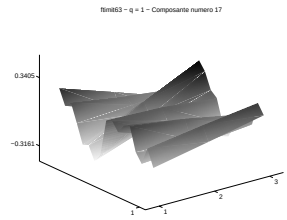

ftimit63 – q = 1 – Composante numero 14
0.2577
−0.2684


ftimit63 – q = 1 – Composante numero 15
0.2707
−0.234

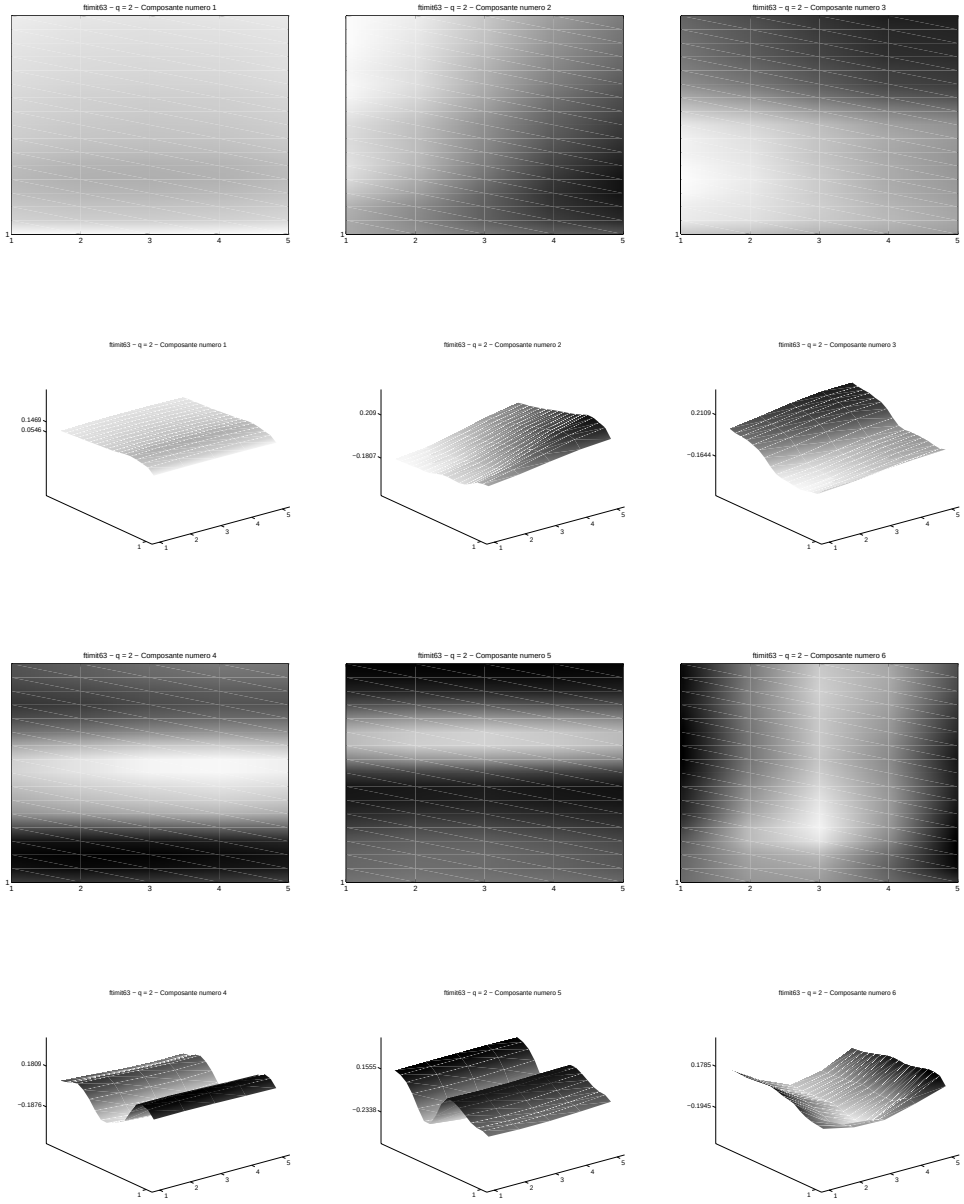

ftimit63 – q = 1 – Composante numero 16

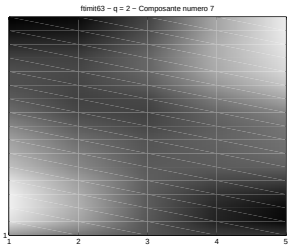

ftimit63 – q = 1 – Composante numero 17

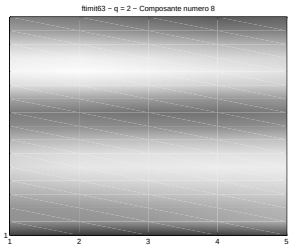

ftimit63 – q = 1 – Composante numero 16
0.2519
−0.2742

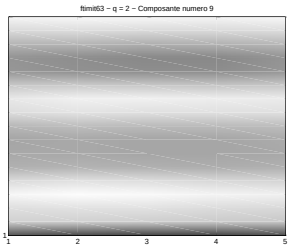

ftimit63 – q = 1 – Composante numero 17
0.3405
−0.3161

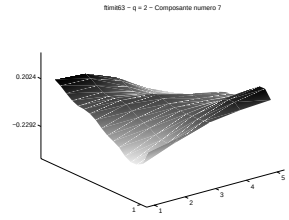
F.3 $q = 2$: 17 premières composantes principales

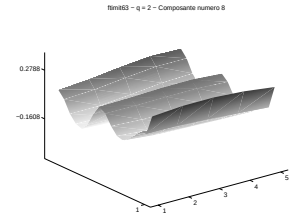


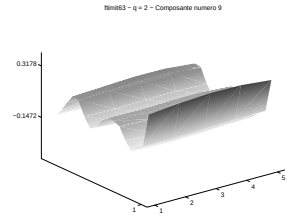

ftimit63 – q = 2 – Composante numero 7

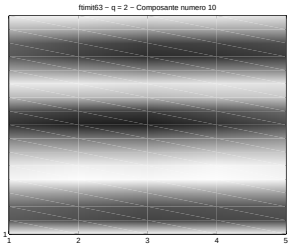

ftimit63 – q = 2 – Composante numero 8

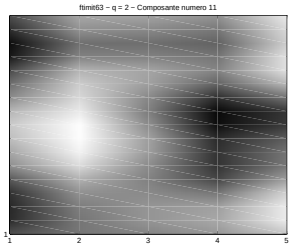

ftimit63 – q = 2 – Composante numero 9

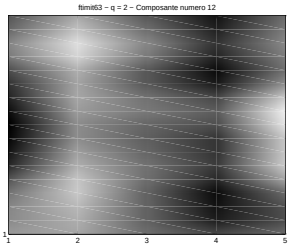

ftimit63 – q = 2 – Composante numero 7
0.2024
−0.2292

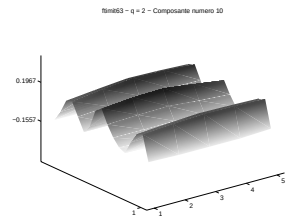

ftimit63 – q = 2 – Composante numero 8
0.2788
−0.1608

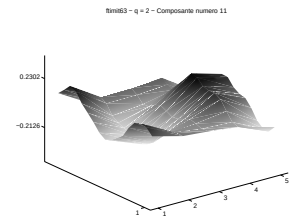

ftimit63 – q = 2 – Composante numero 9
0.3178
−0.1472

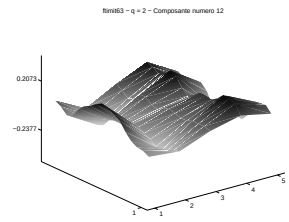

ftimit63 – q = 2 – Composante numero 10

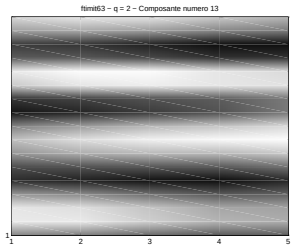

ftimit63 – q = 2 – Composante numero 11

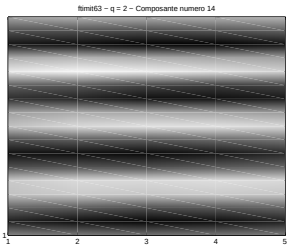

ftimit63 – q = 2 – Composante numero 12

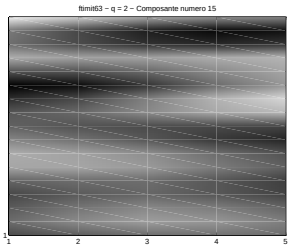

ftimit63 – q = 2 – Composante numero 10
0.1967
−0.1557

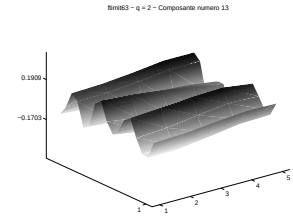

ftimit63 – q = 2 – Composante numero 11
0.2302
−0.2126

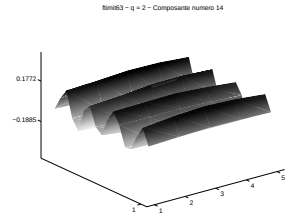

ftimit63 – q = 2 – Composante numero 12
0.2073
−0.2377

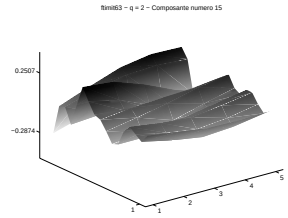

ftimit63 – q = 2 – Composante numero 13

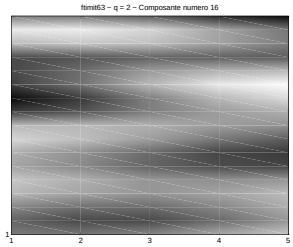

ftimit63 – q = 2 – Composante numero 14

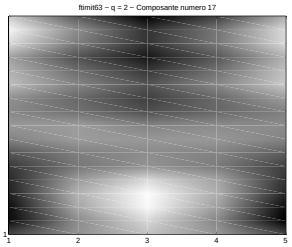

ftimit63 – q = 2 – Composante numero 15

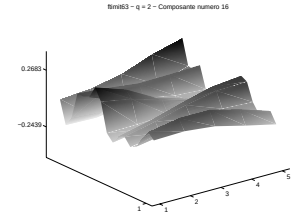

ftimit63 – q = 2 – Composante numero 13
0.1909
−0.1703

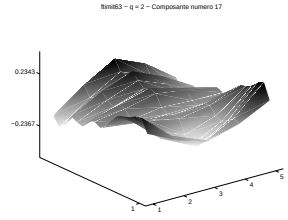

ftimit63 – q = 2 – Composante numero 14
0.1772
−0.1885


ftimit63 – q = 2 – Composante numero 15
0.2507
−0.2874


ftimit63 – q = 2 – Composante numero 16


ftimit63 – q = 2 – Composante numero 17


ftimit63 – q = 2 – Composante numero 16
0.2683
−0.2439

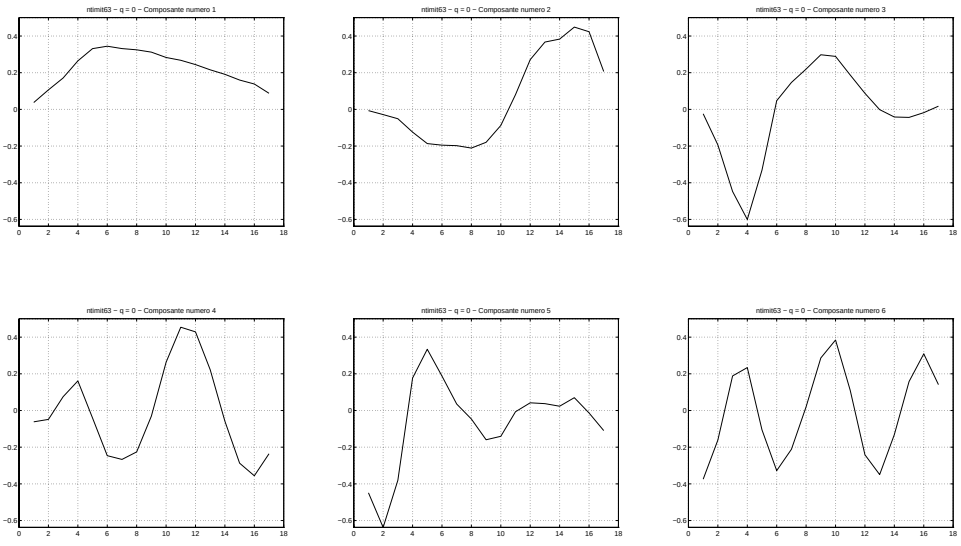

ftimit63 – q = 2 – Composante numero 17
0.2343
−0.2367

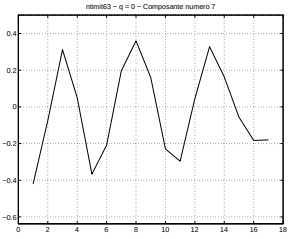
Annexe G

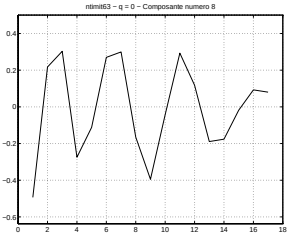
Composantes principales pour la base NTIMIT63

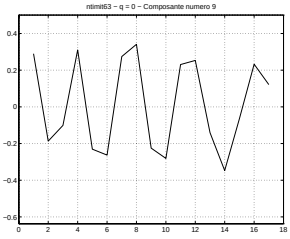
Nous donnons dans cette annexe la représentation graphique de nombreuses composantes principales pour la base NTIMIT63. Cela permet notamment de voir quelles sont celles qui ont été choisies dans le chapitre 7, et de comparer leurs formes avec celles qui ont été écartées. Les représentations sont données pour les valeurs de $q = 0$, $q = 1$ et $q = 2$. Enfin, nous ne représentons, pour chaque valeur de q, que les 17 premières composantes.

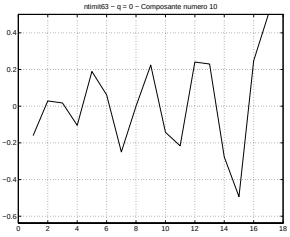
G.1 $q = 0$: 17 composantes principales

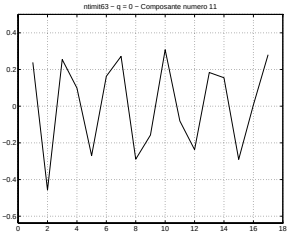


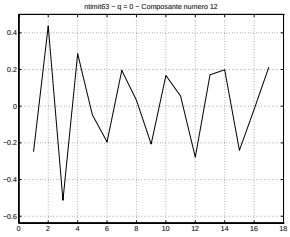

ntimit63 − q = 0 − Composante numero 7

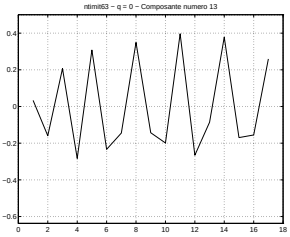

ntimit63 − q = 0 − Composante numero 8

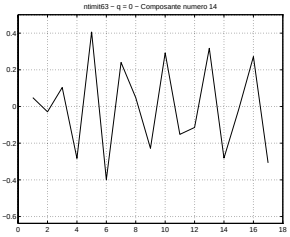

ntimit63 − q = 0 − Composante numero 9

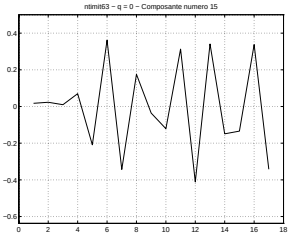

ntimit63 − q = 0 − Composante numero 10

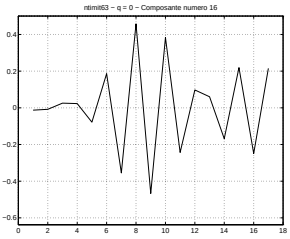

ntimit63 − q = 0 − Composante numero 11

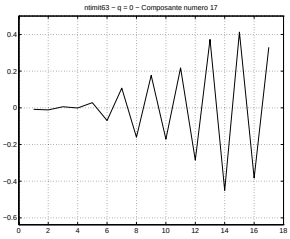

ntimit63 − q = 0 − Composante numero 12


ntimit63 − q = 0 − Composante numero 13

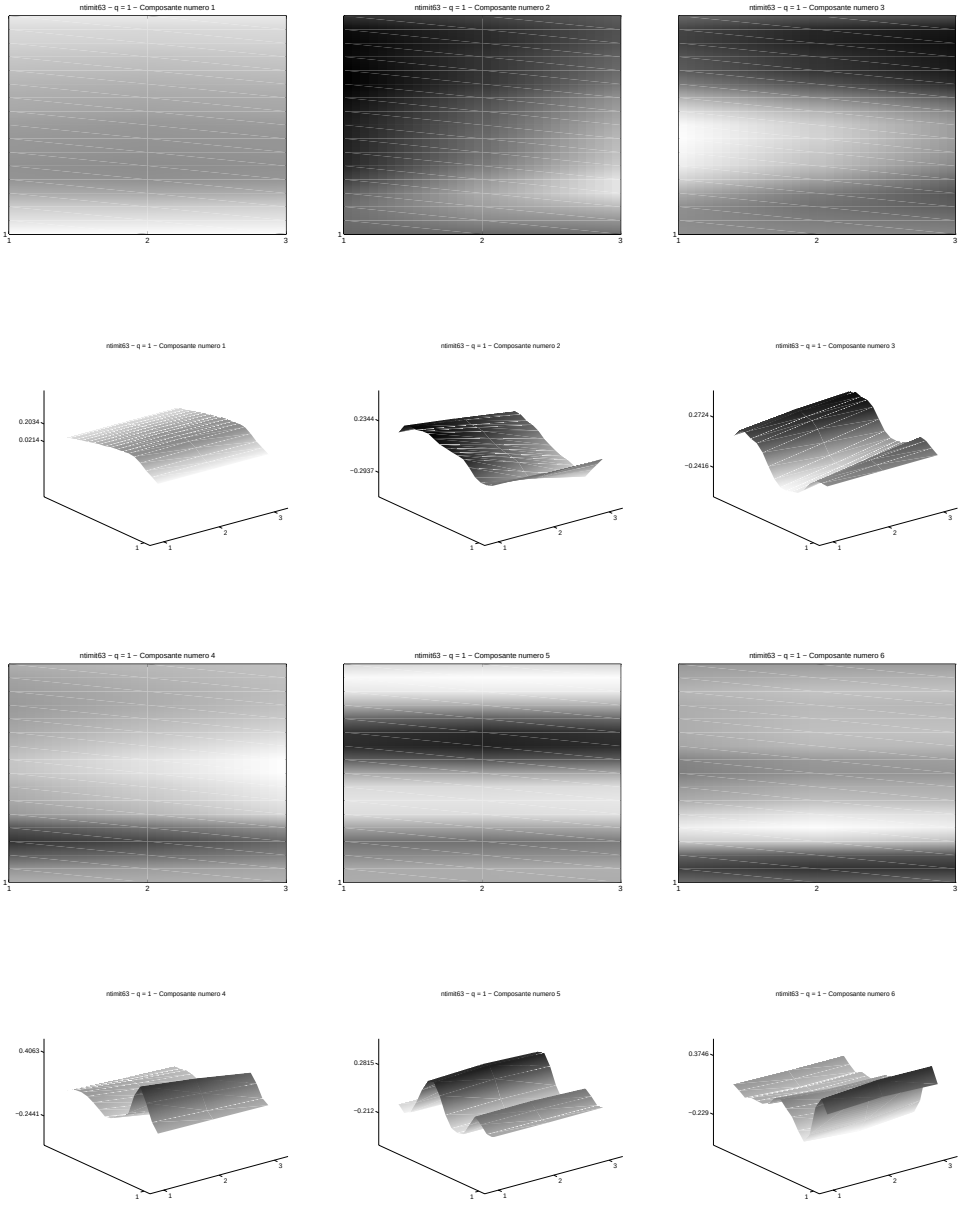

ntimit63 − q = 0 − Composante numero 14

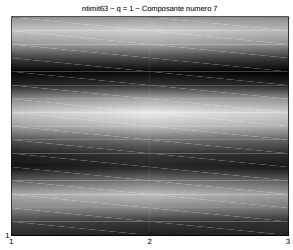

ntimit63 − q = 0 − Composante numero 15

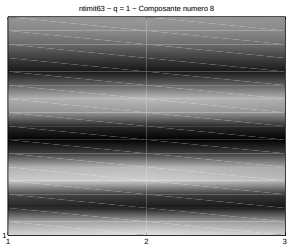

ntimit63 − q = 0 − Composante numero 16

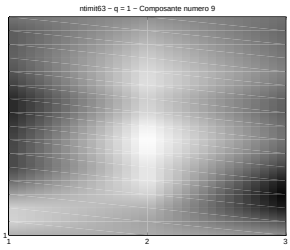

ntimit63 − q = 0 − Composante numero 17

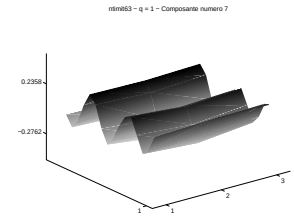
G.2 $q = 1$: 17 premières composantes principales

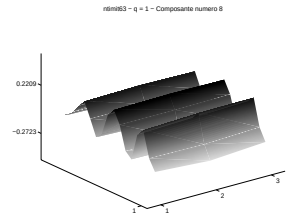


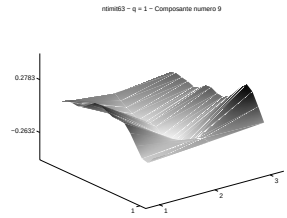

ntimit63 – q = 1 – Composante numero 7

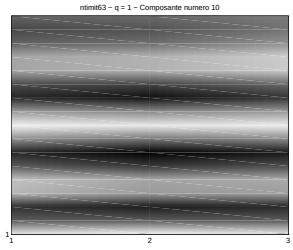

ntimit63 – q = 1 – Composante numero 8

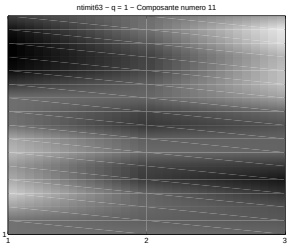

ntimit63 – q = 1 – Composante numero 9

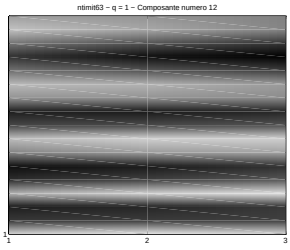

ntimit63 – q = 1 – Composante numero 7
0.2358
−0.2762

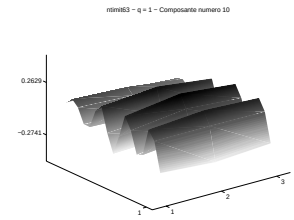

ntimit63 – q = 1 – Composante numero 8
0.2209
−0.2723

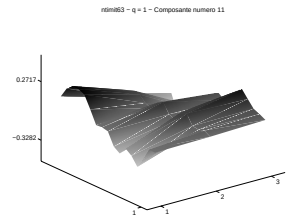

ntimit63 – q = 1 – Composante numero 9
0.2783
−0.2632

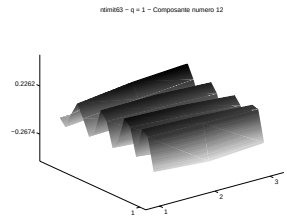

ntimit63 – q = 1 – Composante numero 10

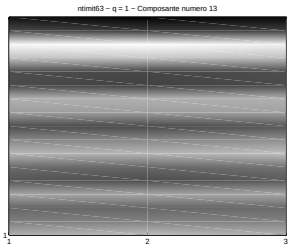

ntimit63 – q = 1 – Composante numero 11

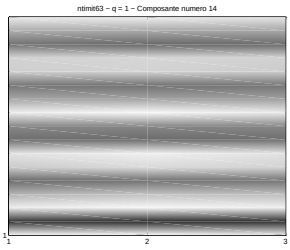

ntimit63 – q = 1 – Composante numero 12

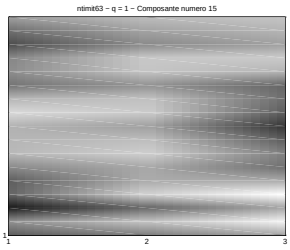

ntimit63 – q = 1 – Composante numero 10
0.2629
−0.2741

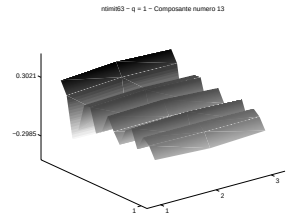

ntimit63 – q = 1 – Composante numero 11
0.2717
−0.3282

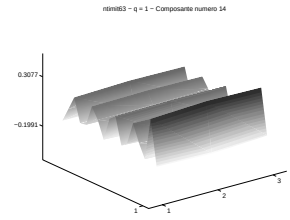

ntimit63 – q = 1 – Composante numero 12
0.2262
−0.2674

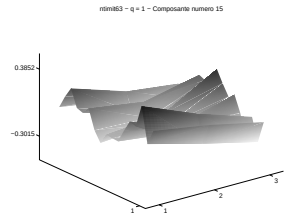

ntimit63 − q = 1 − Composante numero 13

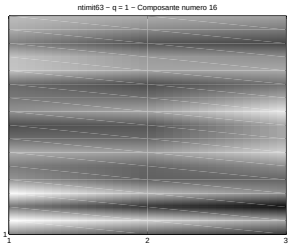

ntimit63 − q = 1 − Composante numero 14

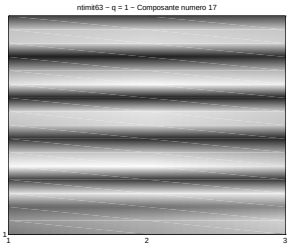

ntimit63 − q = 1 − Composante numero 15

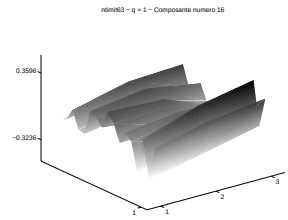

ntimit63 − q = 1 − Composante numero 13
0.3021
−0.2985

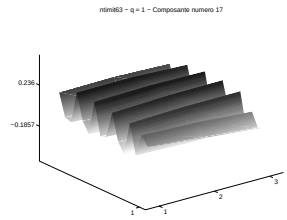

ntimit63 − q = 1 − Composante numero 14
0.3077
−0.1991


ntimit63 − q = 1 − Composante numero 15
0.3852
−0.3015

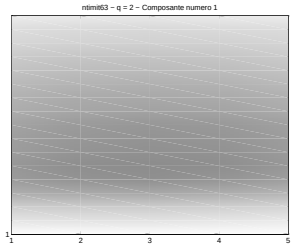

ntimit63 − q = 1 − Composante numero 16

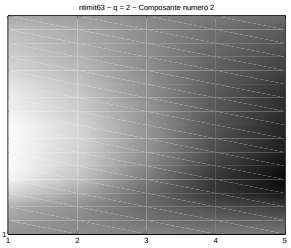

ntimit63 − q = 1 − Composante numero 17

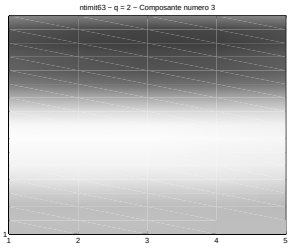

ntimit63 − q = 1 − Composante numero 16
0.3596
−0.3236

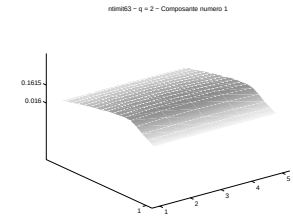

ntimit63 − q = 1 − Composante numero 17
0.236
−0.1857

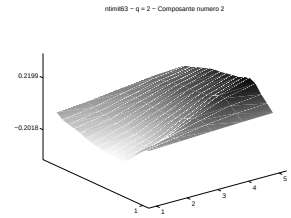
G.3 $q = 2$: 17 premières composantes principales

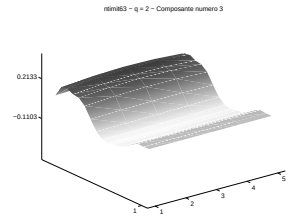


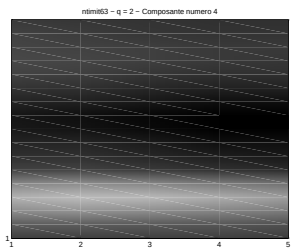


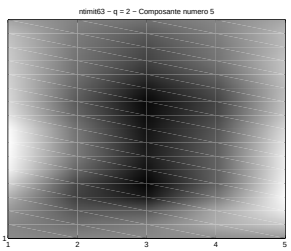


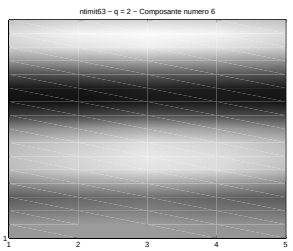


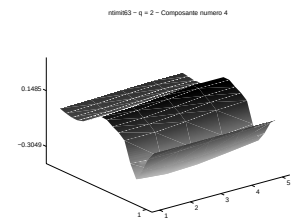


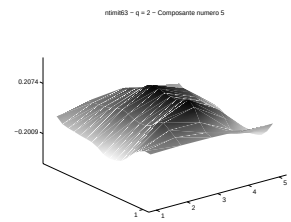


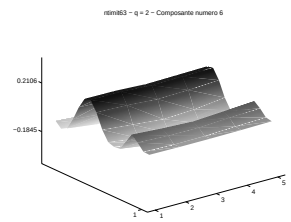


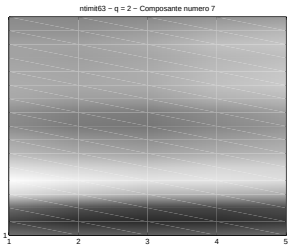


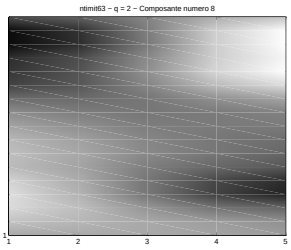


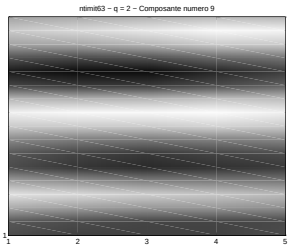


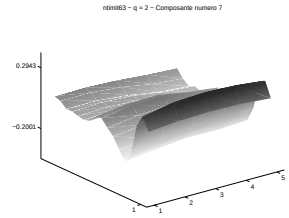


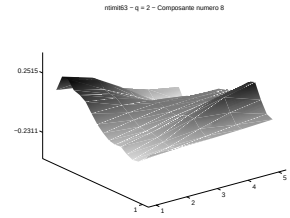


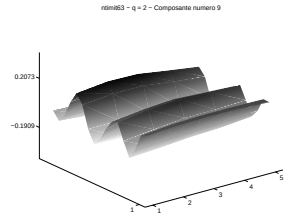

ntimit63 − q = 2 − Composante numero 7

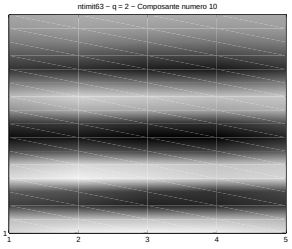

ntimit63 − q = 2 − Composante numero 8

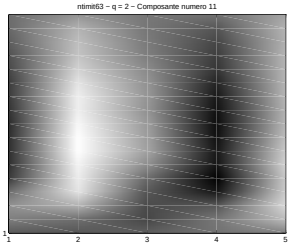

ntimit63 − q = 2 − Composante numero 9

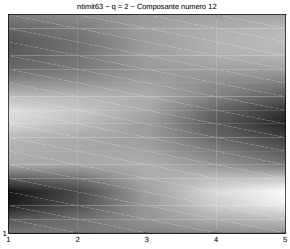

ntimit63 − q = 2 − Composante numero 7
0.2943
−0.2001

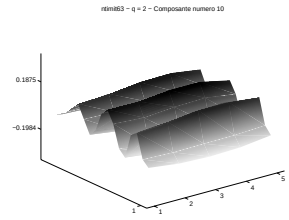

ntimit63 − q = 2 − Composante numero 8
0.2515
−0.2311

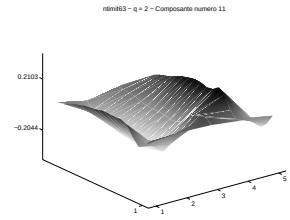

ntimit63 − q = 2 − Composante numero 9
0.2073
−0.1909

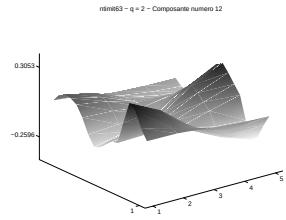

ntimit63 − q = 2 − Composante numero 10

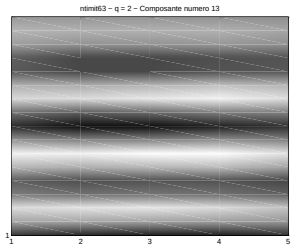

ntimit63 − q = 2 − Composante numero 11

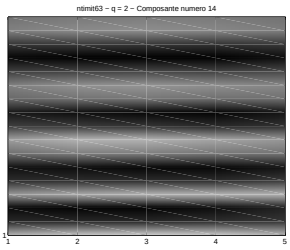

ntimit63 − q = 2 − Composante numero 12

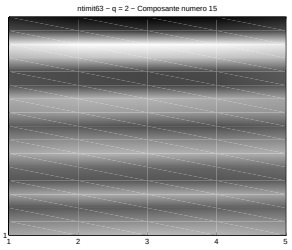

ntimit63 − q = 2 − Composante numero 10
0.1875
−0.1984

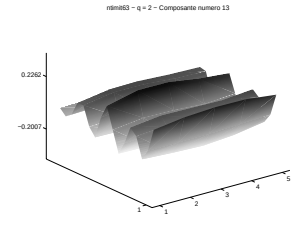

ntimit63 − q = 2 − Composante numero 11
0.2103
−0.2044

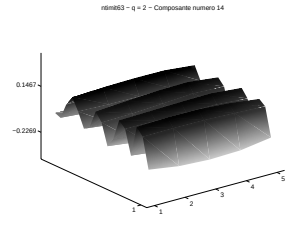

ntimit63 − q = 2 − Composante numero 12
0.3053
−0.2596

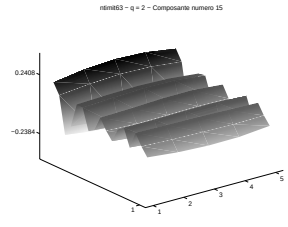

ntimit63 − q = 2 − Composante numero 13

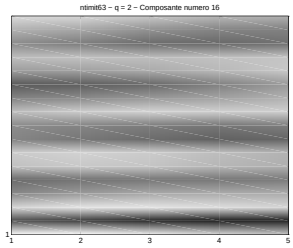

ntimit63 − q = 2 − Composante numero 14

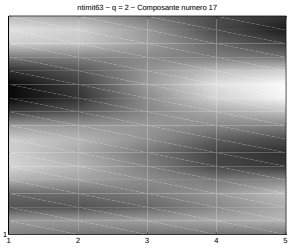

ntimit63 − q = 2 − Composante numero 15

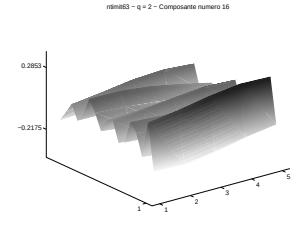

ntimit63 − q = 2 − Composante numero 13
0.2282
−0.2007

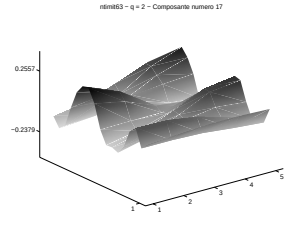

ntimit63 − q = 2 − Composante numero 14
0.1467
−0.2269


ntimit63 − q = 2 − Composante numero 15
0.2408
−0.2384


ntimit63 − q = 2 − Composante numero 16


ntimit63 − q = 2 − Composante numero 17


ntimit63 − q = 2 − Composante numero 16
0.2853
−0.2175


ntimit63 − q = 2 − Composante numero 17
0.2557
−0.2379

Annexe H

Publications

Second-Order Statistical Measures for Text-Independent Speaker Identification.
Frédéric Bimbot, Ivan Magrin-Chagnolleau and Luc Mathan.
Speech Communication, vol. 17, no. 1–2 : pp. 177–192, August 1995.

Effect of Utterance Duration and Phonetic Content on Text-Independent Speaker Identification using Second-Order Statistical Measures.
Ivan Magrin-Chagnolleau, Jean-François Bonastre and Frédéric Bimbot.
Proceedings of EUROSPEECH 95, vol. 1, pp. 337–340, Madrid, SPAIN, September 1995.

A Further Investigation on AR-Vector Models for Text-Independent Speaker Identification.
Ivan Magrin-Chagnolleau, Joachim Wilke and Frédéric Bimbot.
Proceedings of ICASSP 96, vol. 1, pp. 401–404, Atlanta, USA, May 1996.

Second-Order Statistical Measures for Text-Independent Speaker Identification

Frédéric BIMBOT, Ivan MAGRIN-CHAGNOLLEAU and Luc MATHAN

Ecole Nationale Supérieure des Télécommunications
E.N.S.T. / Télécom Paris – Département Signal
C.N.R.S. – URA 820
46, rue Barrault
75634 PARIS cedex 13
FRANCE, European Union

E-mail: bimbot@sig.enst.fr and ivan@sig.enst.fr

Speech Communication, Vol. 17, No. 1–2, pp. 177–192, August 1995.

Second-Order Statistical Measures for Text-Independent Speaker Identification

Abstract

This article presents an overview of several measures for speaker recognition. These measures relate to second-order statistical tests, and can be expressed under a common formalism. Alternate formulations of these measures are given and their mathematical properties are studied. In their basic form, these measures are asymmetric, but they can be symmetrized in various ways. All measures are tested in the framework of text-independent closed-set speaker identification, on 3 variants of the TIMIT database (630 speakers) : TIMIT (high quality speech), FTIMIT (a restricted bandwidth version of TIMIT) and NTIMIT (telephone quality). Remarkable performances are obtained on TIMIT but the results naturally deteriorate with FTIMIT and NTIMIT. Symmetrization appears to be a factor of improvement, especially when little speech material is available. The use of some of the proposed measures as a reference benchmark to evaluate the intrinsic complexity of a given database under a given protocol is finally suggested as a conclusion to this work.

Abstandsmaße basierend auf statistischen Methoden zweiter Ordnung zur textunabhängigen Sprecheridentifizierung

Zusammenfassung

Dieser Artikel beschreibt mehrere Abstandsmaße der Sprechererkennung. Diese Abstandsmaße beziehen sich auf Tests basierend auf statistischen Methoden zweiter Ordnung und können unter einem gemeinsamen Formalismus betrachtet werden. Alternative Formalismen werden vorgestellt und ihre mathematischen Eigenschaften untersucht. In ihrer ursprünglichen Form sind diese Abstandsmaße asymetrisch. Sie können jedoch auf vielfältige Weise in eine symetrische Form umgewandelt werden. Alle Abstandsmaße werden im Rahmen einer textunabhängigen Sprechererkennung einer geschlossenen Sprechermenge an drei Variationen der TIMIT-Sprachdatenbank (630 Sprecher) getestet : TIMIT (Sprache mit hoher Aufnahmequalität), FTIMIT (eine Version von TIMIT mit eingeschränkter Bandbreite) und NTIMIT (Telephonqualität). Beachtenswerte Ergebnisse wurden mit TIMIT erreicht, die sich mit FTIMIT und NTIMIT verschlechtern. Es stellt sich heraus, daß die Symetrisierung einen Verbesserungsfaktor darstellt, vor allem, wenn wenig Sprachmaterial vorhanden ist. Die Verwendung einiger der vorgeschlagenen Abstandsmaße als Referenzvergleich zur Evaluierung der Komplexität einer gegebenen Sprachdatenbank unter einem gegebenen Protokoll wird am Ende dieser Arbeit vorgeschlagen.

Mesures statistiques du second ordre pour l'identification du locuteur indépendante du texte

Résumé

Cet article présente un ensemble de mesures pour la reconnaissance du locuteur. Ces mesures reposent sur des tests statistiques du second ordre, et peuvent être exprimées sous un formalisme commun. Différentes expressions de ces mesures sont proposées et leurs propriétés mathématiques sont étudiées. Dans leur forme la plus simple, ces mesures ne sont pas symétriques, mais elles peuvent être symétrisées de différentes façons. Toutes les mesures sont testées dans le cadre de l'identification du locuteur indépendante du texte en ensemble fermé, sur 3 versions de la base de données TIMIT (630 locuteurs) : TIMIT (parole de très bonne qualité), FTIMIT (version filtrée de TIMIT) et NTIMIT (qualité téléphonique). Des performances remarquables sont obtenues sur TIMIT, mais les résultats se dégradent naturellement avec FTIMIT et NTIMIT. La symétrisation apparaît comme un facteur d'amélioration, plus particulièrement lorsque l'on dispose de peu de parole. Il est finalement suggéré, comme conclusion à ce travail, d'utiliser certaines mesures proposées comme méthodes de référence pour évaluer la complexité intrinsèque d'une base de données quelconque, sous un protocole donné.

1 Introduction

1.1 A brief overview

Recent experiments [2] [16] [5] [17] using vector Auto-Regressive models for speaker recognition confirm and further develop work carried out by Grenier [13]. The vector AR approach provides excellent results on a subset of the TIMIT database (420 speakers), in a text-independent mode : with a training of 5 sentences (approximately 15 seconds) and tests of 1 sentence (approximately 3 seconds), closed-set identification scores reported by Montacié [17] are of 98.4 %, and reach 100 % when using 5 sentences for testing. By incorporating a discriminant analysis, the 98.4 % score improves to 99.3 %.
On the same database, other approaches have been recently tested, in particular Neural Network based methods. For instance, Rudasi and Zahorian propose binary discriminative networks [20], and reach a 100 % identification score, with 47 speakers, 5 sentences for training and 5 others for testing. Bennani [3] reports experiments with a modular TDNN-based architecture which provides 100 % correct identification for more than 100 TIMIT speakers, using about 15 seconds for training and less than 1 second for testing.
An other method used by Hattori [14] is based on predictive networks. Under this approach, a neural network is trained, for each speaker, to predict a speech frame given the 2 previous ones. During recognition, the identified speaker is the one corresponding to the network with the lowest prediction error. With the best variant, Hattori obtains 100 % correct identifications on 24 speakers (from TIMIT), with about 15 seconds for training and 9 seconds for testing.
Still on TIMIT database, Reynolds [19] shows that a Gaussian Mixture speaker model (with 32 Gaussian distributions with diagonal covariance matrices) leads to a very high level of identification performance : 99.7 % for 168 speakers, using 8 sentences for training and 2 for testing. As discussed by Furui [9], the Gaussian Mixture approach shares strong similarities with the Vector Quantization based approaches [22] and with the Ergodic HMM based methods [18] [21]. It is therefore very likely that these approaches would also provide excellent results on TIMIT.

1.2 Motivation

In spite of the fact that all approaches mentioned in this brief overview were tested on the same database, it is still difficult to have a clear idea of their relative performances. Among the factors of variability between the experiments are the speech signal pre-processing, the type of acoustic analysis, the length of training and test utterances and of course the number of speakers for which the results are reported.
A systematic comparison of any new approach with all pre-existing methods, under the exact same protocol, is theoretically possible but practically unfeasible ; not only owing to the amount of work involved, but also because it may be very difficult to reproduce in detail a specific algorithm for which all needed information may not be publicly available, or which is sensitive to initialization conditions. Moreover, it can be argued that such or such database is easy and non-discriminant (which may very well be the case for TIMIT), but we lack reliable tools to evaluate the intrinsic difficulty of a database.

A possible way to address this problem of evaluation is the use of a common algorithm as a reference benchmark to evaluate the complexity of a given database under a given protocol [7]. Desirable properties for such a reference method are its relative efficiency and robustness, but also its easy implementation and its absolute reproductibility [6].
The work reported in this article is dedicated to similarity measures between speakers which are derived from statistical tests, with an underlying Gaussian speaker model. The theoretical formulation of these measures illustrates their straightforward reproductibility, while the experimental results evaluate their efficiency on several databases : namely TIMIT (high quality speech), FTIMIT (a 0-4 kHz version of TIMIT) and NTIMIT (telephone quality speech).
In parallel to this large scale evaluation, we discuss the possibility of using one or two of the proposed approaches as systematical benchmarks, in order to provide baseline performance for any database and protocol. Such reference scores would give an idea of the degree of complexity of a given task, and the improvement obtained by any other method would indicate the benefits of a more elaborate speaker model.

1.3 Outline

Three families of measures are investigated in this paper, namely :

- log-likelihood based measures
- sphericity test based measures
- relative eigenvalue deviation measures

In section 3, we present all measures under a common formalism (defined in section 2), and we study their mathematical properties. In their original forms, these measures are not symmetric, and we describe, in section 4, some possibilities to symmetrize them. Section 5 is dedicated to the description of our evaluation protocol, and to the corresponding results. In section 6, we discuss, the possibility of using some of the measures as reference methods.

2 Notation, definitions, properties

2.1 A Gaussian model per speaker

Let $\{x_t\}_{1 \leq t \leq M}$ be a sequence of M vectors resulting from the p-dimensional acoustic analysis of a speech signal uttered by a speaker $\mathcal{X}$. For instance : filter-bank coefficients, linear prediction coefficients, cepstrum coefficients,... Under the hypothesis of a Gaussian speaker model, the vector sequence $\{x_t\}$ can be summarized by its mean vector $\bar{x}$ and its covariance matrix X, i.e.

$$\bar{x} = \frac{1}{M} \sum_{t=1}^{M} x_t \quad \textit{and} \quad X = \frac{1}{M} \sum_{t=1}^{M} (x_t - \bar{x}) \cdot (x_t - \bar{x})^T \tag{1}$$

Similarly, for a speaker $\mathcal{Y}$, a parameterized speech utterance $\{y_t\}$ of N vectors can be modeled by $\bar{y}$ and Y, with

$$\bar{y} = \frac{1}{N} \sum_{t=1}^{N} y_t \quad \textit{and} \quad Y = \frac{1}{N} \sum_{t=1}^{N} (y_t - \bar{y}) \cdot (y_t - \bar{y})^T \tag{2}$$

Vectors $\bar{x}$ and $\bar{y}$ are p-dimensional, while X and Y are $p \times p$ symmetric matrices. Throughout this article, a speaker $\mathcal{X}$ (respectively $\mathcal{Y}$) will be represented by $\bar{x}$, X and M, (respectively $\bar{y}$, Y and N). We will also denote

$$\delta = \bar{y} - \bar{x}$$

$$\Gamma = X^{-\frac{1}{2}} \, Y \, X^{-\frac{1}{2}}$$

$$\rho = \frac{N}{M}$$

where $X^{\frac{1}{2}}$ is the symmetric square root matrix of X. Note that, when swapping $\mathcal{X}$ and $\mathcal{Y}$, vector δ becomes $-\delta$, matrix Γ becomes Γ^{-1} and real number ρ becomes $1/\rho$.

2.2 Second-order statistical measures

We focus on similarity measures μ between speakers $\mathcal{X}$ and $\mathcal{Y}$ which can be expressed as a function

$$\mu\,(\mathcal{X},\ \mathcal{Y}) = \phi\,(\bar{x},\ X,\ M,\ \bar{y},\ Y,\ N) \tag{3}$$

The measures μ that we investigate are derived from statistical hypothesis testing. They are constructed so that they are non-negative, i.e.

$$\forall\ \mathcal{X},\ \forall\ \mathcal{Y}, \qquad \mu\,(\mathcal{X},\ \mathcal{Y})\ \geq\ 0 \tag{4}$$

and they satisfy the property

$$\forall\ \mathcal{X}, \qquad \mu\,(\mathcal{X},\ \mathcal{X})\ =\ 0 \tag{5}$$

In their basic forms, the measures are non-symmetric, but we propose several ways to symmetrize them, so that

$$\forall\ \mathcal{X},\ \forall\ \mathcal{Y}, \qquad \mu\,(\mathcal{X},\ \mathcal{Y})\ =\ \mu\,(\mathcal{Y},\ \mathcal{X}) \tag{6}$$

2.3 Relative eigenvalues

We will denote as $\{\lambda_i\}_{1 \leq i \leq p}$ the eigenvalues of matrix Γ, i.e. the roots of the equation

$$det\,[\Gamma\ -\ \lambda I] = 0 \tag{7}$$

where det denotes the determinant, and I the unit matrix. Matrix Γ can be decomposed as

$$\Gamma\ =\ \Theta\,\Lambda\,\Theta^{-1} \tag{8}$$

where Λ is the $p \times p$ diagonal matrix of the eigenvalues, and Θ the $p \times p$ matrix of the eigenvectors. Classically, the eigenvalues λ_i are sorted in decreasing order when i increases.
Solutions of equation (7) are known as the eigenvalues of Y relative to X. Because X and Y are positive matrices, all eigenvalues λ_i are positive. Note also that the eigenvalues of X relative to Y (i.e. the eigenvalues of Γ^{-1}) are $\{1/\lambda_i\}_{1 \leq i \leq p}$.

2.4 Mean functions of the eigenvalues

Three particular functions of the eigenvalues λ_i are used in this article :

$$\text{The arithmetic mean :} \quad a\,(\lambda_1, ..., \lambda_p) = \frac{1}{p}\sum_{i=1}^{p}\lambda_i \tag{9}$$

$$\text{The geometric mean :} \quad g\,(\lambda_1, ..., \lambda_p) = \left(\prod_{i=1}^{p}\lambda_i\right)^{1/p} \tag{10}$$

$$\text{The harmonic mean :} \quad h\,(\lambda_1, ..., \lambda_p) = \left(\frac{1}{p}\sum_{i=1}^{p}\frac{1}{\lambda_i}\right)^{-1} \tag{11}$$

Because all eigenvalues λ_i are positive, it can be shown that

$$a \geq g \geq h \tag{12}$$

with equality if and only if all λ_i are equal. Moreover, swapping $\mathcal{X}$ and $\mathcal{Y}$ turns a into $1/h$, g into $1/g$ and h into $1/a$. In other words,

$$a\,(\frac{1}{\lambda_1}, ..., \frac{1}{\lambda_p}) = \frac{1}{h\,(\lambda_1, ..., \lambda_p)} \tag{13}$$

$$g\,(\frac{1}{\lambda_1}, ..., \frac{1}{\lambda_p}) = \frac{1}{g\,(\lambda_1, ..., \lambda_p)} \tag{14}$$

$$h\,(\frac{1}{\lambda_1}, ..., \frac{1}{\lambda_p}) = \frac{1}{a\,(\lambda_1, ..., \lambda_p)} \tag{15}$$

2.5 Computation of a, g and h

Given that the trace (denoted tr) satisfies $tr\,(AB) = tr\,(BA)$ and that the determinant (denoted det) verifies $det\,(AB) = det\,A \cdot det\,B$, we have the following properties :

$$a\,(\lambda_1, ..., \lambda_p) = \frac{1}{p}\,tr\,\Lambda = \frac{1}{p}\,tr\,\Gamma = \frac{1}{p}\,tr\,(YX^{-1}) \tag{16}$$

$$g\,(\lambda_1, ..., \lambda_p) = (det\,\Lambda)^{1/p} = (det\,\Gamma)^{1/p} = \left(\frac{det\,Y}{det\,X}\right)^{1/p} \tag{17}$$

$$h\,(\lambda_1, ..., \lambda_p) = \frac{p}{tr\,(\Lambda^{-1})} = \frac{p}{tr\,(\Gamma^{-1})} = \frac{p}{tr\,(XY^{-1})} \tag{18}$$

These equations show that functions a, g and h can be computed directly from X, Y, X^{-1}, Y^{-1}, $det\,X$ and $det\,Y$, without extracting explicitely the eigenvalues λ_i, nor calculating the matrix square roots of X and Y. Moreover, $tr\,(YX^{-1})$ and $tr\,(XY^{-1})$ can be computed without calculating the full matrix product, but only the diagonal elements of the product.

3 Second-order statistical measures

3.1 Gaussian likelihood measure

3.1.1 Definition

By supposing that all acoustic vectors extracted from the speech signal uttered by speaker $\mathcal{X}$ are distributed like a Gaussian function, the likelihood of a single acoustic vector y_t uttered by speaker $\mathcal{Y}$ is classically

$$G\,(y_t \mid \mathcal{X}) \;=\; \frac{1}{(2\pi)^{\frac{p}{2}}\;(det\;X)^{\frac{1}{2}}}\;e^{-\frac{1}{2}(y_t-\bar{x})^T X^{-1}(y_t-\bar{x})} \tag{19}$$

If we assume that all vectors y_t are independent observations, the average log-likelihood of $\{y_t\}_{1\leq t\leq N}$ can be written

$$\begin{aligned}\overline{G_{\mathcal{X}}}\,(y_1^N) \;&=\; \frac{1}{N}\,log\;G\,(y_1\;...\;y_n \mid \mathcal{X}) \;=\; \frac{1}{N}\;\sum_{t=1}^{N}\;log\;G\,(y_t \mid \mathcal{X}) \qquad (20)\\ &=\; -\frac{1}{2}\;\left[\;p\,log\;2\pi\;+\;log\;(det\;X)\;+\;\frac{1}{N}\;\sum_{t=1}^{N}\;(y_t-\bar{x})^T X^{-1}(y_t-\bar{x})\;\right]\end{aligned}$$

By replacing $y_t-\bar{x}$ by $y_t-\bar{y}+\bar{y}-\bar{x}$ and using the property

$$\frac{1}{N}\;\sum_{t=1}^{N}\;(y_t-\bar{y})^T X^{-1}(y_t-\bar{y}) \;=\; tr\;(Y X^{-1}) \tag{21}$$

we get

$$\overline{G_{\mathcal{X}}}\,(y_1^N)\;+\;\frac{p}{2}\,log\;2\pi \;=\; -\frac{1}{2}\;\left[\;log\;(det\;X)\;+\;tr\;(Y X^{-1})\;+\;(\bar{y}-\bar{x})^T X^{-1}(\bar{y}-\bar{x})\;\right] \tag{22}$$

and

$$\begin{aligned}&\frac{2}{p}\;\overline{G_{\mathcal{X}}}\,(y_1^N)\;+\;log\;2\pi\;+\frac{1}{p}\;log\;(det\;Y)\;+\;1 \qquad (23)\\ &\quad=\;\frac{1}{p}\;\left[log\left(\frac{det\,Y}{det\,X}\right)\;-\;tr\,(Y X^{-1})\;-\;(\bar{y}-\bar{x})^T X^{-1}(\bar{y}-\bar{x})\right]\;+\;1\end{aligned}$$

Therefore, if we define the Gaussian likelihood measure μ_G as

$$\begin{aligned}\mu_G\,(\mathcal{X},\,\mathcal{Y}) \;&=\; \frac{1}{p}\left[\,tr\,(Y X^{-1})\;-\;log\;\left(\frac{det\,Y}{det\,X}\right)\;+\;(\bar{y}-\bar{x})^T X^{-1}(\bar{y}-\bar{x})\right]\;-\;1 \qquad (24)\\ &=\; \frac{1}{p}\left[\;tr\;\Gamma\;-\;log\;(det\;\Gamma)\;+\;\delta^T X^{-1}\delta\;\right]\;-\;1 \qquad (25)\\ &=\; a\;-\;log\;g\;+\;\frac{1}{p}\;\delta^T X^{-1}\delta\;-\;1 \qquad (26)\end{aligned}$$

we have

$$\underset{\mathcal{X}}{Argmax}\;\;\overline{G_{\mathcal{X}}}\,(y_1^N) \;=\; \underset{\mathcal{X}}{Argmin}\;\;\mu_G\,(\mathcal{X},\,\mathcal{Y}) \tag{27}$$

3.1.2 Properties of μ_G

Matrix X^{-1} being, like X, positive definite, $\delta^T X^{-1}\delta \geq 0$. Moreover, we have $log\, g \leq g - 1$ and $a \geq g$. Therefore, $a - log\, g - 1 \geq 0$ and $\mu_G(\mathcal{X}, \mathcal{Y}) \geq 0$. Measure $\mu_G(\mathcal{X}, \mathcal{Y}) = 0$ if and only if all eigenvalues λ_i are equal to 1 and δ is the null vector, i.e. if and only if $X = Y$ and $\bar{x} = \bar{y}$. However, $\mu_G(\mathcal{X}, \mathcal{Y})$ is non-symmetric, its dual term being

$$\mu_G(\mathcal{Y}, \mathcal{X}) = \frac{1}{h} + log\, g + \frac{1}{p}\,\delta^T Y^{-1}\delta - 1 \neq \mu_G(\mathcal{X}, \mathcal{Y}) \tag{28}$$

3.1.3 A variant of μ_G

When dealing with noisy or distorted speech, the mean vectors $\bar{x}$ and $\bar{y}$ may be strongly influenced by the channel characteristics, while covariance matrices X and Y are usually more robust to variations between recording conditions and transmission lines [11]. Thus, the difference $\delta = \bar{y} - \bar{x}$ may be a misleading term in μ_G.
A Gaussian likelihood measure on the covariance matrices only, denoted here μ_{Gc}, can therefore be derived from the previous likelihood measure as

$$\mu_{Gc}(\mathcal{X}, \mathcal{Y}) = \frac{1}{p}\left[tr\,(YX^{-1}) - log\left(\frac{det\, Y}{det\, X}\right)\right] - 1 \tag{29}$$

$$= \frac{1}{p}\,[\, tr\,\Gamma - log\,(det\,\Gamma)\,] - 1 \tag{30}$$

$$= a - log\, g - 1 \tag{31}$$

This measure can be expressed as a function of the eigenvalues λ_i of matrix Γ. However, it does not require an explicit extraction of the eigenvalues. It has the same properties as measure μ_G. In particular, it is still non-symmetric, since

$$\mu_{Gc}(\mathcal{Y}, \mathcal{X}) = \frac{1}{h} + log\, g - 1 \neq \mu_{Gc}(\mathcal{X}, \mathcal{Y}) \tag{32}$$

3.2 Arithmetic-geometric sphericity measure

3.2.1 Definition

As presented by Anderson [1], a likelihood function for testing the proportionality of a covariance matrix Y to a given covariance matrix X is

$$S(Y \mid X) = \left[\frac{det\,(X^{-\frac{1}{2}}YX^{-\frac{1}{2}})}{\left(\frac{1}{p}\,tr\,(X^{-\frac{1}{2}}YX^{-\frac{1}{2}})\right)^p}\right]^{\frac{N}{2}} = \left[\frac{det\,\Gamma}{\left(\frac{1}{p}\,tr\,\Gamma\right)^p}\right]^{\frac{N}{2}} \tag{33}$$

This expression results from the combination of two criteria : one on the diagonality of matrix Γ, and a second one on the equality of the diagonal elements of Γ, given that Γ is diagonal.

By denoting as $\overline{S_{\mathcal{X}}}\,(y_1^N)$ the average likelihood function for the sphericity test,

$$\overline{S_{\mathcal{X}}}\,(y_1^N) \;=\; \frac{1}{N}\,log\;S\,(Y \mid X) \tag{34}$$

and by defining

$$\mu_{Sc}\,(\mathcal{X},\,\mathcal{Y}) \;=\; log\;\left[\frac{\frac{1}{p}\;tr\;\Gamma}{(det\;\Gamma)^{1/p}}\right] \tag{35}$$

$$\;=\; log\;\left[\frac{\frac{1}{p}\;tr\;(Y\,X^{-1})}{\left(\frac{det\;Y}{det\;X}\right)^{1/p}}\right] \tag{36}$$

$$\;=\; log\;\left(\frac{a}{g}\right) \tag{37}$$

we have

$$\underset{\mathcal{X}}{Argmax}\;\;\overline{S_{\mathcal{X}}}\,(y_1^N) \;=\; \underset{\mathcal{X}}{Argmin}\;\;\mu_{Sc}\,(\mathcal{X},\,\mathcal{Y}) \tag{38}$$

Measure μ_{Sc} appears as the logarithm of the ratio of the arithmetic and the geometric means of the eigenvalues of Y relative to X. As for measure μ_{Gc}, μ_{Sc} derives from a test on the covariance matrices only. It can be expressed as a function of the eigenvalues λ_i, but it does not require the search for the eigenvalues. The use of the arithmetic-geometric sphericity test for speaker recognition was initially proposed by Grenier [12], in the framework of text-dependent experiments.

3.2.2 Properties of μ_{Sc}

Since $a \;\geq\; g$, it is obvious that $\mu_{Sc}\,(\mathcal{X},\,\mathcal{Y}) \;\geq\; 0$. Measure $\mu_{Sc}\,(\mathcal{X},\,\mathcal{Y}) \;=\; 0$ if and only if all eigenvalues λ_i are equal, i.e. if and only if X and Y are proportional. In particular, $\mu_{Sc}\,(\mathcal{X},\,\mathcal{X}) \;=\; 0$, but $X \;=\; Y$ is not a necessary condition. Finally, μ_{Sc} is not symmetric, and

$$\mu_{Sc}\,(\mathcal{Y},\,\mathcal{X}) \;=\; log\;\left(\frac{g}{h}\right) \;\neq\; \mu_{Sc}\,(\mathcal{X},\,\mathcal{Y}) \tag{39}$$

3.3 Absolute deviation measure

3.3.1 Definition

The expression of μ_{Gc} and μ_{Sc} as functions of the eigenvalues λ_i are :

$$\mu_{Gc}\,(\mathcal{X},\,\mathcal{Y}) \;=\; \frac{1}{p}\;\sum_{i=1}^{p}\,(\lambda_i \;-\; log\;\lambda_i \;-\; 1) \tag{40}$$

$$\mu_{Sc}\;(\mathcal{X},\,\mathcal{Y}) \;=\; log\;\left(\;\frac{1}{p}\;\sum_{i=1}^{p}\;\lambda_i\;\right) \;-\; \frac{1}{p}\;\sum_{i=1}^{p}\;log\;\lambda_i \tag{41}$$

As a matter of fact, it is possible to construct other metrics to measure the dissimilarity between speakers, through their covariance matrices. Any function of the eigenvalues λ_i, which is non-negative, and which takes the zero value when all eigenvalues are equal to unity, is a possible choice.
This approach was proposed by Gish [10], who constructed a measure which is based on the total absolute deviation of the eigenvalues from unity. The generic expression of this measure, which we will denote as μ_{Dc}, is

$$\mu_{Dc}\,(\mathcal{X},\,\mathcal{Y}) \;=\; \frac{1}{p}\,\sum_{i=1}^{p}\;|\,\lambda_i \;-\; 1\,| \tag{42}$$

In this formulation, measure μ_{Dc} is the average absolute deviation of the eigenvalues λ_i from unity. Gish showed that robustness can be gained by removing large eigenvalues from the summation, because they may correspond to "abnormalities in small dimensional subspaces".

3.3.2 Properties of μ_{Dc}

It can be easily checked that measure μ_{Dc} is non-negative, and that it is null if and only if covariance matrices X and Y are equal. The measure is non-symmetric, since

$$\mu_{Dc}\,(\mathcal{Y},\,\mathcal{X}) \;=\; \frac{1}{p}\,\sum_{i=1}^{p}\;|\,\frac{1}{\lambda_i} \;-\; 1\,| \neq\; \mu_{Dc}\,(\mathcal{X},\,\mathcal{Y}) \tag{43}$$

4 Symmetrization

4.1 Motivation

All measures reviewed in the previous section have the common property of being non-symmetric. In other words, the roles played by the training data and by the test data are not interchangeable. However, our intuition would be that a similarity measure should be symmetric.
The asymmetry of measures μ_G, μ_{Gc} and μ_{Sc} can be explained by the following fact. These measures are based on statistical tests which suppose that the reference speaker model $\mathcal{X}$ is exact, while the test model $\mathcal{Y}$ is an estimation. But in practice, both reference and test models are estimates. Therefore, it is natural to search for a symmetric expression of originally asymmetric tests.
Moreover, it can be foreseen that the reliability of a reference model is dependent on the number of data that was used to estimate its parameters. This is experimentally confirmed by the discrepancies that can be observed in speaker identification performances, between $\mu\,(\mathcal{X},\,\mathcal{Y})$ and $\mu\,(\mathcal{Y},\,\mathcal{X})$, all the more as M and N, the number of reference and test vectors, are disproportionate (i.e. when $\rho = N/M$ is very different from 1).

4.2 Symmetrization procedures

A first possibility for symmetrizing a measure $\mu\ (\mathcal{X},\ \mathcal{Y})$, is to construct the average between the measure and its dual term :

$$\mu_{[0.5]}\ (\mathcal{X},\ \mathcal{Y})\ =\ \frac{1}{2}\,\mu\ (\mathcal{X},\ \mathcal{Y})\ +\ \frac{1}{2}\,\mu\ (\mathcal{Y},\ \mathcal{X})\ =\ \mu_{[0.5]}\ (\mathcal{Y},\ \mathcal{X}) \tag{44}$$

For instance, the Gaussian likelihood measure, symmetrized in this manner, becomes

$$\mu_{G_{[0.5]}}\ (\mathcal{X},\ \mathcal{Y})\ =\ \frac{1}{2}\ \left[\ a\ +\ \frac{1}{h}\ +\ \frac{1}{p}\,\delta^T\,(X^{-1}+Y^{-1})\,\delta\ \right]\ -\ 1 \tag{45}$$

which, for the covariance only measure, simplifies into

$$\mu_{Gc_{[0.5]}}\ (\mathcal{X},\ \mathcal{Y})\ =\ \frac{1}{2}\ \left(\ a\ +\ \frac{1}{h}\right)\ -\ 1 \tag{46}$$

while the arithmetic-geometric sphericity measure becomes proportional to the arithmetic-harmonic sphericity measure [4] :

$$\mu_{Sc_{[0.5]}}\ (\mathcal{X},\ \mathcal{Y})\ =\ \frac{1}{2}\,log\ \left(\ \frac{a}{h}\right) \tag{47}$$

This procedure of symmetrization can improve the classification performance, compared to both asymmetric terms taken individually. This is the case when training and test patterns have comparable length. However, we observed an inefficiency, or a degradation of the performance when the lengths differed significantly ($\rho\ \not\approx\ 1$). When training and test patterns are obtained from speech utterances with very different lengths, it turns out that $\mu\ (\mathcal{X},\ \mathcal{Y})$ performs better than $\mu\ (\mathcal{Y},\ \mathcal{X})$ when $\rho\ \leq\ 1$, and conversely. In other words, when the amount of training data is significantly lower than the amount of test data, it is preferable to model the test data and compute the average likelihood of the training data for the test model, rather than doing the opposite.
In the lack of a rigorous theoretical framework, we have limited our investigations to empirical trials. We have postulated an arbitrary form for more general symmetric measures μ, i.e. linear combinations of the asymmetric terms, weighted by coefficients that are function of the number of training and test vectors (respectively M and N) :

$$\mu_{[\psi_{MN}]}\ (\mathcal{X},\ \mathcal{Y})\ =\ \psi_{MN}\ \cdot\ \mu\ (\mathcal{X},\ \mathcal{Y})\ +\ \psi_{NM}\ \cdot\ \mu\ (\mathcal{Y},\ \mathcal{X}) \tag{48}$$

with

$$\psi_{MN}\ +\ \psi_{NM}\ =\ 1 \tag{49}$$

We have limited our tests to 2 particular functions ψ_{MN} and ψ_{NM}, namely :

$$\psi_{MN}\ =\ \ \ \alpha_{MN}\ \ \ =\ \frac{\sqrt{M}}{\sqrt{M}+\sqrt{N}}\ \ =\ \frac{1}{1+\sqrt{\rho}} \tag{50}$$

$$\psi_{NM}\ =\ 1-\alpha_{MN}\ =\ \frac{\sqrt{N}}{\sqrt{M}+\sqrt{N}}\ \ =\ \frac{\sqrt{\rho}}{1+\sqrt{\rho}} \tag{51}$$

and

$$\psi_{MN} \;=\; \beta_{MN} \;=\; \tfrac{M}{M+N} \;=\; \frac{1}{1+\rho} \tag{52}$$

$$\psi_{NM} \;=\; 1-\beta_{MN} \;=\; \tfrac{N}{M+N} \;=\; \frac{\rho}{1+\rho} \tag{53}$$

A similar approach was used by Montacié on AR-vector model residuals [17]. Note that, when $M \geq N$, $\rho \leq 1$ and therefore $0.5 \leq \alpha_{MN} \leq \beta_{MN}$.
We will not give the detailed expression of each measure, for each set of weights in this text. As an example, measure μ_G weighted by β_{MN} becomes

$$\mu_{G_{[\beta_{MN}]}}(\mathcal{X},\,\mathcal{Y}) \;=\; \frac{M\,\cdot\,\mu_G(\mathcal{X},\,\mathcal{Y}) \;+\; N\,\cdot\,\mu_G(\mathcal{Y},\,\mathcal{X})}{M+N} \tag{54}$$

$$\begin{aligned} &= \frac{1}{1+\rho}\,a \;-\; \frac{1-\rho}{1+\rho}\,log\;g \;+\; \frac{\rho}{1+\rho}\,\frac{1}{h} \\ &\qquad + \frac{1}{p}\,\delta^T\left(\frac{X^{-1}+\rho\,Y^{-1}}{1+\rho}\right)\delta \;-\; 1 \end{aligned} \tag{55}$$

$$\begin{aligned} &= \frac{1}{p}\left[\frac{1}{1+\rho}\,tr\,(YX^{-1}) \;-\; \frac{1-\rho}{1+\rho}\,log\,\left(\frac{detY}{detX}\right) \;+\; \frac{\rho}{1+\rho}\,tr\,(XY^{-1})\right] \\ &\qquad + \frac{1}{p}\left[(\bar{y}-\bar{x})^T\left(\frac{X^{-1}+\rho\,Y^{-1}}{1+\rho}\right)(\bar{y}-\bar{x})\right] \;-\; 1 \end{aligned} \tag{56}$$

The symmetry of this expression can easily be checked.
Even though they are empirical, the symmetrizations using α_{MN} and β_{MN} provide generally better results than the symmetrization with weights equal to $\frac{1}{2}$. The optimal expression for symmetrized measures can certainly be derived from estimation theory, but it is not a trivial problem.
An exception to the general approach was applied to measure μ_{Dc}, since we experienced that it was slightly more efficient to symmetrize $log\;\mu_{Dc}$ as above, instead of μ_{Dc} itself. However, $log\;\mu_{Dc}$ can not be considered as a measure in the mathematical sense, since it is not non-negative. Therefore,

$$log\left[\mu_{Dc_{[\psi_{MN}]}}(\mathcal{X},\,\mathcal{Y})\right] \;=\; \psi_{MN}\,\cdot\,log\left[\mu_{Dc}(\mathcal{X},\,\mathcal{Y})\right] \;+\; \psi_{NM}\,\cdot\,log\left[\mu_{Dc}(\mathcal{Y},\,\mathcal{X})\right] \tag{57}$$

which is equivalent to

$$\mu_{Dc_{[\psi_{MN}]}}(\mathcal{X},\,\mathcal{Y}) \;=\; \mu_{Dc}(\mathcal{X},\,\mathcal{Y})^{\,\psi_{MN}} \;\cdot\; \mu_{Dc}(\mathcal{Y},\,\mathcal{X})^{\,\psi_{NM}} \tag{58}$$

5 Experiments and results

5.1 Task

We have tested the measures described in this article, in the framework of closed-set text-independent speaker identification. There is a single reference per speaker (composed of a mean vector $\bar{x}$, a covariance matrix X and a number of data M). All test utterances are different from all training utterances, and all training utterances are different from one another. Each measure is evaluated as regards its classification ability using a 1-nearest neighbour decision rule. The possibility of rejection is not taken into account : the test speaker is always part of the set of references.

5.2 Databases

For our experiments, we used TIMIT and NTIMIT databases. TIMIT [8] contains 630 speakers (438 male and 192 female), each of them having uttered 10 sentences. Two sentences have the prefix "sa" (sa1 and sa2). Sentences sa1 and sa2 are different, but they are the same across speakers. Three sentences have the prefix "si" and five have the prefix "sx". These 8 sentences are different from one another, and different across speakers. Sentences "sa" and "si" have an average duration of 2.9 seconds. Sentences "sx" have an average duration of 3.2 seconds. The speech signal is recorded through a high quality microphone, in a very quiet environment, with a 0-8 kHz bandwidth. The signal is sampled at 16 kHz, on 16 bits, on a linear amplitude scale. Moreover, all recordings took place in a single session (contemporaneous speech).
The NTIMIT database [15] was obtained by playing TIMIT speech signal through an artificial mouth installed in front of the microphone of a fixed handset frame and transmitting this input signal through a different telephone line for each sentence (local or long distance network). The signal is sampled at 16 kHz, but its useful bandwidth is limited to telephone bandwidth (approximately 300-3400 Hz). Each sample is represented on 16 bits (linear).

5.3 Signal analysis

Each sentence is analysed as followed : the speech signal is decomposed in frames of 504 samples (31.5 ms) at a frame rate of 160 samples (10 ms). A Hamming window is applied to each frame. The signal is not pre-emphasized. For each frame, a Winograd Fourier Transform is computed and provides 252 square module values representing the short term power spectrum in the 0-8 kHz band.
This Fourier power spectrum is then used to compute 24 filter bank coefficients. Each filter is triangular (except the first and last ones which have a rectangle trapezoidal shape). They are placed on a non-uniform frequency scale, similar to the Bark/Mel scale. The central frequency of the 24 filters are, in Hz : 47, 147, 257, 378, 510, 655, 813, 987, 1178, 1386, 1615, 1866, 2141, 2442, 2772, 3133, 3529, 3964, 4440, 4961, 5533, 6159, 6845, and 7597. Each filter covers a spectral range from the central frequency of the previous filter to the central frequency of the

next filter, with a maximum value of 1 for its own central frequency. For each frequency, only 2 filters (maximum) are non-zero, and their magnitudes add up to 1.
We finally take the base 10 logarithm of each filter output and multiply the result by 10, to form a 24-dimensional vector of filter bank coefficients in dB. For the TIMIT database, all 24 coefficients are kept, from which we compute, for each utterance a 24-dimensional mean vector and a 24×24 (symmetric) covariance matrix.
In order to simulate, for some of the experiments, a low-pass filtering of the speech signal in the 0-4 kHz band, we have simply discarded the last 7 coefficients of the 24-dimensional vectors obtained from the full band signal. The last filter, with index 17, has a central frequency of 3529 Hz, and becomes zero above 3964 Hz. This is the approach we used for NTIMIT database, since the useful bandwidth does not exceed 4000 Hz for these data. We also used this approach on TIMIT, in order to obtain results corresponding to a 0-4 kHz bandwidth, without the telephone line variability. We will refer to these data as FTIMIT data. Under these analysis conditions, each mean vector is 17 dimensional, while covariance matrices are 17×17 (symmetric) matrices.

5.4 Training and test protocols

We use 2 training protocols, namely a "long training" and a "short training".

- For the **"long training"**, we use all 5 "sx" sentences concatenated together as a single reference pattern for each speaker. The average total duration of a "long training" pattern is **14.4 seconds**. A single reference (mean vector $\bar{x}$, covariance matrix X and number of vectors M) is computed for each speaker from all speech frames, represented as filter bank coefficients. In particular, no speech activity detector is used to remove silent speech portions.

- For the **"short training"**, we only use the first 2 "sx" sentences in alphanumeric order, in the same way as for the "long training". The average total duration of a "short training" is **5.7 seconds** (including silences).

For the tests, we also have 2 distinct protocols : a "long test" and a "short test".

- For the **"long test"**, all "sa" and "si" sentences (5 in total) are concatenated together as a single test pattern, for each speaker. In this framework, we therefore have a single test pattern per speaker, i.e. **630 test patterns** altogether. In average, each pattern lasts **15.9 seconds**.

- For the **"short test"**, each "sa" and "si" sentences are tested separately. The whole test set thus consists of $630 \times 5 =$ **3150 test patterns**, of **3.2 seconds** each, in average.

Even though the "sa" sentences are the same for each speaker, these sentences are used in the test set. Therefore, the experiments can be considered as totally text-independent.

5.5 Experiments

In the experiments reported in this article, we have systematically tested the 4 families of measures :

- μ_G
- μ_{Gc}
- μ_{Sc}
- μ_{Dc}

in two asymmetric forms :

- $\mu\ (\mathcal{X},\ \mathcal{Y})$
- $\mu\ (\mathcal{Y},\ \mathcal{X})$

as well as in the three symmetric forms proposed in section 4 :

- $\mu_{[0.5]}\ (\mathcal{X},\ \mathcal{Y})$
- $\mu_{[\alpha_{MN}]}\ (\mathcal{X},\ \mathcal{Y})$
- $\mu_{[\beta_{MN}]}\ (\mathcal{X},\ \mathcal{Y})$

These evaluations were carried out on :

- TIMIT (24 filter bank coefficients between 0 and 8000 Hz)
- FTIMIT (first 17 filter bank coefficients from TIMIT between 0 and 4000 Hz)
- NTIMIT (17 filter bank coefficients between 0 and 4000 Hz)

It is reasonable to expect that the 3 databases, in this order, correspond to an increasing degree of difficulty.
In each case, we give the results for 4 possible training $\times$ test protocols, corresponding to various typical values of the total amount of speech material per speaker $T = M + N$, of the ratio ρ between test and training material, and therefore to different weighting factors α_{MN} and β_{MN} :

- **"long-long"** protocol : long training $\times$ long test
  $\bar{T} \approx 3000\,cs$, $\bar{\rho} \approx 1.10$, $\bar{\alpha}_{MN} \approx 0.48$, $\bar{\beta}_{MN} \approx 0.49$
- **"short-long"** protocol : short training $\times$ long test
  $\bar{T} \approx 2150\,cs$, $\bar{\rho} \approx 2.79$, $\bar{\alpha}_{MN} \approx 0.26$, $\bar{\beta}_{MN} \approx 0.37$
- **"long-short"** protocol : long training $\times$ short test
  $\bar{T} \approx 1750\,cs$, $\bar{\rho} \approx 0.22$, $\bar{\alpha}_{MN} \approx 0.82$, $\bar{\beta}_{MN} \approx 0.68$
- **"short-short"** protocol : short training $\times$ short test
  $\bar{T} \approx\ \ 900\,cs$, $\bar{\rho} \approx 0.56$, $\bar{\alpha}_{MN} \approx 0.64$, $\bar{\beta}_{MN} \approx 0.57$

5.6 Results

The results are organized in 3 sets of 4 tables. The first set of tables (numbered I.1, I.2, I.3 and I.4) corresponds to results for TIMIT, the second set (Tables II.1 to II.4) for FTIMIT and the third set (Tables III.1 to III.4) for NTIMIT. The first table of each set (i.e. Tables I.1, II.1 and III.1) reports the results obtained for the "long-long" protocol, while the second one (I.2, II.2 and III.2) reports those for the "short-long" protocol. Similarly, the third and fourth tables of each set correspond respectively to the "long-short" and "short-short" protocols. In each table, the results relative to a given family of measures are organized in columns. The first line corresponds to the scores of both asymmetric terms (each cell is subdivided into 2), while the second, third and fourth lines show the results for the various symmetric forms. All results are given in terms of percentage of correct identification. Depending on this percentage S, and on the number of test patterns n, we give in Table 0 the half-width of the 95 % confidence interval, which is calculated as :

$$\pm 2 \sqrt{\frac{S \cdot (100 - S)}{n}}$$

Note that this quantity is the same for a score S and for $100 - S$.

score : S | 100 - S	95 | 5	85 | 15	75 | 25	65 | 35	55 | 45
long test, n = 630	± 1.7 %	± 2.8 %	± 3.5 %	± 3.8 %	± 4.0%
short test, n = 3150	± 0.8 %	± 1.3 %	± 1.5 %	± 1.7 %	± 1.8%

Table 0 : *Half-width of the 95 % confidence interval for different values of the identification score S in %, corresponding to the long and short test protocols.*

We will not comment in detail each performance figure in Tables I, II and III, but we will rather try to underline several global trends.

For all measures, $\mu\ (\mathcal{X},\ \mathcal{Y})$ and $\mu\ (\mathcal{Y},\ \mathcal{X})$ perform differently. The term $\mu\ (\mathcal{X},\ \mathcal{Y})$ performs better when the training speech material has a longer duration than the test material, and conversely. The discrepancy between the performances of the asymmetric terms is especially obvious for measure μ_{Dc}

With non-distorted speech (TIMIT and FTIMIT), measure μ_G outperforms measure μ_{Gc} and all other measures on covariance matrices only. On the opposite, when channel variability is present (NTIMIT), the use of the mean vectors is, as expected, detrimental to the results.

In their asymmetric forms, the most efficient measure among the covariance-only measures is measure μ_{Sc}. However, when symmetrisation is applied, the performances tend to level-off, with a slight advantage for μ_{Dc}.

Among the symmetrization procedures that we tested, the most efficient one seems to be the one using weights β_{MN} and β_{NM} for μ_G, μ_{Gc} and μ_{Sc}, whereas α_{MN} and α_{NM} appear to be preferable for $log\ \mu_{Dc}$.

Measures	μ_G		μ_{Gc}		μ_{Sc}		μ_{Dc}	
$\mu(\mathcal{X},\mathcal{Y})$ | $\mu(\mathcal{Y},\mathcal{X})$	100 %	100 %	100 %	99.8 %	100 %	100 %	99.5 %	99.8 %
$\mu_{[0.5]}(\mathcal{X},\mathcal{Y})$	100 %		100 %		100 %		100 %	
$\mu_{[\alpha_{MN}]}(\mathcal{X},\mathcal{Y})$	100 %		100 %		100 %		100 %	
$\mu_{[\beta_{MN}]}(\mathcal{X},\mathcal{Y})$	100 %		100 %		100 %		100 %	

Table I.1: *long training (5 sentences $\approx$ 14.4 s) – long test (5 sentences $\approx$ 15.9 s)*
$\bar{T} \approx 3000\ cs$, $\bar{\rho} \approx 1.10$, $\bar{\alpha}_{MN} \approx 0.48$, $\bar{\beta}_{MN} \approx 0.49$

Measures	μ_G		μ_{Gc}		μ_{Sc}		μ_{Dc}	
$\mu(\mathcal{X},\mathcal{Y})$ | $\mu(\mathcal{Y},\mathcal{X})$	93.2 %	99.4 %	86.7 %	97.1 %	94.9 %	96.4 %	73.3 %	92.1 %
$\mu_{[0.5]}(\mathcal{X},\mathcal{Y})$	98.1 %		94.6 %		95.7 %		95.1 %	
$\mu_{[\alpha_{MN}]}(\mathcal{X},\mathcal{Y})$	98.7 %		95.7 %		96.2 %		97.0 %	
$\mu_{[\beta_{MN}]}(\mathcal{X},\mathcal{Y})$	99.2 %		96.5 %		96.0 %		97.0 %	

Table I.2: *short training (2 sentences $\approx$ 5.7 s) – long test (5 sentences $\approx$ 15.9 s)*
$\bar{T} \approx 2150\ cs$, $\bar{\rho} \approx 2.79$, $\bar{\alpha}_{MN} \approx 0.26$, $\bar{\beta}_{MN} \approx 0.37$

Measures	μ_G		μ_{Gc}		μ_{Sc}		μ_{Dc}	
$\mu(\mathcal{X},\mathcal{Y})$ | $\mu(\mathcal{Y},\mathcal{X})$	97.9 %	89.7 %	96.2 %	78.8 %	97.3 %	93.6 %	83.6 %	59.4 %
$\mu_{[0.5]}(\mathcal{X},\mathcal{Y})$	97.2 %		93.9 %		97.0 %		97.3 %	
$\mu_{[\alpha_{MN}]}(\mathcal{X},\mathcal{Y})$	98.4 %		97.1 %		97.3 %		97.6 %	
$\mu_{[\beta_{MN}]}(\mathcal{X},\mathcal{Y})$	98.4 %		97.6 %		97.6 %		94.8 %	

Table I.3: *long training (5 sentences $\approx$ 14.4 s) – short test (1 sentence $\approx$ 3.2 s)*
$\bar{T} \approx 1750\ cs$, $\bar{\rho} \approx 0.22$, $\bar{\alpha}_{MN} \approx 0.82$, $\bar{\beta}_{MN} \approx 0.68$

Measures	μ_G		μ_{Gc}		μ_{Sc}		μ_{Dc}	
$\mu(\mathcal{X},\mathcal{Y})$ | $\mu(\mathcal{Y},\mathcal{X})$	83.8 %	78.2 %	73.5 %	64.9 %	81.9 %	77.7 %	52.9 %	45.4 %
$\mu_{[0.5]}(\mathcal{X},\mathcal{Y})$	89.7 %		82.2 %		82.7 %		84.4 %	
$\mu_{[\alpha_{MN}]}(\mathcal{X},\mathcal{Y})$	90.1 %		83.4 %		83.0 %		84.2 %	
$\mu_{[\beta_{MN}]}(\mathcal{X},\mathcal{Y})$	89.7 %		83.6 %		83.3 %		80.1 %	

Table I.4: *short training (2 sentences $\approx$ 5.7 s) – short test (1 sentence $\approx$ 3.2 s)*
$\bar{T} \approx\ \ 900\ cs$, $\bar{\rho} \approx 0.56$, $\bar{\alpha}_{MN} \approx 0.64$, $\bar{\beta}_{MN} \approx 0.57$

Tables I.1, I.2, I.3, I.4 :
Text-independent speaker identification – TIMIT database (630 speakers).
The results are given in percentage of correct identification.

Measures	μ_G		μ_{Gc}		μ_{Sc}		μ_{Dc}	
$\mu(\mathcal{X},\mathcal{Y})$ | $\mu(\mathcal{Y},\mathcal{X})$	98.4 %	99.4 %	96.1 %	98.3 %	97.6 %	98.1 %	90.5 %	95.4 %
$\mu_{[0.5]}(\mathcal{X},\mathcal{Y})$	99.4 %		97.9 %		97.9 %		98.3 %	
$\mu_{[\alpha_{MN}]}(\mathcal{X},\mathcal{Y})$	99.5 %		97.9 %		97.9 %		98.6 %	
$\mu_{[\beta_{MN}]}(\mathcal{X},\mathcal{Y})$	99.5 %		97.9 %		97.8 %		98.4 %	

Table II.1: *long training (5 sentences $\approx$ 14.4 s) – long test (5 sentences $\approx$ 15.9 s)* $\bar{T} \approx 3000\ cs$, $\bar{\rho} \approx 1.10$, $\bar{\alpha}_{MN} \approx 0.48$, $\bar{\beta}_{MN} \approx 0.49$

Measures	μ_G		μ_{Gc}		μ_{Sc}		μ_{Dc}	
$\mu(\mathcal{X},\mathcal{Y})$ | $\mu(\mathcal{Y},\mathcal{X})$	78.7%	88.6%	63.2 %	77.0 %	72.9 %	76.4 %	44.1 %	65.6 %
$\mu_{[0.5]}(\mathcal{X},\mathcal{Y})$	87.0 %		76.8 %		76.4 %		76.7 %	
$\mu_{[\alpha_{MN}]}(\mathcal{X},\mathcal{Y})$	87.9 %		77.5 %		76.2 %		77.6 %	
$\mu_{[\beta_{MN}]}(\mathcal{X},\mathcal{Y})$	89.0 %		77.8 %		76.4 %		76.2 %	

Table II.2: *short training (2 sentences $\approx$ 5.7 s) – long test (5 sentences $\approx$ 15.9 s)* $\bar{T} \approx 2150\ cs$, $\bar{\rho} \approx 2.79$, $\bar{\alpha}_{MN} \approx 0.26$, $\bar{\beta}_{MN} \approx 0.37$

Measures	μ_G		μ_{Gc}		μ_{Sc}		μ_{Dc}	
$\mu(\mathcal{X},\mathcal{Y})$ | $\mu(\mathcal{Y},\mathcal{X})$	81.4 %	67.9 %	70.0 %	49.8 %	70.7 %	66.3 %	48.1 %	33.3 %
$\mu_{[0.5]}(\mathcal{X},\mathcal{Y})$	81.8 %		67.3 %		70.4 %		72.2 %	
$\mu_{[\alpha_{MN}]}(\mathcal{X},\mathcal{Y})$	84.2 %		71.8 %		71.7 %		73.1 %	
$\mu_{[\beta_{MN}]}(\mathcal{X},\mathcal{Y})$	83.6 %		72.6 %		72.0 %		64.4 %	

Table II.3: *long training (5 sentences $\approx$ 14.4 s) – short test (1 sentence $\approx$ 3.2 s)* $\bar{T} \approx 1750\ cs$, $\bar{\rho} \approx 0.22$, $\bar{\alpha}_{MN} \approx 0.82$, $\bar{\beta}_{MN} \approx 0.68$

Measures	μ_G		μ_{Gc}		μ_{Sc}		μ_{Dc}	
$\mu(\mathcal{X},\mathcal{Y})$ | $\mu(\mathcal{Y},\mathcal{X})$	54.7%	49.7%	39.8 %	32.2 %	42.6 %	41.2 %	23.1 %	20.6 %
$\mu_{[0.5]}(\mathcal{X},\mathcal{Y})$	61.4 %		43.9 %		44.4 %		46.5 %	
$\mu_{[\alpha_{MN}]}(\mathcal{X},\mathcal{Y})$	61.8 %		45.3 %		44.5 %		46.8 %	
$\mu_{[\beta_{MN}]}(\mathcal{X},\mathcal{Y})$	61.4 %		45.8 %		44.4 %		43.6 %	

Table II.4: *short training (2 sentences $\approx$ 5.7 s) – short test (1 sentence $\approx$ 3.2 s)* $\bar{T} \approx\ 900\ cs$, $\bar{\rho} \approx 0.56$, $\bar{\alpha}_{MN} \approx 0.64$, $\bar{\beta}_{MN} \approx 0.57$

Tables II.1, II.2, II.3, II.4 :
Text-independent speaker identification – FTIMIT database (630 speakers).
The results are given in percentage of correct identification.

Measures	μ_G		μ_{Gc}		μ_{Sc}		μ_{Dc}	
$\mu(\mathcal{X},\mathcal{Y})$ | $\mu(\mathcal{Y},\mathcal{X})$	45.3 %	50.3 %	59.5 %	63.0 %	66.0 %	64.9 %	41.0 %	51.0 %
$\mu_{[0.5]}(\mathcal{X},\mathcal{Y})$	49.4 %		63.0 %		66.4 %		67.9 %	
$\mu_{[\alpha_{MN}]}(\mathcal{X},\mathcal{Y})$	49.0 %		63.0 %		66.5 %		68.4 %	
$\mu_{[\beta_{MN}]}(\mathcal{X},\mathcal{Y})$	49.4 %		63.6 %		66.5 %		68.6 %	

Table III.1: *long training (5 sentences $\approx$ 14.4 s) – long test (5 sentences $\approx$ 15.9 s)*
$\bar{T} \approx 3000\ cs$, $\bar{\rho} \approx 1.10$, $\bar{\alpha}_{MN} \approx 0.48$, $\bar{\beta}_{MN} \approx 0.49$

Measures	μ_G		μ_{Gc}		μ_{Sc}		μ_{Dc}	
$\mu(\mathcal{X},\mathcal{Y})$ | $\mu(\mathcal{Y},\mathcal{X})$	17.6 %	24.9 %	22.2 %	31.0 %	28.4 %	29.7 %	12.4 %	22.5 %
$\mu_{[0.5]}(\mathcal{X},\mathcal{Y})$	24.4 %		29.5 %		29.8 %		30.3 %	
$\mu_{[\alpha_{MN}]}(\mathcal{X},\mathcal{Y})$	24.8 %		30.5 %		30.0 %		30.8 %	
$\mu_{[\beta_{MN}]}(\mathcal{X},\mathcal{Y})$	25.7 %		31.3 %		30.5 %		30.8 %	

Table III.2: *short training (2 sentences $\approx$ 5.7 s) – long test (5 sentences $\approx$ 15.9 s)*
$\bar{T} \approx 2150\ cs$, $\bar{\rho} \approx 2.79$, $\bar{\alpha}_{MN} \approx 0.26$, $\bar{\beta}_{MN} \approx 0.37$

Measures	μ_G		μ_{Gc}		μ_{Sc}		μ_{Dc}	
$\mu(\mathcal{X},\mathcal{Y})$ | $\mu(\mathcal{Y},\mathcal{X})$	20.7 %	13.8 %	25.4 %	13.5 %	26.3 %	23.0 %	14.1 %	5.2 %
$\mu_{[0.5]}(\mathcal{X},\mathcal{Y})$	19.3 %		23.4 %		25.2 %		25.2 %	
$\mu_{[\alpha_{MN}]}(\mathcal{X},\mathcal{Y})$	21.1 %		25.4 %		26.1 %		26.4 %	
$\mu_{[\beta_{MN}]}(\mathcal{X},\mathcal{Y})$	21.4 %		26.1 %		26.7 %		20.5 %	

Table III.3: *long training (5 sentences $\approx$ 14.4 s) – short test (1 sentence $\approx$ 3.2 s)*
$\bar{T} \approx 1750\ cs$, $\bar{\rho} \approx 0.22$, $\bar{\alpha}_{MN} \approx 0.82$, $\bar{\beta}_{MN} \approx 0.68$

Measures	μ_G		μ_{Gc}		μ_{Sc}		μ_{Dc}	
$\mu(\mathcal{X},\mathcal{Y})$ | $\mu(\mathcal{Y},\mathcal{X})$	10.1 %	9.0 %	12.0 %	8.9 %	13.7 %	12.8 %	6.4 %	3.1 %
$\mu_{[0.5]}(\mathcal{X},\mathcal{Y})$	11.7 %		13.7 %		14.3 %		14.4 %	
$\mu_{[\alpha_{MN}]}(\mathcal{X},\mathcal{Y})$	11.7 %		14.2 %		14.4 %		14.8 %	
$\mu_{[\beta_{MN}]}(\mathcal{X},\mathcal{Y})$	11.6 %		15.0 %		14.4 %		12.7 %	

Table III.4: *short training (2 sentences $\approx$ 5.7 s) – short test (1 sentence $\approx$ 3.2 s)*
$\bar{T} \approx\ 900\ cs$, $\bar{\rho} \approx 0.56$, $\bar{\alpha}_{MN} \approx 0.64$, $\bar{\beta}_{MN} \approx 0.57$

Tables III.1, III.2, III.3, III.4 :
Text-independent speaker identification – NTIMIT database (630 speakers).
The results are given in percentage of correct identification.

The positive effect of symmetrization is important when little speech material is available. The most significant differences are observed for the short training × short test protocol. Table IV gives orders of magnitude of the relative error rate reduction between the asymmetric measures and their best symmetric version. If S is the percentage of correct identification for the asymmetric measure and S' is the percentage of correct identification for the symmetric measure, the relative error rate reduction is calculated as :

$$\frac{S' - S}{100 - S}$$

This relative improvement is given for the two protocols using short duration test data only. For the two others, the statistical significance of the observed differences are too small to be conclusive, given the larger confidence interval.

measure	μ_G	μ_{Gc}	μ_{Sc}	μ_{Dc}
TIMIT	~ 30 %	~ 40 %	~ 10 %	~ 75 %
FTIMIT	~ 15 %	~ 10 %	~ 5 %	~ 40 %
NTIMIT	~ 1 %	~ 1 %	< 1 %	~ 10 %

Table IV : *Order of magnitude of the relative error rate reduction between asymmetric and symmetric measures. Results for short test protocols only.*

These results show that symmetrization improves covariance-only measures (μ_{Gc}, μ_{Sc} and μ_{Dc}) as the task becomes intrinsically less difficult (TIMIT > FTIMIT > NTIMIT), and as the original asymmetric measures perform less well ($\mu_{Dc} < \mu_{Gc} < \mu_{Sc}$). On the other hand, when the Gaussian speaker model is not powerful enough for the task (NTIMIT), or when the asymmetric measure is quite efficient (μ_{Sc}), symmetrization is less useful.

6 Discussion

Our evaluations show that remarkable performances can be obtained on the TIMIT database for text-independent closed-set speaker identification (630 speakers) by second-order statistical measures, i.e. with a very simple underlying speaker model. Therefore, TIMIT is certainly an easy database for speaker recognition, and the measures exposed in this article work very well, on this database. Naturally, their overall performances degrade with more adverse conditions : a significant amount of speaker characteristics seems to be contained in the 4–8 kHz band, since FTIMIT results are significantly worse than TIMIT results. The effect of telephone channel distortion and variability are the cause of an even more severe drop on NTIMIT recognition scores. The effect of temporal drift owed to multisession recordings can not be studied with TIMIT derived data, but it is easy to predict an additional negative role of this factor on the performances. If second-order statistical measures are clearly efficient for relatively simple tasks, they are obviously not the ultimate solution to speaker recognition for any kind of applications.

6.1 Beyond the performances

However, second-order statistical measures have several advantages. They are simple to implement and easy to reproduce. Moreover, Gaussian likelihood measures (μ_G and μ_{Gc}) in their asymmetric forms are particular cases of several general approaches frequently used in text-independent speaker recognition. A 1-Gaussian speaker model is equivalent to a Vector Quantization codebook with 1 entry associated with a Mahalanobis distance. It is also equivalent to any kind of Hidden Markov Model (Left-to-Right, Ergodic,...) with 1 state and 1 Gaussian distribution. It is a particular case of a k-Gaussian Mixture model with k = 1. Finally, the likelihood criterion is often used on vector prediction residuals obtained from linear or connectionist models for which the identity model (0^{th}-order prediction) is a particular case.
Therefore, μ_G and μ_{Gc} are at the intersection of several classical approaches, which are extensions of this basic model in various directions (variations of the distance measure, use of more or less strong temporal constraints, refinement of the speaker distribution model, filtering of the acoustic parameters,...). Given the extreme simplicity of the second-order statistical measures, we therefore suggest that any speaker recognition task could be systematically benchmarked by one or two of these measures, in order to obtain a reference score indicating the intrinsic complexity of the chosen database and protocol. In particular, the preprocessings, the acoustic analysis, the training and test splitting of the data, and the decision strategy for the method under test should be identically used for the benchmark method.

6.2 A possible reference approach

Even though asymmetric Gaussian likelihood based measures do not systematically perform better than other second-order statistical measures, μ_G and μ_{Gc} may be preferable as reference benchmark measures in two cases : when they are compared with other asymmetric approaches (which is the case for VQ, HMM and Gaussian Mixtures), and when the length of training material is significantly higher than the length of test material. The choice between μ_G and μ_{Gc} should be guided by the processing that is applied to the data for the system under evaluation : whether, for this particular protocol, the long term average is substracted or not to the acoustic parameters. Measures $\mu_{G_{[\beta_{MN}]}}$ or $\mu_{Gc_{[\beta_{MN}]}}$ can also be implemented simply and could be systematically tested. However, the lack of theoretical justification for these measures, and the relatively small improvement they provide as soon as a reasonable amount of speech material is available, make it more debatable. Nevertheless, if the approach under test is formally symmetric, it would be fair to compare it to a symmetric reference measure.

7 Conclusion

The goal of this work has been multiple. Firstly, to investigate the properties and performances of simple speaker recognition approaches, to compare them and to identify their limits. Our large scale evaluation on TIMIT, on a low-pass filtered version of TIMIT and on NTIMIT illustrates clearly that speech quality and quantity are major factors of performance, and that on high quality contemporaneous speech, simple and fast methods can be extremely efficient. For instance, this type of approach may prove sufficient for applications such as the automatic speaker labeling of radio or television recordings, for which the signal quality is constant and the voice drift relatively marginal.
Secondly, our work illustrates the extreme caution with which any conclusion can be drawn on the merit of a given method outside of any point of comparison. Since it may not be feasible to compare any new method with all state-of-the-art approaches, it is at least desirable to benchmark the task with a simple and general reference approach.
Thirdly, we believe that second-order statistical tests and measures, based on the Gaussian likelihood scoring are a good compromise as reference measures, since they are easy to implement, simple to reproduce, inexpensive in computation and light in storage requirements. Moreover, they appear, in their asymmetric forms, as simpler versions of more elaborate approaches. Though symmetrization is not a systematic factor of improvement, symmetric versions of the measures could be tested as well, especially in comparison with other symmetric methods. More theoretical work on symmetrization is however needed to find optimal symmetric forms.
The systematical use of a reference method in order to calibrate the complexity of a speaker recognition task can only result from a consensus between researchers both on the concept of a benchmark evaluation by a common approach and on the choice of the reference algorithm itself. We hope that this article will contribute to widen the concertation that had started during the SAM-A European ESPRIT project, dedicated to speech assessment methodology.

References

[1] T. W. Anderson. *An Introduction to Multivariate Analysis*. John Wiley and Sons, 1958.

[2] T. Artières, Y. Bennani, P. Gallinari, and C. Montacié. Connectionnist and conventional models for text-free talker identification tasks. In *Proceedings of NEURONIMES 91*, 1991. Nîmes, France.

[3] Y. Bennani. Speaker identification through a modular connectionist architecture: evaluation on the TIMIT database. In *ICSLP 92*, volume 1, pages 607–610, Oct. 1992. Banff, Canada.

[4] F. Bimbot and L. Mathan. Text-free speaker recognition using an arithmetic-harmonic sphericity measure. In *EUROSPEECH 93*, volume 1, pages 169–172, Sept. 1993. Berlin, Germany.

[5] F. Bimbot, L. Mathan, A. de Lima, and G. Chollet. Standard and target-driven AR-vector models for speech analysis and speaker recognition. In *ICASSP 92*, volume 2, pages II.5–II.8, Mar. 1992. San Francisco, United-States.

[6] F. Bimbot, A. Paoloni, and G. Chollet. *Assessment Methodology for Speaker Identification and Verification Systems*. Technical report – Task 2500 – Report I9, SAM-A ESPRIT Project 6819, 1993.

[7] G. Chollet and C. Gagnoulet. On the evaluation of speech recognizers and data bases using a reference system. In *ICASSP 82*, volume 3, pages 2026–2029, May 1982. Paris, France.

[8] W. M. Fisher, G. R. Doddington, and K. M. Goudie-Marshall. The DARPA speech recognition research database : specifications and status. In *Proceedings of the DARPA workshop on speech recognition*, pages 93–99, Feb. 1986.

[9] S. Furui. An overview of speaker recognition technology. In *Workshop on automatic speaker recognition, identification and verification*, pages 1–9, Apr. 1994. Martigny, Switzerland.

[10] H. Gish. Robust discrimination in automatic speaker identification. In *ICASSP 90*, volume 1, pages 289–292, Apr. 1990. New Mexico, United-States.

[11] H. Gish, M. Krasner, W. Russell, and J. Wolf. Methods and experiments for text-independent speaker recognition over telephone channels. In *ICASSP 86*, volume 2, pages 865–868, Apr. 1986. Tokyo, Japan.

[12] Y. Grenier. *Identification du locuteur et adaptation au locuteur d'un système de reconnaissance phonémique*. PhD thesis, ENST, 1977.

[13] Y. Grenier. Utilisation de la prédiction linéaire en reconnaissance et adaptation au locuteur. In *XIèmes Journées d'Etude sur la Parole*, pages 163–171, May 1980. Strasbourg, France.

[14] H. Hattori. Text-independent speaker recognition using neural networks. In *ICASSP 92*, volume 2, pages 153–156, Mar. 1992. San Francisco, United-States.

[15] C. Jankowski, A. Kalyanswamy, S. Basson, and J. Spitz. NTIMIT: a phonetically balanced, continuous speech, telephone bandwidth speech database. *ICASSP 90*, Apr. 1990. New Mexico, United-States.

[16] C. Montacié, P. Deléglise, F. Bimbot, and M.-J. Caraty. Cinematic techniques for speech processing: temporal decomposition and multivariate linear prediction. In *ICASSP 92*, volume 1, pages 153–156, Mar. 1992. San Francisco, United-States.

[17] C. Montacié and J.-L. Le Floch. AR-vector models for free-text speaker recognition. In *ICSLP 92*, volume 1, pages 611–614, Oct. 1992. Banff, Canada.

[18] A. B. Poritz. Linear predictive Hidden Markov Models and the speech signal. In *ICASSP 82*, pages 1291–1294, May 1982. Paris, France.

[19] D. A. Reynolds. Speaker identification and verification using Gaussian Mixture speaker models. In *Workshop on automatic speaker recognition, identification and verification*, pages 27–30, Apr. 1994. Martigny, Switzerland.

[20] L. Rudasi and S. A. Zahorian. Text-independent talker identification with neural networks. In *ICASSP 91*, volume 1, pages 389–392, 1991. Toronto, Canada.

[21] M. Savic and S. K. Gupta. Variable parameter speaker verification system based on hidden markov modeling. In *ICASSP 90*, volume 1, pages 281–284, Apr. 1990. Albuquerque, New Mexico, United-States.

[22] F. K. Soong, A. E. Rosenberg, and B.-H. Juang. A Vector Quantization approach to speaker recognition. Technical Report 3, AT&T, Mar. 1987.

EFFECT OF UTTERANCE DURATION AND PHONETIC CONTENT ON SPEAKER IDENTIFICATION USING SECOND-ORDER STATISTICAL METHODS

Ivan MAGRIN-CHAGNOLLEAU[◇] *Jean-François BONASTRE*[⋆] *Frédéric BIMBOT*[◇]

[◇]Télécom Paris (E.N.S.T.) – Dépt. Signal, C.N.R.S. URA 820
46, rue Barrault – F-75634 Paris Cedex 13 – FRANCE – European Union
email: ivan@sig.enst.fr and bimbot@sig.enst.fr

[⋆]Laboratoire d'Informatique Université d'Avignon
33, rue Louis Pasteur – F-84000 Avignon – FRANCE – European Union
email: jfb@univ-avignon.fr

ABSTRACT

Second-order statistical methods show very good results for automatic speaker identification in controlled recording conditions [2]. These approaches are generally used on the entire speech material available. In this paper, we study the influence of the content of the test speech material on the performances of such methods, i.e. under a more analytical approach [3]. The goal is to investigate on the kind of information which is used by these methods, and where it is located in the speech signal. Liquids and glides together, vowels, and more particularly nasal vowels and nasal consonants, are found to be particularly speaker specific: test utterances of 1 second, composed in majority of acoustic material from one of these classes provide better speaker identification results than phonetically balanced test utterances, even though the training is done, in both cases, with 15 seconds of phonetically balanced speech. Nevertheless, results with other phoneme classes are never dramatically poor. These results tend to show that the speaker-dependent information captured by long-term second-order statistics is consistently common to all phonetic classes, and that the homogeneity of the test material may improve the quality of the estimates.

1. INTRODUCTION

In this paper, the influence of phonetic content on the performance of a speaker identification system is investigated. Second-order statistical methods are chosen because they provide very good results with a low quantity of computations and with a restricted quantity of speech material, provided recording conditions and channel distortions are controlled [2]. The goal of this work is to study how the performances of this family of approaches vary with the phonetic content of the test material.
Several experiments were previously reported on the relative speaker discriminating properties of phonemes. In particular, Eatock et al. [5] used a VQ codebook based approach and concluded that nasals and vowels provided the best performances on an English language database. Le Floch et al. [9] investigated the properties of AR-vector models and concluded that vowels, diphtongs and nasals provided the best performances, also on an English language database.
The specificity of the work reported in this paper comes from the fact that second-order statistical methods are used and that the training and test material are heterogeneous. The general experimental framework is the following: 15 seconds of training material coming from phonetically balanced sentences, are used to build a reference model for each speaker. Then, specific speech segments are selected from other phonetically balanced sentences in order to build a test pattern which is strongly biased towards a particular phoneme or phoneme class.
The database used for our experiments contains 67 cooperative speakers recorded in a slightly noisy environment. For each speaker, 50 French phonetically balanced sentences, i.e approximately 3 minutes of speech are recorded in a single session. This volume of speech allows to have a sufficient number of occurences for most phonemes.
The three second-order statistical measures used in the experiments are described in section 2.. Details on the speech database and on the signal analysis are given in section 3.. Section 4. reports preliminary experiments on utterance duration, and section 5. the experiments on the phonetic content. Finally, section 6. concludes on this work and gives some perspectives.

2. SPEAKER IDENTIFICATION MEASURES

The 3 speaker identification methods used in this work are inspired from statistical tests on covariance matrices [1], computed on acoustic parameters. The first two speaker similarity measures are directly derived from maximum likelihood Gaussian classifiers [7]. The third one is based on a sphericity test between covariance matrices [8]. The measures, asymmetric in their original form, are symmetrised as a weighted sum of the asymetric measure and its dual term. In previous work [2], this procedure was shown to improve performances.
Let X and Y denote two covariance matrices of a reference speaker and of a test speaker respectively, corresponding to the covariance of some spectral vectors computed along a sentence. Let $\bar{x}$ and $\bar{y}$ denote the means of the spectral vectors. Let M and N denote the number of spectral vectors used to estimate the covariance matrices and mean vectors, and p the dimension of the spectral vectors. The mathematical expression of the three measures that we used in our experiments, are given in Table 1. Note that, once the covariance matrices X and Y are inverted, and that their determinant is evaluated, the computation of the measures requires very few operations.

3. DATABASE AND SIGNAL ANALYSIS

3.1. Database

The corpus is composed of read phonetically balanced sentences [4], the phonetic transcription of which can be found in the BDSON database. The sentences are prompted on a screen. Recording begins and ends automatically using a speech activity detector. Each sentence is recorded through a SHURE SM10A microphon, and digitized at a 16 kHz

$$\mu_G(X,Y) = \frac{1}{p}\left[\frac{M}{M+N}\ tr(YX^{-1}) + \frac{N}{M+N}\ tr(XY^{-1}) - \frac{M-N}{M+N}\ \log\left(\frac{detY}{detX}\right)\right] + \frac{1}{p}\left[(\bar{y}-\bar{x})^T\left[\frac{M}{M+N}\ X^{-1} + \frac{N}{M+N}\ Y^{-1}\right](\bar{y}-\bar{x})\right] - 1$$

$$\mu_{Gc}(X,Y) = \frac{1}{p}\left[\frac{M}{M+N}\ tr(YX^{-1}) + \frac{N}{M+N}\ tr(XY^{-1}) - \frac{M-N}{M+N}\ \log\left(\frac{detY}{detX}\right)\right] - 1$$

$$\mu_{Sc}(X,Y) = \frac{M}{M+N}\ \log\ tr(YX^{-1}) + \frac{N}{M+N}\ \log\ tr(XY^{-1}) - \frac{1}{p}\frac{M-N}{M+N}\ \log\left(\frac{detX}{detY}\right) - \log p$$

Table 1. *Expressions of the three symmetrized second-order statistical measures ("tr" denote the trace and "det" the determinant of a matrix).*

sampling frequency on 16 bits by an OROS AU22 board. The recording equipment was set up in the corridor of a university, i.e with a non-negligible background noise. The recordings are single-session. 67 speakers took part to the experiments, mostly students. They each recorded approximately 3 minutes of speech.

3.2. Signal Analysis

The speech analysis module extracts filterbank coefficients in the following way: a Winograd Fourier Transform is computed on Hamming windowed signal frames of 31.5 ms (i.e 504 samples) at a frame rate of 10 ms (160 samples). For each frame, spectral vectors of 24 Mel-Scale Triangular-Filter Bank coefficients are then calculated from the Fourier Transform power spectrum, and expressed in logarithmic scale. Covariance matrices and mean vectors are finally computed from these spectral vectors.

4. UTTERANCE DURATION

The first set of experiments investigates on the influence of utterance duration for second-order statistical methods. This allowed us to choose meaningful training and test durations for the second part of the work. Several durations for training and test are chosen: 15, 10, 6, 3 and 2 seconds for training; 10, 6, 3, 2 and 1 second for testing. For each speaker, all sentences are randomly concatenated together. The silences at the beginning and the end of sentences are not removed, but they generally do not exceed 0.1 second. First, a certain amount of speech is selected for training (for example 15 seconds). Then, the rest is segmented in several test portions, each portion having a predetermined duration, until 20 test portions are obtained unless the speech material is exhausted. As a result of this experimental design, the experiments are text-independent.

Percentages of correct identifications are given in Table 2. The best measure is always μ_G, which confirms once again that the average of the spectral vectors is a significant source of speaker specific information, for good quality contemporaneous speech.

Even though the measures are symmetric, a clear asymmetry of the results can be observed: for example, with a training of 10 seconds and a test of 2 seconds, the percentage of correct identifications with μ_G is 95.3 %, while with a training of 2 seconds and a test of 10 seconds, it only reaches 88.4 %. In fact, if a training of 2 seconds is poorly representative of the speaker (for instance, if it contains a lot of silence), this has an effect on all the tests utterances from this very speaker, which affects considerably the overall score. If, conversely, a test of 2 seconds is of poor quality, it only causes 1 mistake, which has little impact on the global performance.

In the second part of our experiments, we chose a 15 second training duration, which is a guaranty to have a reliable training and a sufficient phonetic coverage. The test duration is chosen to 1 second for two reasons. Firstly, if a longer duration is chosen, the number of tests for a given phoneme is less significant. Secondly, if the percentage of correct identifications is too high, we felt that comparisons may lack statistical significance.

Test Duration		Training Duration 15 s	10 s	6 s	3 s	2 s
10 s (1095)	μ_G	**99.9**	**99.7**	**99.3**	**94.3**	**88.4**
	μ_{Gc}	99.8	99.7	98.1	92.0	85.5
	μ_{Sc}	99.8	99.7	98.3	91.1	85.3
6 s (1294)	μ_G	**99.9**	**99.5**	**98.8**	**93.0**	**87.2**
	μ_{Gc}	99.6	99.0	96.8	88.6	82.6
	μ_{Sc}	99.5	99.1	97.0	88.5	82.2
3 s (1340)	μ_G	**98.7**	**97.3**	**95.3**	**87.0**	**80.2**
	μ_{Gc}	97.8	96.0	92.0	82.5	74.8
	μ_{Sc}	98.1	96.5	93.0	82.4	74.0
2 s (1340)	μ_G	**97.5**	**95.3**	**91.7**	**83.3**	**74.3**
	μ_{Gc}	94.8	92.9	86.3	75.6	66.1
	μ_{Sc}	95.7	92.9	86.8	76.3	66.0
1 s (1340)	μ_G	**87.5**	**83.9**	**76.0**	**65.3**	**57.0**
	μ_{Gc}	79.5	73.5	65.1	53.3	47.8
	μ_{Sc}	83.6	79.0	71.7	57.9	50.7

Table 2. *Speaker identification: results in percentage of correct identifications for different training and test durations. The numbers of tests for each test duration is indicated in parentheses.*

5. PHONETIC CONTENT

5.1. Segmentation

In order to study the influence of phonetic content on the performances of the second-order statistical methods, we used an automatic system for segmenting the speech material into specific phonemes or phonetic classes. This system of automatic localisation is based on a bottom-up acoustic-phonetic decoder, which is speaker independent [3], [10], [6]. For each sentence, this decoder proposes a set of weighted phonetic hypotheses. These hypotheses are then aligned with the phonetic transcription of the sentence, by a left-right alignment algorithm. In order to obtain a high localisation accuracy, the algorithm is tuned with a high level of rejection for uncertain alignments: in this experiment, 55 % of the sentences were rejected. As a consequence, the phonetic events selected by this procedure can be considered as highly reliable, and quite typical in their category. The localisation algorithm gives a small kernel for a recognized phoneme. So the phoneme segments were extended to 5 frames before and 5 frames after the kernel. Therefore, the segments are not only composed with frames of the given phoneme, as they may also include a small proportion of transitions.

5.2. Experimental Protocol

For this set of experiments, training material consists of all speech frames derived from 15 seconds of phonetically balanced sentences, i.e. no specific phonetic events are selected. It is in fact the exact same material as the one used in the previous experiment on the influence of utterance duration (see section 4.), with training durations of 15 seconds.
For the tests, specific phonetic classes and phonemes are selected on the rest of the speech material, by the procedure described in section 5.1.. For a particular phoneme, all the frames labeled as this given phoneme are concatenated together, and this material is divided into as many tests of 1 second as possible.

5.3. Description of the phoneme classes

Results for all phoneme classes and phonemes are not presented in this article: we chose to report only on those for which more than 40 tests were carried out.
A first class, refered to as *All*, contains all the phonemes. The experiments with test data from this class is however slightly different from the one in section 4., with 15 seconds for training and 1 second for testing: speech material in the class *All* is composed of acoustic material clearly identified as a phoneme by the segmentation process, and in particular, it does not contain silences, pauses, or other non-linguistic events.
The other classes are: *Vowels* (which contains oral and nasal vowels but not glides), *Oral Vowels*, *Nasal Vowels*, *Consonants* (which contains also glides), *Non-Nasal Consonants* (which contains all the consonants except the nasal consonants), *Nasal Consonants*, *Stop Consonants*, *Fricatives* and *Liquids+Glides* (which form a single class).

5.4. Results

Results for various phoneme classes and individual phonemes are given in Table 3 and Table 4 respectively. The percentage of correct identifications on all the tests is given, as well as the average of the correct identifications per speaker. The number of tests is also indicated.

5.5. Comments

Quite surprisingly, measure μ_G still performs better than μ_{Gc} and μ_{Sc}: even though the mean vector within a phonetic class is expected to be strongly class dependent (and therefore not to match the training mean vector), it still keeps some consistence across phonetic classes.
The results for the class *All* outperform slightly those for the 15 s × 1 s experiment of section 4., probably because the speech material is, in the second case, more reliable.
A second observation is that the results for each phoneme class is higher than the result of the class *All*. It is also the case for most of the phonemes. This tends to show that a phonetically homogeneous test material benefits to the overall speaker identification performance, even though the training material does not share the same character.
In more detail, *Vowels* give better results than *Consonants*. *Nasal Vowels* outperform *Vowels* and *Oral Vowels*. *Non-Nasal Consonants*, and more particularly *Stop Consonants* or *Fricatives*, give lower performances than all *Consonants* together, whereas *Liquids+Glides* and *Nasal Consonants* yield higher scores. Note that the class *Liquids+Glides* gives the best results altogether, except with μ_{Sc} for which the classes *Nasal Consonants* and *Nasal Vowels* perform better.
In what concerns individual phonemes, the best scores with

Phonemes	*All* (1334)		*Vowels* (1247)	
	I	M	I	M
μ_G	**90.6**	90.5	**97.1**	97.0
μ_{Gc}	80.8	80.6	92.3	92.1
μ_{Sc}	83.5	83.4	92.0	91.7
Phonemes	*Oral Vowels* (1206)		*Nasal Vowels* (262)	
	I	M	I	M
μ_G	**96.0**	95.9	**98.1**	98.5
μ_{Gc}	91.3	91.2	89.3	90.5
μ_{Sc}	90.5	90.5	93.1	92.9
Phonemes	*Consonants* (1247)		*Non-Nasal Cons.* (1186)	
	I	M	I	M
μ_G	**96.2**	96.2	**94.7**	94.8
μ_{Gc}	91.1	91.0	89.4	89.3
μ_{Sc}	91.6	91.3	89.0	89.0
Phonemes	*Nasal Cons.* (390)		*Stop Cons.* (693)	
	I	M	I	M
μ_G	**96.9**	97.7	**94.2**	95.7
μ_{Gc}	85.9	87.6	91.2	92.6
μ_{Sc}	95.1	93.9	91.3	92.7
Phonemes	*Fricatives* (486)		*Liquids + Glides* (277)	
	I	M	I	M
μ_G	**92.2**	92.8	**98.9**	98.8
μ_{Gc}	83.7	84.8	92.4	92.3
μ_{Sc}	86.2	86.9	92.4	91.4

Table 3. *Speaker identification with speech material selected from specific phonetic classes. Training is composed of 15 s of phonetically balanced speech and test is composed of 1 s of phonetically biased speech. I = Global correct identification score, M = Average correct identification score over all speakers. The number of tests for each test configuration is indicated in parentheses.*

μ_G are obtained with the phonemes /o/, /d/, /ɛ/, and /ɑ̃/, whereas /k/, /ʒ/, /u/, /s/ and /y/ give the poorest performance levels.

6. CONCLUSION

Our experiments on the effect of the phonetic content on speaker identification using second-order statistics, tend to show that, on our database, the phonetic homogeneity of the test material is usually a significant factor of improvement, even if the training material is heterogeneous. The phonetic classes that yield particularly good results are *Liquids+Glides*, *Vowels* (and more particularly *Nasal Vowels*), and *Nasal Consonants*, but results for other classes are never dramatically poor. There may therefore exist some kind of speaker-dependent *tie* between acoustic distributions across phonemes that is captured by the second-order statistical methods. This hypothesis has to be confirmed on other types of speech data, in particular noisy speech and non-contemporaneous recordings.

REFERENCES

[1] T. W. Anderson. *An Introduction to Multivariate Analysis.* John Wiley and Sons, 1958.

[2] F. Bimbot, I. Magrin-Chagnolleau, and L. Mathan. Second-order statistical measures for text-independent

speaker identification. *Speech Communication*, 17(1–2):177–192, Aug. 1995.

[3] J.-F. Bonastre and H. Méloni. Automatic speaker recognition and analytic process. In *Proceedings of EUROSPEECH 93*, volume 1, pages 441–444, Sept. 1993. Berlin, Germany.

[4] P. Combescure. Vingt listes de dix phrases françaises phonétiquement équilibrées. Note technique, CNET (Centre Lannion A), Oct. 1980.

[5] J. P. Eatock and J. S. Mason. A quantitative assessment of the relative speaker discriminating properties of phonemes. In *Proceedings of ICASSP 94*, volume 1, pages 133–136, Apr. 1994. Adelaïde, Australia.

[6] P. Gilles. *Décodage phonétique de la parole et adaptation au locuteur*. PhD thesis, Université d'Avignon et des Pays de Vaucluse, Jan. 1993.

[7] H. Gish. Robust discrimination in automatic speaker identification. In *Proceedings of ICASSP 90*, volume 1, pages 289–292, Apr. 1990. New Mexico, United-States.

[8] Y. Grenier. *Identification du locuteur et adaptation au locuteur d'un système de reconnaissance phonémique*. PhD thesis, ENST, 1977.

[9] J.-L. Le Floch, C. Montacié, and M.-J. Caraty. Investigations on speaker characterization from Orphée system technics. In *Proceedings of ICASSP 94*, volume 1, pages 149–152, Apr. 1994. Adelaïde, Australia.

[10] H. Méloni and P. Gilles. Décodage acoustico-phonétique ascendant. *Traitement du Signal*, 8(2):107–114, 1991.

Phonemes	/i/ (174)		/e/ (105)	
	I	M	I	M
μ_G	**92.0**	92.2	**96.2**	97.5
μ_{Gc}	79.9	81.3	86.7	88.0
μ_{Sc}	81.0	82.5	87.6	88.9
Phonemes	/ɛ/ (258)		/y/ (66)	
	I	M	I	M
μ_G	**97.3**	98.2	**87.9**	90.6
μ_{Gc}	89.9	90.8	84.8	89.8
μ_{Sc}	90.7	93.0	86.4	92.2
Phonemes	/ə/ (329)		/a/ (378)	
	I	M	I	M
μ_G	**96.7**	97.1	**95.8**	95.7
μ_{Gc}	91.8	91.8	87.8	86.9
μ_{Sc}	91.2	90.9	90.7	89.3
Phonemes	/o/ (87)		/u/ (75)	
	I	M	I	M
μ_G	**97.7**	97.3	**85.3**	87.9
μ_{Gc}	87.4	88.3	73.3	77.4
μ_{Sc}	89.7	89.6	76.0	81.2
Phonemes	/ɑ̃/ (108)		/p/ (113)	
	I	M	I	M
μ_G	**97.2**	97.4	**91.2**	92.9
μ_{Gc}	89.8	89.4	82.3	84.3
μ_{Sc}	94.4	92.9	87.6	88.7
Phonemes	/t/ (196)		/k/ (102)	
	I	M	I	M
μ_G	**90.8**	92.3	**82.4**	86.0
μ_{Gc}	86.7	87.1	78.4	84.4
μ_{Sc}	89.3	90.4	78.4	81.0
Phonemes	/d/ (111)		/s/ (194)	
	I	M	I	M
μ_G	**97.3**	97.5	**87.1**	89.0
μ_{Gc}	91.0	91.9	78.9	80.4
μ_{Sc}	92.8	93.8	84.0	84.3
Phonemes	/v/ (42)		/ʒ/ (48)	
	I	M	I	M
μ_G	**90.5**	89.2	**83.3**	82.9
μ_{Gc}	76.2	77.0	77.1	76.0
μ_{Sc}	76.2	75.7	75.0	73.6
Phonemes	/m/ (158)		/n/ (149)	
	I	M	I	M
μ_G	**95.6**	96.4	**96.6**	94.4
μ_{Gc}	79.7	80.3	87.2	88.0
μ_{Sc}	95.6	94.5	94.6	94.8

Table 4. *Speaker identification with speech material selected from specific phoneme realisations. Training is composed of 15 s of phonetically balanced speech and test is composed of 1 s of phonetically biased speech. I = Global correct identification score, M = Average correct identification score over all speakers. The number of tests for each test configuration is indicated in parentheses.*

A FURTHER INVESTIGATION ON AR-VECTOR MODELS FOR TEXT-INDEPENDENT SPEAKER IDENTIFICATION

Ivan MAGRIN-CHAGNOLLEAU *Joachim WILKE* *Frédéric BIMBOT*

Télécom Paris (E.N.S.T.), Dépt. Signal – C.N.R.S., URA 820
46, rue Barrault – 75634 Paris cedex 13 – FRANCE – European Union
email: ivan@sig.enst.fr and bimbot@sig.enst.fr

ABSTRACT

In this paper, we investigate on the role of dynamic information on the performances of AR-vector models for speaker recognition. To this purpose, we design an experimental protocol that destroys the time structure of speech frame sequences, which we compare to a more conventional one, i.e. keeping the natural time order. These results are also compared with those obtained with a (single) Gaussian model. Several measures are systematically investigated in the three cases, and different ways of symmetrisation are tested. We observe that the destruction of the time order can be a factor of improvement for the AR-vector models, and that results obtained with the Gaussian model are merely always better. In most cases, symmetrisation is beneficial.

1. INTRODUCTION

Auto-Regressive (AR) Vector Models have been a significant subject of interest in the field of Speaker Recognition [1] [2] [3] [4] [5] [6] [7]. Whereas the idea of modeling a speaker by an AR-vector model estimated on sequences of speech frames is common to these works, the way to measure the similarity between two speaker models is addressed very differently. Secondly, the use of AR-vector model is often motivated by the belief that such an approach is an efficient way to extract dynamic speaker characteristics, as opposed to static characteristics such as the distribution of speech frame parameters.
In this paper we report on a systematic investigation on similarity measures between AR-vector speaker models obtained as simple combinations of canonical quantities. We also design a protocol in order to examine the role of dynamic information on the performance of the AR-vector approach : we destroy the natural time order of speech frames by shuffling them randomly, and we evaluate the AR-vector approach on these temporally disorganised data. We finally compare both previous approaches to a (single) Gaussian Model [8] [9] [10] [11].

2. DEFINITIONS AND NOTATION

Let $\{\mathbf{x}_t\}_{1\leq t\leq M}$ be a sequence of p-dimensional vectors. Let us define the centered vectors $\mathbf{x}_t^* = \mathbf{x}_t - \bar{\mathbf{x}}$ where $\bar{\mathbf{x}}$ is the mean vector of $\{\mathbf{x}_t\}$.
Let us denote $\mathcal{X}_0$ the covariance matrix of $\{\mathbf{x}_t\}$:

$$\mathcal{X}_0 = \frac{1}{M}\sum_{t=1}^{M}(\mathbf{x}_t - \bar{\mathbf{x}})\cdot(\mathbf{x}_t - \bar{\mathbf{x}})^T = \frac{1}{M}\sum_{t=1}^{M}\mathbf{x}_t^*\cdot\mathbf{x}_t^{*T}$$

We also define as $\mathcal{X}_k$ the lagged covariance matrices :

$$\mathcal{X}_k = \frac{1}{M}\sum_{t=k+1}^{M}\mathbf{x}_t^*\cdot\mathbf{x}_{t-k}^{*T} \text{ with } k = 1, ..., q$$

and the Toeplitz matrix X :

$$X = \begin{bmatrix} \mathcal{X}_0 & \mathcal{X}_1 & ... & \mathcal{X}_q \\ \mathcal{X}_1^T & \mathcal{X}_0 & ... & \mathcal{X}_{q-1} \\ \vdots & \vdots & & \vdots \\ \mathcal{X}_q^T & \mathcal{X}_{q-1}^T & ... & \mathcal{X}_0 \end{bmatrix}$$

A q-th order AR-vector model of sequence $\{\mathbf{x}_t^*\}$ is classically written as :

$$\sum_{i=0}^{q} A_i\cdot\mathbf{x}_{t-i}^* = \mathbf{e}_t \text{ with } A_0 = I_p$$

where $\{A_i\}$ is a set of $q+1$ matrix prediction coefficients, and $\mathbf{e}_t$ is the prediction error vector. $\{A_1, ..., A_q\}$ are obtained by solving the vector Yule-Walker equation [12]. With $A = [A_0 \ ... \ A_q]$, the covariance matrix of the residual of $\{\mathbf{x}_t^*\}$ filtered by A is :

$$E_X^{(A)} = AXA^T$$

Similarily, for a signal $\{\mathbf{y}_t\}_{1\leq t\leq N}$ with model B, we will denote :

$$E_Y^{(B)} = BYB^T$$

If we now consider :

$$\begin{aligned} E_X^{(B)} &= BXB^T \\ E_Y^{(A)} &= AYA^T \end{aligned}$$

these matrices can be interpreted as the covariance matrix of the filtering of $\{\mathbf{x}_t^*\}$ by B, and vice-versa. As A is obtained by minimising $tr(E_X^{(A)})$ and B by minimising $tr(E_Y^{(B)})$, we have $tr(E_X^{(B)}) \geq tr(E_X^{(A)})$ and $tr(E_Y^{(A)}) \geq tr(E_Y^{(B)})$.

Let us finally define $\Gamma_X^{(B/A)}$ and $\Gamma_{Y/X}^{(A)}$ as :

$$\begin{aligned} \Gamma_X^{(B/A)} &= \left(E_X^{(A)}\right)^{-\frac{1}{2}}\cdot E_X^{(B)}\cdot\left(E_X^{(A)}\right)^{-\frac{1}{2}} \\ \Gamma_{Y/X}^{(A)} &= \left(E_X^{(A)}\right)^{-\frac{1}{2}}\cdot E_Y^{(A)}\cdot\left(E_X^{(A)}\right)^{-\frac{1}{2}} \end{aligned}$$

function f	a	$\log a$	g	$\log g$	$a - \log g - 1$	$\log(a/g)$	$a - g$
AR-vector model - spectral frames in their natural time order							
$f_X^{(B/A)} \mid f_Y^{(A/B)}$	16.8 | 8.6	16.8 | 8.6	16.2 | 7.6	16.2 | 7.6	19.1 | 10.8	23.8 | 19.4	22.2 | 17.5
symmetrised	3.5 $^\bullet$	4.1 $^\bullet$	4.1 $^\bullet$	4.1 $^\bullet$	3.2 $^\bullet$	7.9 $^\bullet$	7.3 $^\bullet$
$f_{Y/X}^{(A)} \mid f_{X/Y}^{(B)}$	75.6 | 51.4	75.6 | 51.4	88.3 | 73.0	88.3 | 73.0	15.2 | 34.3	7.6 | 18.7	15.2 | 14.6
symmetrised	6.0 *	4.8 *	12.4 *	4.8 *	5.4 $^\diamond$	7.0 $^\diamond$	6.0 $^\diamond$
AR-vector model - spectral frames in a random time order							
$f_{X'}^{(B'/A')} \mid f_{Y'}^{(A'/B')}$	2.5 | 56.5	2.5 | 56.5	4.1 | 58.1	4.1 | 58.1	2.5 | 56.2	4.1 | 55.9	3.5 | 54.6
symmetrised	3.5 $^\diamond$	3.5 $^\diamond$	5.7 $^\diamond$	5.7 $^\diamond$	2.5 $^\diamond$	4.1 $^\diamond$	4.1 $^\diamond$
$f_{Y'/X'}^{(A')} \mid f_{X'/Y'}^{(B')}$	42.5 | 45.4	42.5 | 45.4	98.1 | 82.9	98.1 | 82.9	**1.3** | 22.9	**1.0** | 6.7	3.2 | 8.9
symmetrised	4.8 *	2.2 *	46.7 *	12.7 *	2.9 $^\diamond$	**1.0** $^\diamond$	**1.6** $^\diamond$
Gaussian model							
$f_{Y_o/X_o}^{(I)} \mid f_{X_o/Y_o}^{(I)}$	37.5 | 47.0	37.5 | 47.0	98.4 | 98.4	98.4 | 98.4	**0.6** | 7.9	**0.6** | 3.2	2.9 | 6.4
symmetrised	3.8 *	**1.3** *	97.1 *	99.4 *	**1.0** $^\diamond$	**0.6** $^\diamond$	**1.0** $^\diamond$

Table 1. TIMIT - Speaker identification error rates

where $E^{\frac{1}{2}}$ is the symmetric square root matrix of E.
The first matrix can be interpreted as the covariance matrix of $\{\mathbf{x}_t^*\}$ filtered by B relative to the one of $\{\mathbf{x}_t^*\}$ filtered by A, and the second one as the covariance matrix of $\{\mathbf{y}_t^*\}$ filtered by A relative to the one of $\{\mathbf{x}_t^*\}$ filtered by A.

3. SPEAKER MODELS

The purpose of this paper is to investigate on different ways of using an AR-vector model for speaker identification. A speaker is characterised by a second-order AR-vector model ($q = 2$) estimated on some speech material training. The matrix prediction coefficients $\{A_1, A_2\}$ are obtained by solving the vector Yule-Walker equation in the case $q = 2$:

$$[A_1\ A_2] \cdot \begin{bmatrix} \mathcal{X}_0 & \mathcal{X}_1 \\ \mathcal{X}_1^T & \mathcal{X}_0 \end{bmatrix} = -\left[\mathcal{X}_1^T\ \mathcal{X}_2^T\right]$$

- A first model is a 2nd-order AR-vector model trained on speech frames presented in their natural time order. Therefore, the model of $\mathcal{X}$ is $\{A, X\}$.
- A second model is a 2nd-order AR-vector model trained on the same speech frames as previously, but presented in a random time order. Each speaker $\mathcal{X}$ is characterised by $\{A', X'\}$ which are obtained in the same way as $\{A, X\}$, after speech frames have been randomly shuffled.

Gaussian speaker model is also tested as a reference model. In this second framework, a speaker $\mathcal{X}$ is represented by the covariance matrix $\mathcal{X}_0$. It is equivalent to a 0th-order AR-vector model, i.e. $A = [A_0] = I_p$ and $X = [\mathcal{X}_0]$, which we will denote as $\{I, X_0\}$.

4. SIMILARITY MEASURES

We consider now 2 speakers $\mathcal{X}$ and $\mathcal{Y}$, and we present a general formalism for expressing similarity measures between their AR-vector models.

Two families of similarity measures are investigated :

$$\begin{aligned} f_X^{(B/A)}(\mathcal{X}, \mathcal{Y}) &= f\left(\Gamma_X^{(B/A)}\right) \\ f_{Y/X}^{(A)}\ (\mathcal{X}, \mathcal{Y}) &= f\left(\Gamma_{Y/X}^{(A)}\right) \end{aligned}$$

The first family can be interpreted as a measure between two models (A and B), via their influence on the same vector signal (X). This family of measures (which we will refer to as VI), generalises the Itakura measure to the vector case [13]. Examples of such measures are proposed in [4] and [6]. On the opposite, the second family can be viewed as a measure between two signals (X and Y) filtered by a common model (A). Some of the IS measures proposed in [3] [5] belong to this family. Note also that setting $\{A, X\} = \{I, X_0\}$ allows to construct a similar family of measures for the Gaussian model.
The function f is chosen equal to a combination of the following canonical quantities :

$$\begin{aligned} a(\Gamma) &= \tfrac{1}{p}\, tr(\Gamma) \\ g(\Gamma) &= [det(\Gamma)]^{\frac{1}{p}} \end{aligned}$$

It can be shown that a and g are positive and that $a \geq g$. Moreover these quantities can be computed very efficiently [11]. The composed functions $a - \log g - 1$ and $\log(a/g)$ are respectively the Maximum-Likelihood measure [9] and the Arithmetic-Geometric Sphericity measure [8].
As these measures are not symmetric, different symmetrisations can be applied on the original measures. Given $f_X^{(B/A)}$ and $f_Y^{(A/B)}$, we define :

$$\begin{aligned} f_X^{(B/A)^\star} &= \frac{1}{2}\, f_X^{(B/A)} + \frac{1}{2}\, f_Y^{(A/B)} \\ f_X^{(B/A)^\diamond} &= \frac{\bar{M}}{\bar{M}+\bar{N}}\, f_X^{(B/A)} + \frac{\bar{N}}{\bar{M}+\bar{N}}\, f_Y^{(A/B)} \\ f_X^{(B/A)^\bullet} &= \frac{\bar{N}}{\bar{M}+\bar{N}}\, f_X^{(B/A)} + \frac{\bar{M}}{\bar{M}+\bar{N}}\, f_Y^{(A/B)} \end{aligned}$$

function f	a	$\log a$	g	$\log g$	$a - \log g - 1$	$\log(a/g)$	$a - g$
AR-vector model - spectral frames in their natural time order							
$f_X^{(B/A)} \mid f_Y^{(A/B)}$	38.7 | 30.2	38.7 | 30.2	37.1 | 29.5	37.1 | 29.5	42.5 | 35.2	51.1 | 50.8	49.5 | 49.5
symmetrised	24.8 $^\bullet$	25.1 $^\bullet$	24.8 $^\bullet$	24.4 $^\bullet$	26.3 $^\bullet$	35.6 $^\bullet$	33.3 $^\bullet$
$f_{Y/X}^{(A)} \mid f_{X/Y}^{(B)}$	93.3 | 86.0	93.3 | 86.0	96.5 | 94.6	96.5 | 94.6	44.1 | 69.8	41.6 | 39.1	49.2 | 39.1
symmetrised	23.5 *	21.3 *	32.4 *	25.4 *	24.4 $^\diamond$	34.6 $^\diamond$	33.0 $^\diamond$
AR-vector model - spectral frames in a random time order							
$f_{X'}^{(B'/A')} \mid f_{Y'}^{(A'/B')}$	35.9 | 82.2	35.9 | 82.2	36.8 | 81.3	36.8 | 81.3	32.4 | 83.5	34.6 | 82.2	34.3 | 81.6
symmetrised	39.1 $^\diamond$	39.1 $^\diamond$	40.0 $^\diamond$	40.0 $^\diamond$	34.3 $^\diamond$	33.3 $^\diamond$	33.3 $^\diamond$
$f_{Y'/X'}^{(A')} \mid f_{X'/Y'}^{(B')}$	78.7 | 71.4	78.7 | 71.4	98.4 | 93.7	98.4 | 93.7	15.9 | 43.8	**13.3** | 21.6	20.3 | 27.3
symmetrisation	21.9 *	**14.6** *	69.8 *	52.4 *	**14.0** $^\diamond$	**13.3** $^\diamond$	**14.3** $^\diamond$
Gaussian model							
$f_{Y_o/X_o}^{(I)} \mid f_{X_o/Y_o}^{(I)}$	77.1 | 71.8	77.1 | 71.8	98.4 | 98.4	98.4 | 98.4	**14.6** | 27.3	**12.7** | 17.1	20.3 | 21.3
symmetrised	15.6 *	**11.8** *	97.8 *	98.4 *	**12.7** $^\diamond$	**12.4** $^\diamond$	**14.3** $^\diamond$

Table 2. FTIMIT - Speaker identification error rates

$\bar{M}$ is the average number of frames for the training sentences across all speakers, and $\bar{N}$ is the average number of frames for the test sentences. The same symmetrisations are applied to $f_{Y/X}^{(A)}$ and $f_{X/Y}^{(B)}$.

5. DATABASE AND SIGNAL ANALYSIS

We use the first 63 speakers of TIMIT [14] and NTIMIT [15] for our experiments (19 females and 44 males)[1]. Each of them has read 10 sentences. The signal is sampled at 16 kHz, on 16 bits, on a linear amplitude scale. NTIMIT is a telephone-channel version of TIMIT.

Each sentence is analysed as follows : for each speech token, the speech signal is kept in its integrality; it is decomposed into frames of 31.5 ms at a frame rate of 10 ms, with no pre-emphasis. A Hamming window is applied to each frame. Then the module of a 504 point Fourier Transform is computed, from which 24 Mel-scale triangular filter bank coefficients are extracted. The spectral vectors $\{\mathbf{x}_t\}$ (of dimension $p = 24$) are formed from the logarithm of each filter output. These analysis conditions are identical to those used in [11].

For the TIMIT database, all 24 coefficients of $\{\mathbf{x}_t\}$ are kept. For NTIMIT, 24-dimensional vectors are also extracted, but we keep only the first 17 coefficients, which corresponds to the telephone bandwidth. Experiments are also made on "FTIMIT", obtained by taking the 17 first coefficients of the vectors $\{\mathbf{x}_t\}$ extracted from TIMIT.

6. EXPERIMENTS

A common training/test protocol is used for all the experiments. It is described in detail in [11] (as protocol "long-short"). Training material consists of 5 sentences (i.e $\approx$ 14.4 s) which are concatenated into a single reference per speaker. Tests are carried out on 5×1 sentence per speaker (i.e ≈ 3.2 s per sentence) which are tested separately. The total number of independent tests is therefore $63 \times 5 = 315$. The decision rule is the 1-nearest neighbour.

Results of the experiments are given by database (Tables 1 2 and 3). Performances are reported in terms of closed-set speaker identification error rates on the test set for the canonical measures and various combined measures in their asymmetric and their best symmetric form. For the symmetrised measures, a superscript indicates to which symmetrisation ($\star$, $\diamond$ or $\bullet$) does the result correspond.

7. DISCUSSION

The following observations can be made :

- Symmetrisation is generally a factor of improvement. However, the appropriate symmetrisation is difficult to predict. It depends on the type of asymmetric measure, and whether the data are in a natural or in a random time order.
- For each database (TIMIT, FTIMIT and NTIMIT), we have underlined the best 10 (or 11) measures. They are (almost) the same ones for all 3 databases. The best one is always obtained with the Gaussian Model.
- With spectral frames in their natural order, VI measures globally outperform IS measures in canonical forms, but the trend is inverted with composed forms.
- With spectral frames in a random order, symmetric composed IS measures outperform all other AR-vector measures, in spite of the loss of the dynamic spectral characteristics.

8. CONCLUSION

In our experiments, we did not succeed in obtaining better speaker identification results with an AR-vector model based measure than with a single Gaussian model classifier.

[1] More precisely, we have kept all female and male speakers of "train/dr1" and "test/dr1", the first female speaker of "train/dr2", and the first 13 male speakers of "train/dr2".

function f	a	$\log a$	g	$\log g$	$a - \log g - 1$	$\log(a/g)$	$a - g$
AR-vector model - spectral frames in their natural time order							
$f_X^{(B/A)} \mid f_Y^{(A/B)}$	71.8 | 54.6	71.8 | 54.6	67.3 | 54.3	67.3 | 54.3	78.1 | 58.4	83.8 | 69.5	82.9 | 67.9
symmetrised	51.8 $^\bullet$	52.1 $^\bullet$	50.5 $^\bullet$	50.2 $^\bullet$	57.5 $^\bullet$	66.0 $^\bullet$	65.1 $^\bullet$
$f_{Y/X}^{(A)} \mid f_{X/Y}^{(B)}$	96.8 | 92.4	96.8 | 92.4	97.1 | 95.6	97.1 | 95.6	67.3 | 88.9	66.0 | 78.7	75.2 | 76.8
symmetrised	61.9 *	56.5 *	68.3 *	53.0 *	59.7 $^\diamond$	63.2 $^\diamond$	66.4 $^\diamond$
AR-vector model - spectral frames in a random time order							
$f_{X'}^{(B'/A')} \mid f_{Y'}^{(A'/B')}$	64.4 | 92.1	64.1 | 92.1	65.4 | 91.8	65.4 | 91.8	61.9 | 92.4	64.8 | 93.3	64.4 | 93.0
symmetrised	65.4 $^\diamond$	65.1 $^\diamond$	67.9 $^\diamond$	68.3 $^\diamond$	62.2 $^\diamond$	64.4 $^\diamond$	64.1 $^\diamond$
$f_{Y'/X'}^{(A')} \mid f_{X'/Y'}^{(B')}$	94.0 | 94.3	94.0 | 94.3	98.4 | 97.5	98.4 | 97.5	**47.0** | 86.4	**46.0** | 63.2	56.8 | 77.1
symmetrisation	61.9 *	52.4 *	88.3 *	72.4 *	50.2 $^\diamond$	**44.1** $^\diamond$	**48.6** $^\diamond$
Gaussian model							
$f_{Y_o/X_o}^{(I)} \mid f_{X_o/Y_o}^{(I)}$	93.0 | 94.6	93.0 | 94.6	98.4 | 98.4	98.4 | 98.4	**44.1** | 75.9	**42.5** | 59.7	56.2 | 73.3
symmetrised	58.1 *	**49.8** *	97.8 *	98.4 *	**47.6** $^\diamond$	**44.1** $^\diamond$	**49.2** $^\diamond$

Table 3. NTIMIT - Speaker identification error rates

This observation is in contradiction with results reported in [7], but this divergence may be due to different signal pre-processing and analysis.
Moreover, we globally obtained better performances with the AR-vector model on spectral frames in a random time order rather than when we kept the natural time order. Therefore, the role of dynamic speaker characteristics in the success of the AR-vector model can be questioned, as our results suggest that AR-vector models tend to extract indirectly speaker characteristics of a static nature.
Finally, the influence of symmetrisation can be crucial, but its theoretical basis remains to be understood.

REFERENCES

[1] Yves Grenier. Utilisation de la prédiction linéaire en reconnaissance et adaptation au locuteur. In *XIèmes Journées d'Etude sur la Parole*, pages 163–171, May 1980. Strasbourg, France.

[2] T. Artières, Y. Bennani, P. Gallinari, and C. Montacié. Connectionnist and conventional models for text-free talker identification tasks. In *Proceedings of NEURONIMES 91*, 1991. Nîmes, France.

[3] C. Montacié, P. Deléglise, F. Bimbot, and M.-J. Caraty. Cinematic techniques for speech processing: temporal decomposition and multivariate linear prediction. In *Proceedings of ICASSP 92*, volume 1, pages 153–156, March 1992. San Francisco, United-States.

[4] F. Bimbot, L. Mathan, A. de Lima, and G. Chollet. Standard and target-driven AR-vector models for speech analysis and speaker recognition. In *Proceedings of ICASSP 92*, volume 2, pages II.5–II.8, March 1992. San Francisco, United-States.

[5] Claude Montacié and Jean-Luc Le Floch. AR-vector models for free-text speaker recognition. In *Proceedings of ICSLP 92*, volume 1, pages 611–614, October 1992. Banff, Canada.

[6] Chintana Griffin, Tomoko Matsui, and Sadoaki Furui. Distance measures for text-independent speaker recognition based on MAR model. In *Proceedings of ICASSP 94*, volume 1, pages 309–312, April 1994. Adelaïde, Australia.

[7] J.-L. Le Floch, C. Montacié, and M.-J. Caraty. Speaker recognition experiments on the NTIMIT database. In *Proceedings of EUROSPEECH 95*, volume 1, pages 379–382, September 1995. Madrid, Spain.

[8] Yves Grenier. *Identification du locuteur et adaptation au locuteur d'un système de reconnaissance phonémique*. PhD thesis, ENST, 1977.

[9] Herbert Gish, Michael Krasner, William Russell, and Jared Wolf. Methods and experiments for text-independent speaker recognition over telephone channels. In *Proceedings of ICASSP 86*, volume 2, pages 865–868, April 1986. Tokyo, Japan.

[10] Frédéric Bimbot and Luc Mathan. Text-free speaker recognition using an arithmetic-harmonic sphericity measure. In *Proceedings of EUROSPEECH 93*, volume 1, pages 169–172, September 1993. Berlin, Germany.

[11] Frédéric Bimbot, Ivan Magrin-Chagnolleau, and Luc Mathan. Second-order statistical measures for text-independent speaker identification. *Speech Communication*, 17(1–2):177–192, August 1995.

[12] P. Whittle. On the fitting of multivariate autoregressions, and the approximate canonical factorization of a spectral density matrix. *Biometrika*, 50(1–2):129–134, 1963.

[13] Fumitada Itakura. Minimum prediction residual principle applied to speech recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 23(1):67–72, February 1975.

[14] William M. Fisher, George R. Doddington, and Kathleen M. Goudie-Marshall. The DARPA speech recognition research database : specifications and status. In *Proceedings of the DARPA workshop on speech recognition*, pages 93–99, February 1986.

[15] Charles Jankowski, Ashok Kalyanswamy, Sara Basson, and Judith Spitz. NTIMIT: a phonetically balanced, continuous speech, telephone bandwidth speech database. *Proceedings of ICASSP 90*, April 1990. New Mexico, United-States.

Bibliographie

[1] T. W. Anderson. *An Introduction to Multivariate Analysis*. John Wiley and Sons, 1958.

[2] T. Artières, Y. Bennani, P. Gallinari & C. Montacié. Connectionnist and conventional models for text-free talker identification tasks. In *Proceedings of NEURONIMES 91*. 1991. Nîmes, France.

[3] T. Artières & P. Gallinari. Neural models for extracting speaker characteristics in speech modelization systems. In *Proceedings of EUROSPEECH 93*, pp. 2263–2266. 1993. Berlin, Germany.

[4] T. Artières & P. Gallinari. Approches prédictives neuronales pour l'identification. In *XXèmes Journées d'Etude sur la Parole*, pp. 275–280. Jun. 1994. Trégastel, France.

[5] Bishnu S. Atal. Automatic recognition of speakers from their voices. *Proceedings of the IEEE*, vol. 64, no. 4 : pp. 460–475, Apr. 1976.

[6] B.S. Atal. Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. *Journal of the Acoustical Society of America*, vol. 55, no. 6 : pp. 1304–1312, Jun. 1974.

[7] J. Baras & P. Rajasekaran. Robustness study of free-text speaker identification and verification. In *Proceedings of ICASSP 93*. 1993.

[8] Younès Bennani. Speaker identification through a modular connectionist architecture : evaluation on the TIMIT database. In *Proceedings of ICSLP 92*, vol. 1, pp. 607–610. Oct. 1992. Banff, Canada.

[9] Younès Bennani & Patrick Gallinari. On the use of TDNN-extracted features information in talker identification. In *Proceedings of ICASSP 91*, vol. 1, pp. 385–388. May 1991. Toronto, Canada.

[10] Younès Bennani & Patrick Gallinari. Neural networks for discrimination and modelization of speakers. *Speech Communication*, vol. 17, no. 1–2 : pp. 159–175, Aug. 1995.

[11] L. Besacier & J.-F. Bonastre. Subband approach for automatic-speaker recognition : optimal division of the frequency domain. In *Workshop on Audio and Video Biometric Person Authentication*. 1997. Crans-Montana, Switzerland.

[12] F. Bimbot, L. Mathan, A. de Lima & G. Chollet. Standard and target-driven AR-vector models for speech analysis and speaker recognition. In *Proceedings of ICASSP 92*, vol. 2, pp. II.5–II.8. Mar. 1992. San Francisco, United-States.

[13] Frédéric Bimbot, Enrico Bocchieri & Bishnu Atal. Sous-espaces de projection de séquences de trames acoustiques pour l'analyse et la reconnaissance de parole. In *XXIèmes Journées d'Etude sur la Parole*. 1996. Avignon, France.

[14] Frédéric Bimbot, Gérard Chollet & Andrea Paoloni. Assessment methodology for speaker identification and verification systems. In *Workshop on automatic speaker recognition identification verification*, pp. 75–82. Apr. 1994. Martigny, Switzerland.

[15] Frédéric Bimbot, Ivan Magrin-Chagnolleau & Luc Mathan. Second-order statistical measures for text-independent speaker identification. *Speech Communication*, vol. 17, no. 1–2 : pp. 177–192, Aug. 1995.

[16] Frédéric Bimbot & Luc Mathan. Text-free speaker recognition using an arithmetic-harmonic sphericity measure. In *Proceedings of EUROSPEECH 93*, vol. 1, pp. 169–172. Sep. 1993. Berlin, Germany.

[17] Frédéric Bimbot, Andrea Paoloni & Gérard Chollet. *Assessment Methodology for Speaker Identification and Verification Systems.* Technical report – Task 2500 – Report I9, SAM-A ESPRIT Project 6819, 1993.

[18] R. Bolt, F. Cooper, E. David, P. Denes, J. Pickett & K. Stevens. Speaker identification by speech spectrograms : some further observations. *Journal of the Acoustical Society of America*, vol. 54 : pp. 531–534, 1973.

[19] Jean-François Bonastre. *Stratégie analytique orientée connaissances pour la caractérisation et l'identification du locuteur*. Ph.D. thesis, Université d'Avignon et des pays de Vaucluse, Jan. 1994.

[20] Jean-François Bonastre & Henri Méloni. Automatic speaker recognition and analytic process. In *Proceedings of EUROSPEECH 93*, vol. 1, pp. 441–444. Sep. 1993. Berlin, Germany.

[21] Jean-François Bonastre & Henri Méloni. Avantages d'une approche analytique orientée connaissances en reconnaissance du locuteur. In *XXèmes Journées d'Etude sur la Parole*, pp. 259–264. Jun. 1994. Trégastel, France.

[22] Ian Booth, Michael Barlow & Brett Watson. Enhancements to DTW and VQ decision algorithms for speaker recognition. *Speech Communication*, vol. 13, no. 3–4 : pp. 427–433, Dec. 1993.

[23] Peter D. Bricker & Sandra Pruzansky. Effects of stimulus content and duration on talker identification. *Journal of the Acoustical Society of America*, vol. 40, no. 6 : pp. 1441–1449, 1966.

[24] Joseph P. Campbell Jr. Testing with the Yoho cd-rom voice verification corpus. In *Proceedings of ICASSP 95*, vol. 1, pp. 341–344. 1995. Detroit, United-States.

[25] Chi Wei Che, Qiguang Lin & Dong-Suk Yuk. An HMM approach to text-prompted speaker verification. In *Proceedings of ICASSP 96*, vol. 2, pp. 673–676. May 1996. Atlanta, Georgia, United-States.

[26] R.S. Cheung & B.A. Eisenstein. Feature selection via dynamic programming for text independent speaker identification. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 26 : pp. 397–403, Oct. 1978.

[27] G. Chollet. About the ethics of speaker identification. In *Proceedings of ICPhS 91*, vol. 1, p. 397. 1991. Aix-en-Provence, France.

[28] G. Chollet & F. Bimbot. *Handbook of Standards and Resources for Spoken Language Systems*, chap. Assessment of speaker verification systems. Mouton de Gruyter, 1997. Disponible au printemps 97.

[29] F. Clarke & R. Becker. Comparison of techniques for discriminating among talkers. *Journal of Speech and Hearing Research*, vol. 12 : pp. 747–761, 1969.

[30] J.M. Colombi, D.W. Ruck, S.K. Rogers, M. Oxley & T.R. Anderson. Cohort selection and word grammar effects for speaker recognition. In *Proceedings of ICASSP 96*, vol. 1, pp. 85–88. May 1996. Atlanta, Georgia, United-States.

[31] P. Combescure. Vingt listes de dix phrases françaises phonétiquement équilibrées. Note technique, CNET (Centre Lannion A), Oct. 1980.

[32] A. Compton. Effects of filtering and vocal duration upon the identification of speakers, aurally. *Journal of the acoustical Society of America*, vol. 35, no. 11 : pp. 1748–1752, 1963.

[33] David J. Darlington & Douglas R. Campbell. Sub-band adaptive filtering applied to speech enhancement. In *Proceedings of ICSLP 96*. 1996.

[34] Georges R. Doddington. Speaker recognition, identifying people by their voices. *Proceedings of the IEEE*, vol. 73, no. 11 : pp. 1651–1664, November 1985.

[35] J. Eatock & J.S. Mason. Speaker-dependent speech classification in speaker recognition. In *ESCA Workshop on Speaker Characterisation in Speech Technology*, pp. 94–97. Jun. 1990. Edinburgh, SCOTLAND.

[36] J. P. Eatock & J. S. Mason. A quantitative assessment of the relative speaker discriminating properties of phonemes. In *Proceedings of ICASSP 94*, vol. 1, pp. 133–136. Apr. 1994. Adelaïde, Australia.

[37] J. M. Elvira & R. A. Carrasco. Neural networks for speech and speaker recognition through a digital telephone exchange. In *Proceedings of EUROSPEECH 93*, vol. 3, pp. 2291–2294. Sep. 1993. Berlin, Germany.

[38] Nikos Fakotakis & John Sirigos. A high performance text independent speaker recognition system based on vowel spotting and neural nets. In *Proceedings of ICASSP 96*, vol. 2, pp. 661–664. May 1996. Atlanta, Georgia, United-States.

[39] K. R. Farrell, R. J. Mammone & K. T. Assaleh. Speaker recognition using neural networks and conventional classifiers. *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 1 : pp. 194–205, Jan. 1994.

[40] Kevin Farrell. *Speaker recognition using the modified neural tree network*. Ph.D. thesis, University of New Jersey, Oct. 1993.

[41] Kevin R. Farrell & Richard J. Mammone. Speaker identification using neural tree networks. In *Proceedings of ICASSP 94*, vol. 1, pp. 165–168. Apr. 1994. Adelaïde, Australia.

[42] William M. Fisher, George R. Doddington & Kathleen M. Goudie-Marshall. The DARPA speech recognition research data-

base : specifications and status. In *Proceedings of the DARPA workshop on speech recognition*, pp. 93–99. Feb. 1986.

[43] S. Furui. Research on individuality features in speech waves and automatic speaker recognition techniques. *Speech Communication*, vol. 5 : pp. 183–197, 1986.

[44] S. Furui, F. Itakura & S. Saito. Talker recognition by long-time averaged speech spectrum. *Elect. Commun. Japan*, vol. 55-A, no. 10 : pp. 54–61, 1972.

[45] Sadaoki Furui. Cepstral analysis techniques for automatic speaker verification. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 29, no. 2 : pp. 254–272, Aug. 1981.

[46] Sadaoki Furui. Comparison of speaker recognition methods using static features and dynamic features. *IEEE Transactions on Acoustics, Speech, and Signal processing*, vol. 29, no. 3 : pp. 342–350, Jun. 1981.

[47] Sadoaki Furui. An overview of speaker recognition technology. In *Workshop on automatic speaker recognition, identification and verification*, pp. 1–9. Apr. 1994. Martigny, Switzerland.

[48] John S. Garofolo. The structure and format of the DARPA TIMIT CD-ROM prototype. Documentation with a CD-ROM, 1988.

[49] Philippe Gilles. *Décodage phonétique de la parole et adaptation au locuteur*. Ph.D. thesis, Université d'Avignon et des Pays de Vaucluse, Jan. 1993.

[50] Herbert Gish. Robust discrimination in automatic speaker identification. In *Proceedings of ICASSP 90*, vol. 1, pp. 289–292. Apr. 1990. New Mexico, United-States.

[51] Herbert Gish, Michael Krasner, William Russell & Jared Wolf. Methods and experiments for text-independent speaker recognition over telephone channels. In *Proceedings of ICASSP 86*, vol. 2, pp. 865–868. Apr. 1986. Tokyo, Japan.

[52] J. Godfrey, E. Holliman & J. MacDaniel. Switchboard : telephone speech corpus for research and development. In *Proceedings of ICASSP 92*, vol. 1, pp. 517–520. 1992.

[53] U. Goldstein. Speaker-identifying features based on format tracks. *Journal of the Acoustical Society of America*, vol. 59, no. 1 : pp. 176–182, 1975.

[54] Robert M. Gray. Vector quantization. In Alex Waibel & Kai-Fu Lee, editors, *Readings in Speech Recognition*, pp. 75–100. Morgan

Kaufmann Publishers, 1990. IEEE ASSP Magazine, Vol. 1, No. 2, pp. 4–29, April 1984.

[55] Yves Grenier. *Identification du locuteur et adaptation au locuteur d'un système de reconnaissance phonémique*. Ph.D. thesis, ENST, 1977.

[56] Yves Grenier. Utilisation de la prédiction linéaire en reconnaissance et adaptation au locuteur. In *XIèmes Journées d'Etude sur la Parole*, pp. 163–171. May 1980. Strasbourg, France.

[57] Chintana Griffin, Tomoko Matsui & Sadoaki Furui. Distance measures for text-independent speaker recognition based on MAR model. In *Proceedings of ICASSP 94*, vol. 1, pp. 309–312. Apr. 1994. Adelaïde, Australia.

[58] M. I. Hannah, A. T. Sapeluk, R. I. Dampert & I. M. Roger. The effect of utterance length and content on speaker-verifier performance. In *Proceedings of EUROSPEECH 93*, vol. 3, pp. 2299–2302. Sep. 1993. Berlin, Germany.

[59] Hiroaki Hattori. Text-independent speaker recognition using neural networks. In *Proceedings of ICASSP 92*, vol. 2, pp. 153–156. Mar. 1992. San Francisco, United-States.

[60] Hiroaki Hattori. Text-independent speaker recognition using neural networks. *IEICE Trans. Inf. and Syst.*, vol. E76–D, no. 3 : pp. 345–351, Mar. 1993.

[61] Hiroaki Hattori. Text-independent speaker verification using neural networks. In *Workshop on automatic speaker recognition identification verification*, pp. 103–106. Apr. 1994. Martigny, Switzerland.

[62] M. Hecker. Speaker recognition : an interpretive survey of the litterature. Monography, 1971.

[63] H. Hermansky. Perceptual linear predictive (plp) analysis of speech. *Journal of the Acoustical Society of America*, vol. 87 : pp. 1738–1752, 1990.

[64] Hynek Hermansky & Nelson Morgan. RASTA processing of speech. *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4 : pp. 578–589, october 1994.

[65] Hynek Hermansky & Nelson Morgan. RASTA processing of speech. *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4 : pp. 578–589, 1994.

[66] J.A. Hernandez-Mendez & A.R. Figueiras-Vidal. Measuring similarities among speakers by means of neural networks. In *Proceedings of EUROSPEECH 93*, vol. 1, pp. 643–646. Sep. 1993. Berlin, Germany.

[67] Norio Higuchi & Makoto Hashimoto. analysis of acoustic features affecting speaker identification. In *Proceedings of eurospeech 95*, vol. 1, pp. 435–438. 1995.

[68] Norio Higuchi & Makoto Hashimoto. analysis of acoustic features affecting speaker identification. *Journal of the acoustical society of Japan*, vol. 17, no. 1 : pp. 33–35, 1996.

[69] G. Holmgren. Physical and psychological correlates of speaker recognition. *Journal of Speech and Hearing Research*, vol. 10 : pp. 57–66, 1967.

[70] Mehdi Homayounpour & Gérard Chollet. Neural net approaches to speaker verification : comparison with second-order statistic measures. *Proceedings of ICASSP 95*, vol. 1 : pp. 353–356, 1995. Detroit, United-States.

[71] Fumitada Itakura. Minimum prediction residual principle applied to speech recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 23, no. 1 : pp. 67–72, Feb. 1975.

[72] D. James, H.-P. Hutter & F. Bimbot. CAVE - Speaker verification in banking and telecommunications. In *Workshop on Audio- and Video-Based Biometric Person Authentification*. 1997. To Appear.

[73] Charles Jankowski, Ashok Kalyanswamy, Sara Basson & Judith Spitz. NTIMIT : a phonetically balanced, continuous speech, telephone bandwidth speech database. *Proceedings of ICASSP 90*, Apr. 1990. New Mexico, United-States.

[74] C.R. Jankowski Jr., T.F. Quatieri & D.A. Reynolds. Measuring fine structure in speech : application to speaker identification. In *Proceedings of ICASSP 95*, vol. 1, pp. 325–328. 1995. Detroit, United-States.

[75] C.R. Jankowski Jr., T.F. Quatieri & D.A. Reynolds. Fine structure features for speaker identification. In *Proceedings of ICASSP 96*, vol. 2, pp. 689–692. May 1996. Atlanta, Georgia, United-States.

[76] I.T. Jolliffe. *Principal Component Analysis*. Springer-Verlag, 1986.

[77] L. Kersta. Voiceprint identification. *Nature*, vol. 196 : pp. 1253–1257, 1962.

[78] T. Kitamura, E. Hayahara & K. Hatayama. Speaker recognition using dynamic features of speech and a neural network. In *Proceedings of ICSP 90*, vol. 1, pp. 461–464. Oct. 1990. Beijing, China.

[79] Tadashi Kitamura & Shinsai Takei. Speaker recognition model using two-dimensional mel-cepstrum and predictive neural network. In *Proceedings of ICSLP 96*. 1996.

[80] Hermann J. Künzel. Current approaches to forensic speaker recognition. In *Workshop on automatic speaker recognition identification verification*, pp. 135–141. Apr. 1994. Martigny, Switzerland.

[81] P. Ladefoged & R. Vanderslice. The "voiceprint" mystique. In *Working papers in Phonetics*, pp. 126–142. 1967.

[82] Lori F. Lamel, Robert H. Kassel & Stephanie Seneff. Speech database development : design and analysis of the acoustic-phonetic corpus. Documentation with a CD-ROM, 1988.

[83] J.-L. Le Floch, C. Montacié & M.-J. Caraty. Speaker recognition experiments on the NTIMIT database. In *Proceedings of EUROSPEECH 95*, vol. 1, pp. 379–382. Sep. 1995. Madrid, Spain.

[84] J.-L. Le Floch, C. Montacié & M.-J. Caraty. GMM and ARVM cooperation and competition for text-independent speaker recognition on telephone speech. In *Proceedings of ICSLP 96*. 1996.

[85] Jean-Luc Le Floch, Claude Montacié & Marie-José Caraty. Investigations on speaker characterization from Orphée system technics. In *Proceedings of ICASSP 94*, vol. 1, pp. 149–152. Apr. 1994. Adelaïde, Australia.

[86] K.P. Li & G.W. Hughes. Talker differences as they appear in correlation matrices of continuous speech spectra. *Journal of the Acoustical Society of America*, vol. 55, no. 4 : pp. 833–837, Apr. 1974.

[87] K.P. Li, G.W. Hughes & A.S. House. Correlation characteristics and dimensionality of speech spectra. *Journal of the Acoustical Society of America*, vol. 46, no. 4 : pp. 1019–1025, 1969.

[88] Han-Sheng Liou & Richard J. Mammone. A subword neural tree network approach to text-independent speaker verification. In *Proceedings of ICASSP 95*, vol. 1, pp. 357–360. 1995. Detroit, United-States.

[89] Chi-Shi Liu. A general framework of feature extraction : application to speaker recognition. In *Proceedings of ICASSP 96*, vol. 2, pp. 669–672. May 1996. Atlanta, Georgia, United-States.

[90] Li Liu, Jialong He & Günther Palm. Signal modeling for speaker identification. In *Proceedings of ICASSP 96*, vol. 2, pp. 665–668. May 1996. Atlanta, Georgia, United-States.

[91] Ivan Magrin-Chagnolleau, Jean-François Bonastre & Frédéric Bimbot. Effect of utterance duration and phonetic content on speaker identification using second-order statistical methods. In *Proceedings of EUROSPEECH 95*, vol. 1, pp. 337–340. Sep. 1995. Madrid, Spain.

[92] Ivan Magrin-Chagnolleau, Joachim Wilke & Frédéric Bimbot. A further investigation on ar-vector models for text-independent speaker identification. In *Proceedings of ICASSP 96*, vol. 1, pp. 401–404. May 1996. Atlanta, United-States.

[93] Richard J. Mammone, Xiaoyu Zhang & Ravi P. Ramachandran. Robust speaker recognition. *IEEE Signal Processing Magazine*, pp. 58–71, Sep. 1996.

[94] J. Markel, B. Oshika & H. Gray. Long-term feature averaging for speaker recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 25, no. 4 : pp. 330–337, 1977.

[95] John D. Markel & Steven B. Davis. Text-independent speaker recognition from a large linguistically unconstrained time-spaced data base. *IEEE Transactions on Acoutics, Speech and Signal Processing*, vol. 27, no. 1 : pp. 74–82, Feb. 1979.

[96] J.S. Mason, J. Oglesby & L. Xu. Codebooks to optimise speaker recognition. In *Proceedings of EUROSPEECH 89*, pp. 267–270. 1989.

[97] Tomoko Matsui & Sadaoki Furui. A text-independent speaker recognition method robust against utterance variations. In *Proceedings of ICASSP 91*, vol. 1, pp. 377–380. May 1991. Toronto, Canada.

[98] Tomoko Matsui & Sadaoki Furui. Comparaison of text-independent speaker recognition methods using VQ-distortion and discrete/continuous HMM's. *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 3 : pp. 456–459, Jul. 1994.

[99] Tomoko Matsui & Sadoaki Furui. Comparison of text-independent speaker recognition methods using VQ-distorsion and discrete/continuous HMMs. In *Proceedings of ICASSP 92*, vol. 2, pp. 157–160. Mar. 1992. San Francisco, United-States.

[100] F. McGehee. The reliability of the identification of human voice. *Journal of General Psychology*, vol. 17 : pp. 249–271, 1937.

[101] F. McGehee. An experimental study in voice recognition. *Journal of General Psychology*, vol. 31 : pp. 53–65, 1944.

[102] Henri Méloni & Philippe Gilles. Décodage acoustico-phonétique ascendant. *Traitement du Signal*, vol. 8, no. 2 : pp. 107–114, 1991.

[103] Ben Milner. Inclusion of temporal information into features for speech recognition. In *Proceedings of ICSLP 96*, vol. 1. 1996.

[104] Ben P. Milner & Saeed V. Vaseghi. An analysis of cepstral-time matrices for noise and channel robust speech recognition. In *Proceedings of EUROSPEECH 95*, vol. 1, pp. 519–522. 1995.

[105] C. Montacié, P. Deléglise, F. Bimbot & M.-J. Caraty. Cinematic techniques for speech processing : temporal decomposition and multivariate linear prediction. In *Proceedings of ICASSP 92*, vol. 1, pp. 153–156. Mar. 1992. San Francisco, United-States.

[106] Claude Montacié & Jean-Luc Le Floch. AR-vector models for free-text speaker recognition. In *Proceedings of ICSLP 92*, vol. 1, pp. 611–614. Oct. 1992. Banff, Canada.

[107] Climent Nadeu, José B. Marino, Javier Hernando & Albino Nogueiras. Frequency and time filtering of filter-bank energies for HMM speech recognition. In *Proceedings of ICSLP 96*. 1996.

[108] D. Naik, K. Assaleh & R. Mammone. Robust speaker identification using pole filtering. In *ESCA Workshop on Automatic Speaker Recognition Identification and Verification*. 1994. Martigny, Switzerland.

[109] Jayant M. Naik & David M. Lubensky. A hybrid HMM-MLP speaker verification algorithm for telephone speech. In *Proceedings of ICASSP 94*, vol. 1, pp. 153–156. Apr. 1994. Adelaïde, Australia.

[110] J. Oglesby. *Neural models for speaker recognition*. Ph.D. thesis, University College of Swansea, Wales, UK, Mar. 1991.

[111] J. Oglesby & J. S. Mason. Optimisation of neural models for speaker identification. In *Proceedings of ICASSP 90*, vol. 1, pp. 261–264. Apr. 1990. New Mexico, United-States.

[112] J. P. Openshaw, Z. P. Sun & J. S. Mason. A comparison of feature performance under degraded speech in speaker recognition. In *Proceedings of the ESCA workshop on speech processing in adverse conditions*, pp. 119–122. Nov. 1992. Cannes, France.

[113] J. P. Openshaw, Z. P. Sun & J. S. Mason. A comparison of composite feature under degraded speech in speaker recognition. In

Proceedings of ICASSP 93, vol. 2, pp. 371–374. Apr. 1993. Minneapolis, United-States.

[114] J.P. Openshaw & J.S. Mason. Noise robust estimate of speech dynamics for speaker recognition. In *Proceedings of ICSLP 96*. 1996.

[115] A.V. Oppenheim & R.W. Schafer. Homomorphic analysis of speech. *IEEE Transactions on Audio and Electroacoustics*, vol. 16 : pp. 221–226, 1968.

[116] Javier Ortega-Garcia & Joaquin Gonzalez-Rodriguez. Overview of speech enhancement techniques for automatic speaker recognition. In *Proceedings of ICSLP 96*. 1996.

[117] Douglas O'Shaughnessy. Speaker recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 3, no. 1 : pp. 4–17, October 1986.

[118] Douglas O'Shaughnessy. *Speech Communication : Human and Machine*. Addison-Wesley, 1987.

[119] I. Pollack, J.M. Pickett & W.H. Sumby. On the identification of speakers by voice. *Journal of the Acoustical Society of America*, vol. 26, no. 3 : pp. 403–406, May 1954.

[120] Alan B. Poritz. Linear predictive Hidden Markov Models and the speech signal. In *Proceedings of ICASSP 82*, pp. 1291–1294. May 1982. Paris, France.

[121] L. R. Rabiner & B. H. Juang. An introduction to hidden markov models. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, pp. 4–16, January 1986.

[122] Lawrence R. Rabiner. Mathematical foundations of hidden markov models. In H. Niemann et al., editor, *Recent Advances in Speech Understanding and Dialog System*, pp. 183–205. Springer-Verlag, 1988.

[123] Lawrence R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, vol. 77, no. 2 : pp. 257–285, Feb. 1989.

[124] L.R. Rabiner & B.-H. Juang. *Fundamentals of Speech Recognition*. Prentice Hall, 1993.

[125] Ravi P. Ramachandran, Mihailo S. Zilovic & Richard J. Mammone. A comparative study of robust linear predictive analysis methods with applications to speaker identification. *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 2 : pp. 117–125, Mar. 1995.

[126] Wang Ren-Hua, He Lin-Shen & Hiroya Fujisaki. A weighted distance measure based on the fine structure of feature space : Application to speaker recognition. In *Proceedings of ICASSP 90*, vol. 1, pp. 273–276. Apr. 1990. New Mexico, United-States.

[127] D.A. Reynolds, M.A. Zissman, T.F. Quatieri, G.C. O'Leary & B.A. Carlson. The effects of telephone transmission degradations on speaker recognition performance. In *Proceedings of ICASSP 95*, vol. 1, pp. 329–332. 1995. Detroit, United-States.

[128] Douglas A. Reynolds. *A Gaussian Mixture Modeling Approach to Text-Independent Speaker Identification*. Ph.D. thesis, Georgia Institute of Technology, Aug. 1992.

[129] Douglas A. Reynolds. Speaker identification and verification using Gaussian Mixture speaker models. In *Workshop on automatic speaker recognition, identification and verification*, pp. 27–30. Apr. 1994. Martigny, Switzerland.

[130] Douglas A. Reynolds. Speaker identification and verification using Gaussian mixture speaker models. *Speech Communication*, vol. 17, no. 1–2 : pp. 91–108, Aug. 1995.

[131] Douglas A. Reynolds. The effects of handset variability on speaker recognition performance : experiments on the switchboard corpus. In *Proceedings of ICASSP 96*, vol. 1, pp. 113–116. May 1996. Atlanta, Georgia, United-States.

[132] Douglas A. Reynolds & Richard C. Rose. Robust text-independent speaker identification using gaussian mixture speaker models. *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 1 : pp. 72–83, Jan. 1995.

[133] R. Rose, E. Hofstetter & D. Reynolds. Integrated models of signal and background with application to speaker identification in noise. *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 2 : pp. 245–257, 1994.

[134] R. C. Rose, J. Fitzmaurice, E. M. Hofstetter & D. A. Reynolds. Robust speaker identification in noisy environments using noise adaptive speaker models. In *Proceedings of ICASSP 91*, vol. 1, pp. 401–404. May 1991. Toronto, Canada.

[135] Richard C. Rose & Douglas A. Reynolds. Text independent speaker identification using automatic acoustic segmentation. In *Proceedings of ICASSP 90*, vol. 1, pp. 293–296. Apr. 1990. New Mexico, United-States.

[136] Aaron E. Rosenberg. Listener performance in speaker verification tasks. *IEEE Transactions on Audio and Electroacoustics*, vol. 21, no. 3 : pp. 221–225, Jun. 1973.

[137] Aaron E. Rosenberg. Automatic speaker verification : a review. *Proceedins of the IEEE*, vol. 64, no. 4 : pp. 475–487, Apr. 1976.

[138] Aaron E. Rosenberg, Chin-Hui Lee & Sedat Gokcen. Connected word talker verification using whole word hidden markov models. In *Proceedings of ICASSP 91*, vol. 1, pp. 381–384. May 1991. Toronto, Canada.

[139] Aaron E. Rosenberg, Chin-Hui Lee & Frank K. Soong. Subword unit talker verification using hidden markov models. In *Proceedings of ICASSP 90*, vol. 1, pp. 269–272. Apr. 1990. Albuquerque, New Mexico, United-States.

[140] Aaron E. Rosenberg & Frank K. Soong. *Recent Research in Automatic Speaker Recognition (in Advances in Speech Signal Processing)*, chap. 22, pp. 701–738. Marcel-Dekker, 1991.

[141] A.E. Rosenberg & F.K. Soong. Evaluation of a vector quantization talker recognition system in text independent and text dependent modes. In *Proceedings of ICASSP 86*, pp. 873–876. 1986.

[142] A.E. Rosenberg & F.K. Soong. Evaluation of a vector quantization talker recognition system in text independent and text dependent modes. *Computer Speech and Language*, vol. 22 : pp. 143–157, 1987.

[143] Laszlo Rudasi & Stephen A. Zahorian. Text-independent talker identification with neural networks. In *Proceedings of ICASSP 91*, vol. 1, pp. 389–392. 1991. Toronto, Canada.

[144] H. Sakoe & S. Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 26 : pp. 43–49, 1978.

[145] Marvin R. Sambur. Selection of acoustic features for speaker identification. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 23, no. 2 : pp. 176–182, Apr. 1975.

[146] Gilbert Saporta. *Probabilités analyse des données et statistique*. Éditions Technip, 1990.

[147] Michael Savic & Sunil K. Gupta. Variable parameter speaker verification system based on hidden markov modeling. In *Proceedings of ICASSP 90*, vol. 1, pp. 281–284. Apr. 1990. Albuquerque, New Mexico, United-States.

[148] Michael Schmidt & Herbert Gish. Speaker identification via support vector classifiers. In *Proceedings of ICASSP 96*, vol. 1, pp. 105–108. May 1996. Atlanta, Georgia, United-States.

[149] Michael Schmidt, Herbert Gish & Angela Mielke. Covariance estimation methods for channel robust text-independent speaker identification. In *Proceedings of ICASSP 95*, vol. 1, pp. 333–336. 1995. Detroit, United-States.

[150] Stephanie Seneff & Victor W. Zue. Transcription and alignment of the TIMIT database. Documentation with a CD-ROM, 1988.

[151] Anand R. Setlur, Rafid A. Sukkar & Malan B. Gandhi. Speaker verification using mixture likelihood profiles extracted from speaker independent hidden Markov models. In *Proceedings of ICASSP 96*, vol. 1, pp. 109–112. May 1996. Atlanta, Georgia, United-States.

[152] Jia-Lin Shen, Wen-Liang Hwang & Lin-Shan Lee. Robust speech recognition features based on temporal trajectory filtering of frequency band spectrum. In *Proceedings of ICSLP 96*. 1996.

[153] M. Shridhar & N. Mohankrishnan. Text-independent speaker recognition : a revew and some new results. *Speech Communication*, vol. 1, 1982.

[154] M. Shridhar, N. Mohankrishnan & M. Baraniecki. Text-independent speaker recognition using orthogonal linear prediction. In *Proceedings of ICASSP 81*, pp. 197–200. 1981. Atlanta, Georgia, USA.

[155] F.K. Soong & A.E. Rosenberg. On the use of instantaneous and transitional spectral information in speaker recognition. In *Proceedings of ICASSP 86*, pp. 877–880. 1986.

[156] F.K. Soong, A.E. Rosenberg, L.R. Rabiner & B.H. Juang. A vector quantization approach to speaker recognition. In *Proceedings of ICASSP 85*, vol. 1, pp. 387–390. 1985.

[157] Frank K. Soong & Aaron E. Rosenberg. On the use of instantaneous and transitional spectral information in speaker recognition. *IEEE Trans. ASSP*, vol. 36, no. 6 : pp. 871–879, Jun. 1988.

[158] Frank K. Soong, Aaron E. Rosenberg & Biing-Hwang Juang. A Vector Quantization approach to speaker recognition. Tech. Rep. 3, AT&T, Mar. 1987.

[159] K.N. Stevens, C.E. Williams, J.R. Carbonell & Barbara Woods. Speaker authentification and identification : a comparison of

spectrographic and auditory presentations of speech material. *Journal of the Acoustical Society of America*, vol. 44, no. 6 : pp. 1596–1607, 1968.

[160] Z. P. Sun & J. S. Mason. Combining features via LDA in speaker recognition. In *Proceedings of EUROSPEECH 93*, vol. 3, pp. 2287–2290. Sep. 1993. Berlin, Germany.

[161] J. Thompson & J.S. Mason. Within class optimization of cepstra for speaker recognition. In *Proceedings of EUROSPEECH 93*, vol. 1, pp. 165–168. Sep. 1993. Berlin, Germany.

[162] Naftali Z. Tishby. On the application of mixture AR hidden markov models to text independent speaker recognition. *IEEE Transactions on Signal Processing*, vol. 39, no. 3 : pp. 563–570, Mar. 1991.

[163] W. Voiers. Perceptual bases of speaker identity. *Journal of the acoustical Society of America*, vol. 36 : pp. 1065–1073, 1964.

[164] Xin Wang & Guotian Zhao. Text-dependent speaker verification using recurrent time delay neural networks for feature extraction. In *Proceedings of ICSP 93*, vol. 1, pp. 674–677. Oct. 1993. Beijing, CHINA.

[165] J. J. Webb & E. L. Rissanen. Speaker identification experiments using HMMs. In *Proceedings of ICASSP 93*, vol. 2, pp. 387–390. Apr. 1993. Minneapolis, United-States.

[166] P. Whittle. On the fitting of multivariate autoregressions, and the approximate canonical factorization of a spectral density matrix. *Biometrika*, vol. 50, no. 1–2 : pp. 129–134, 1963.

[167] Jared J. Wolf. Efficient acoustic parameters for speaker recognition. *Journal of the Acoustical Society of America*, vol. 51, no. 6 : pp. 2044–2056, 1972. Number 6 (Part 2).

[168] Luoping Xu. *Perceptually-based features for speaker identification*. Ph.D. thesis, University College of Swansea, University of Wales, Feb. 1992.

[169] Kin Yu, John Mason & John Oglesby. Speaker recognition models. In *Proceedings of EUROSPEECH 95*. Sep. 1995. Madrid, Spain.

[170] Kin Yu, John Mason & John Oglesby. Speaker recognition using hidden Markov models, dynamic time warping and vector quantisation. *IEE vision, image and signal processing*, 1995.

[171] X. Zhang & J.S. Mason ad E.C. Andrews. Multiple dynamic features to enhance neural net based speaker verification. In *Proceedings of EUROSPEECH 91*, pp. 1411–1414. 1991.

[172] Yuan-Cheng Zheng & Bao-Zong Yuan. Text-dependent speaker identification using circular hidden markov models. In *Proceedings of ICASSP 88*, vol. 1, pp. 580–582. Apr. 1988. New York, United-States.

[173] Yuan Zhong-Xuan, Xu Bo-Ling & Yu Chong-Zhi. A kind of fuzzy-neural networks for text-independent speaker identification. In *Proceedings of ICASSP 96*, vol. 2, pp. 657–660. May 1996. Atlanta, Georgia, United-States.