Bénéfices et limites des représentations en facteur de variabilité totale pour la reconnaissance du locuteur

Pierre-Michel Bousquet

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

THÈSE

présentée à l'Université d'Avignon et des Pays de Vaucluse
pour obtenir le diplôme de DOCTORAT

SPÉCIALITÉ : Informatique

École Doctorale 536 « Sciences et Agronomie »
Laboratoire d'Informatique (EA 4128)

Bénéfices et limites des représentations en facteur de variabilité totale pour la reconnaissance du locuteur

par
Pierre-Michel Bousquet

Soutenue publiquement le 23 Mai 2014 devant un jury composé de :

M.	Samy BENGIO	Google Inc., California (USA)	Rapporteur
M.	Pietro LAFACE	DAUIN, Politecnico di Torino (Italia)	Rapporteur
M^{me}	Delphine BLANKE	LMA Avignon	Présidente du juty
M.	Sylvain MEIGNIER	LIUM, Université du Maine	Examinateur
M.	Yannick ESTEVE	LIUM, Université du Maine	Examinateur
M.	Driss MATROUF	LIA Avignon	Examinateur
M.	Jean-François BONASTRE	LIA Avignon	Directeur de thèse

Laboratoire d'Informatique d'Avignon - CERI
Centre d'Enseignement et de Recherche en Informatique

Remerciements

La découverte du monde de la recherche n'aura pas été pour moi qu'une aventure scientifique. Elle a constitué aussi une expérience humaine marquante. Le goût du savoir, des lumières et les qualités personnelles des personnes suivantes ont enrichi ma vie et méritent ma sollicitude : Benjamin Bigot, Delphine Blanke, Jean-François Bonastre, Edith Gabriel, Juliette Kahn, Anthony Larcher, Driss Matrouf, Gilles Pouchoulin, Mickaël Rouvier, Eric Sanjuan, Grégory Senay, Florian Verdet.

Merci à Alexandrine, Marie et Cécile. "Sans amour, on n'est rien du tout ...".

Résumé

Le domaine de la reconnaissance automatique du locuteur (RAL) recouvre l'ensemble des techniques visant à discriminer des locuteurs à partir de leurs énoncés de voix. Il se classe dans la famille des procédures d'authentification biométrique de l'identité. La reconnaissance du locuteur a connu ces dernières années une avancée significative avec un nouveau concept de représentation de l'énoncé de voix, désigné sous le terme de *i-vector*. Ce type de représentation s'appuie sur le paradigme de modélisation par mélange de gaussiennes et présente la particularité de se réduire numériquement à un vecteur de dimension faible, au regard des représentations précédentes, et pourtant très discriminant vis à vis du locuteur.

Les travaux présentés dans cette thèse s'inscrivent dans ce nouveau contexte. Orientés autour de cette représentation, ils visent à en comprendre et évaluer les hypothèses, les points fondamentaux, le comportement et les limites.

Nous avons en premier lieu conduit une analyse statistique sur cette nouvelle représentation. L'étude a porté sur l'effet et l'importance relative des différentes étapes de constitution et d'exploitation du concept. Cette analyse a permis de mieux comprendre ses caractéristiques, mais aussi de faire apparaître des défauts de la représentation qui nous ont conduits à mettre en place de nouvelles transformations dans cet espace. L'objectif de ces techniques est de faire converger les données vers des modèles théoriques, à meilleur pouvoir discriminant. Nous recensons et démontrons un certain nombre de propriétés induites par ces transformations, qui justifient leur emploi. En terme de performance, ces techniques réduisent d'un ordre de grandeur de 50% les taux d'erreur des systèmes basés sur les i-vectors et des postulats gaussiens, permettant notamment d'atteindre par la voie du cadre probabiliste gaussien les meilleurs taux de détection dans le domaine.

Une évaluation générale des composants de la méthode est ensuite détaillée dans ce document. Elle met en avant l'importance de certaines étapes, permettant ainsi de dégager, par comparaison à des méthodes alternatives, les approches fondamentales qui confèrent au concept une valeur de paradigme. Nous montrons la primauté de certaines étapes stratégiques dans la chaîne des traitements, parmi lesquelles les transformations que nous avons mises en place, et leur relative indépendance aux méthodes

et hypothèses adoptées.

Des limites de la solution sont mises au jour et exposées dans une étude dite d'*anisotropie*, qui relativise sa capacité à produire une paramétrisation linéaire globale des variabilités qui soit optimale.

En parallèle de ces investigations, nous avons participé à l'exploration d'un nouveau modèle alternatif à la solution la plus usuelle de représentation des énoncés de voix. Conçu par J.F. Bonastre, il produit des vecteurs sous forme de clés binaires et fournit les moyens de les comparer, en suivant une voie semi-paramétrique basée sur une nouvelle approche de la problématique. Cette exploration a contribué à l'amélioration de ce modèle et à l'ouverture de nouvelles pistes. Elle a été également utile à notre évaluation du concept de i-vector.

Les travaux présentés dans ce document contribuent à l'amélioration de ce modèle et à l'ouverture de nouvelles pistes. Ils sont également utiles à notre évaluation du concept de i-vector.

Enfin, quelques aménagements des solutions i-vectors à des cas particuliers ont été mis en place : nous proposons de nouvelles variantes pour gérer la décision sur les énoncés de courte durée (qui constituent l'un des enjeux actuels du domaine) et sur les énoncés présentant une divergence a priori (support, durée, langue distincts).

L'ensemble de ces travaux vise à mieux circonscrire les pistes de recherche les plus porteuses autour de ce nouveau concept de représentation de la voix humaine.

Abstract

The speaker recognition field covers all the techniques intended to authentify the identity by using voice utterances. Speaker recognition has experienced in recent years a significant step forward with a new concept of representation, referred to as the i-vector . This type of representation is based on the Gaussian mixture model paradigm and has the distinguishing feature of being a small size vector compared to previous representations, yet very discriminating towards the speaker.

The works presented in this thesis are within that new context. Focused on this representation, they aim to better understand it and assess its assumptions, highlight its key points, its behaviors and limits.

We first carried out a statistical analysis of this new representation. This analysis helped to better understand its characteristics, but also reveal defaults of the representation that led us to develop new transformations. The goal of these techniques is to move data towards a theoretical model, having a better accuracy for discrimination. We identify and demonstrate a number of properties of these transformations which justify their relevance. In terms of performance, applying these techniques reduce by an order of magnitude of 50% the error rate of systems based on i-vectors and Gaussian assumptions and yield the best detection rate in the field through the Gaussian probabilistic framework.

A complete evaluation of the system components is detailed later in this document. By comparing the fundamental approaches to alternative methods, this evaluation identifies and highlights the fundamental steps that give the concept a value of paradigm. We show the primacy of some strategic steps in the process chain, including our propositions, and their relative independence from methods and assumptions.

Limits of the solution are uncovered and exposed in a study of "anisotropy", which reveals some lack of compliance of i-vector distributions with Gaussian assumptions.

Alongside these investigations, we participated in the exploration of a new model, alternative to the most usual statistical representations of utterances, which relies on a semi- parametric representation. Designed by J.F. Bonastre, it produces binary key vectors and provides the means to compare them. This exploration has contributed to the

improvement of this model and opens new gates. It was also helpful to our evaluation of the concept of i -vector.

Some adaptations of i-vector approach to special speaker recognition tasks are described : we propose new variants to handle short duration utterances (which is one of the current issues in the field) and to deal with a priori mismatch (for example of support, time or distinct language).

We hope that this work will better highlight some of the most promising slopes of research around this new concept of representation for speaker recognition.

Table des matières

Introduction

La vertu intrinsèque du modèle réduit est qu'il compense la renonciation à des dimensions sensibles par l'acquisition de dimensions intelligibles.
Claude Levi-Strauss (La pensée sauvage)

La parole est un moyen de communication universel, que la technologie permet aujourd'hui de diffuser, stocker et restituer à une échelle planétaire. Le traitement du signal de parole est l'enjeu d'un grand nombre d'applications, dans des domaines aussi variés que la sécurité, le pilotage de machines ou l'indexation de documents électroniques.

Les signaux de parole que nous émettons contiennent, en dehors du message linguistique qu'ils transmettent, des caractéristiques particulières à chacun. Ces caractéristiques nous permettent notamment de reconnaître les voix des personnes qui nous sont familières, au moins sur des signaux de voix peu bruités.

La reconnaissance automatique du locuteur (RAL) désigne l'ensemble des procédures automatiques visant à discriminer des locuteurs à partir de leurs énoncés de voix. Elle s'appuie sur la théorie du signal et sur des techniques d'apprentissage automatique. Ces dernières privilégient à la définition de règles la validation statistique. Des représentations numériques du signal de parole, ainsi que des méthodes de détection, sont élaborées afin d'évaluer des hypothèses sur l'identité d'un locuteur présumé.

Une majorité des approches mathématiques mises en oeuvre en RAL reposent sur des approches probabilistes. Parmi les familles de lois utilisées dans le domaine, la plus répandue est la mixture de lois gaussienne multidimensionnelles (GMM) associée à une représentation du signal basée sur des paramètres cepstraux (enveloppe spectrale à court-terme). Un modèle générique, également appelé modèle du monde ou UBM (*Universal Background Model*), permet de représenter le modèle du non-locuteur mais aussi de structurer l'espace acoustique. Par les performances qu'elles permettent d'atteindre, les méthodes basées sur ces approches sont implémentées aujourd'hui dans un très grand nombre de systèmes de RAL.

Problématique

Les principaux questionnements autour de tels systèmes sont les suivants :

- Les représentations vectorielles des énoncés de voix à partir des GMM emploient des paramètres statistiques, des techniques d'adaptation à posteriori et peuvent aussi faire appel à des hypothèses complémentaires, non-gaussiennes voire non-paramétriques. Des transformations vectorielles de ces représentations sont également proposées. L'élaboration de la représentation la mieux adaptée à la nature statistique des paramètres acoustiques de la voix humaine, dans l'objectif de discrimination du locuteur, continue de susciter débats, études et interrogations.
- Pour prendre en compte les variabilités intrinsèque et extrinsèque au locuteur, des facteurs sont extraits de la représentation vectorielle, associés à ces différentes variabilités. La démarche la plus courante, une décomposition additive, isole un facteur supposé invariant pour un locuteur donné. La présence d'une part non négligeable de variabilité indésirable dans ce facteur-locuteur pose la question de la séparabilité des variabilités et continue d'occuper de nombreux travaux de recherche.
- Les modules décisionnels de discrimination du locuteur s'appuient sur les hypothèses sous-jacentes aux représentations et décompositions en facteurs, ainsi que sur les paramètres statistiques estimés par apprentissage. La qualité de la détection dépend essentiellement de la conformité des objets vectoriels à ces hypothèses et à la bonne estimation de leurs paramètres. La difficulté à constituer des systèmes robustes a été souvent mise en avant et de nombreux travaux tentent de pallier ces carences des modèles théoriques.

C'est dans ce contexte qu'est apparu en 2008 une nouvelle approche : (Dehak et al., 2011) introduisent dans les systèmes de RAL basés sur les GMM un objet vectoriel (*i-vector*) capable de résumer numériquement un énoncé de voix complet par un ensemble très réduit de coefficients réels (moins de 600). Ce concept ne vient pas remplacer l'ensemble des acquis précédents. Par contre, il bouscule les intuitions de la communauté, son caractère exploratoire, validé par ses performances mais justifié par quelques rares remarques et observations, nécessitant d'en poursuivre la compréhension ontologique. Alors que cette solution se diffuse rapidement dans les systèmes, l'enquête sur ses propriétés et sur les causes de son efficacité restait à effectuer.

Nos travaux s'orientent autour de ce nouveau concept de représentation. Ils répondent à la nécessité d'en comprendre et évaluer les hypothèses, les points fondamentaux et d'en connaître le comportement et les limites, cela autant pour en améliorer la maîtrise que l'efficacité et la fiabilité.

Contributions

Nous présentons d'abord recensons et dans cette thèse l'analyse que nous avons conduite sur cette nouvelle représentation. L'étude a porté sur ses caractéristiques statistiques et spatiales générales. Un outil de visualisation des variabilités a été conçu,

permettant d'apprécier clairement leurs distributions et l'effet de transformations vectorielles sur celles-ci. La conformité des vecteurs à des hypothèses fondamentales (standardité, gaussianité) a été évaluée.

Ces études ont fait apparaître des défauts de la représentation et nous ont amenés à envisager des transformations des données capables d'en corriger ou atténuer les carences. Il est ressorti de ces investigations une famille de transformations intermédiaires pour les systèmes de RAL basés sur les i-vectors. Ces transformations peuvent être rangées dans la catégorie des normalisations. Leurs particularités, en dehors du fait qu'elle se permettent de modifier les représentations, sont d'être non-linéaires et non-paramétriques. Leurs descriptions ont été publiées dans (Bousquet et al., 2011b) et (Bousquet et al., 2012a).

Les vecteurs après transformations tendent vers un nouveau modèle théorique. Dans cette thèse, nous montrons cette convergence puis recensons et démontrons un ensemble de propriétés vérifiées par des vecteurs de ce modèle, qui éclairent et justifient son opportunité dans le contexte de la RAL.

Une fois ces nouvelles techniques insérées dans les systèmes, leur efficacité nous a amenés à reconsidérer la chaîne de traitement de ces derniers. Le rôle réellement joué, en terme de qualité, par les procédures employées aux différentes étapes méritait d'être réévalué. En particulier, la participation significative de chacune dans la qualité de la modélisation probabiliste semblait remise en question. Nous avons entrepris une évaluation générale d'un système de reconnaissance du locuteur basé sur une représentation compacte de type i-vector. A chacune de ces étapes, la solution majoritairement adoptée et d'efficacité reconnue a été mise en concurrence avec des méthodes alternatives.

La synthèse de ces évaluations a dégagé un ensemble d'étapes fondamentales du paradigme de la représentation compacte des énoncés de voix issue du GMM-UBM. Ces étapes ne sont plus étroitement liées à certaines méthodes et familles statistiques précises et elles isolent les clés de la réussite de ces systèmes.

Une fois définis clairement les fondements du concept, une analyse a montré ses limites. Son approche conduit en effet à une modélisation linéaire, basée sur des paramètres de variabilité universels. L'application de transformations non-linéaires en amont remet en doute la validité d'une telle modélisation. Des mesures dites d'*anisotropie* confirment ce fait et ouvrent de nouvelles perspectives de recherche.

En parallèle de ces investigations, nous avons participé a l'exploration et à l'amélioration d'un nouveau modèle de représentation des énoncés de voix. Présenté dans (Anguera and Bonastre, 2010) (Bonastre et al., 2011b), il produit des clés binaires du locuteur et fournit les moyens de les comparer, en suivant une voie semi-paramétrique basée sur une nouvelle approche de la problématique des densités. Cette exploration a contribué à l'amélioration de ce modèle et à l'ouverture de nouvelles pistes. Elle a été également utile à notre évaluation du concept de i-vector.

Une fois mis en place un système performant de reconnaissance du locuteur basé sur les i-vectors, son application à des systèmes particuliers d'authentification a été

envisagée :

- de nombreuses études tentent de gérer des défauts de support (non-correspondance sur les énoncés à comparer, par exemple téléphone vs microphone).
- la gestion des énoncés de courte durée est également l'un des enjeux actuels du domaine. L'estimation correcte des paramètres d'une voix semble difficile en dessous d'une vingtaine de secondes d'entraînement.
- la mise en place de systèmes capables de traiter les segments de voix fortement bruités doit être étudiée, même si la position intermédiaire du concept i-vectors, loin du signal initial, tendrait plutôt à chercher des aménagements dans la paramétrisation acoustique initiale.

Nous avons cherché à raffiner l'obtention d'une décision basée sur des énoncés courts et également adapté un modèle i-vector au cas de données mixtes, c'est à dire présentant une non-correspondance connue a priori.

Il a été volontairement choisi dans ce document de conserver la terminologie anglo-saxonne pour un certain nombre de méthodes et concepts du domaine. Ces termes appartiennent moins à la langue de Shakespeare qu'à celle d'une communauté d'hommes et femmes dans le monde, qui les emploient pour des échanges constructifs. Quant à la langue de Baudelaire, elle ne m'en voudra pas, je l'espère, de l'exempter de certains termes techniques.

Organisation du document

Le chapitre 1 présente la problématique de la reconnaissance automatique du locuteur. Les différentes tâches de la reconnaissance automatique du locuteur, leurs mesures de performance y sont décrites, ainsi que les corpus et campagnes d'évaluation que nous avons employés dans notre travail.

Le chapitre 2 constitue une description schématique des systèmes de reconnaissance automatique du locuteur. De la paramétrisation acoustique du signal à la production d'une décision, les différentes étapes de traitement y sont présentées, en concentrant nos propos sur les méthodes pré ou postliminaires au concept de i-vectors.

L'investigation statistique du système au coeur de notre travail est présentée au chapitre 3. Cette analyse nous a conduit à introduire plusieurs transformations, qui sont exposées dans le chapitre 4. Nous énumérons et démontrons les propriétés des données après application de ces transformations, justifiant leur emploi et expliquant leur efficacité.

Le chapitre 5 est une évaluation générale du système par i-vectors majoritairement implémenté dans les applications de reconnaissance du locuteur. Nous mesurons principalement l'impact de chacune de ses étapes dans la qualité de la modélisation probabiliste. Ces étapes sont mises en concurrence avec des méthodes alternatives, déterministes voire non-paramétriques. Il ressort de cette évaluation l'émergence d'un concept, implicite jusque là et partiellement indépendant des approches adoptées à

chaque étape. Ce concept prend valeur de paradigme, les différentes évaluations menées dans le chapitre montrant sa primauté dans la réussite de la démarche.

Le chapitre 6 décrit l'analyse de la mesure d'anisotropie qui montre les limites de la modélisation linéaire.

Le chapitre 8 expose nos contributions au modèle dit de clés binaires du locuteur, alternative semi-paramétrique à la représentation des énoncés de voix.

Le chapitre 9 expose nos contributions à des contextes d'applications particuliers : énoncés de courte durée, énoncés en non-correspondance (court vs long, téléphone vs microphone, etc...).

Nous concluons par une synthèse de notre approche et les inévitables questions que soulève toute réponse ou initiative.

Première partie

Introduction au domaine de recherche

Chapitre 1

La reconnaissance automatique du locuteur

L'homme est par nature un système ouvert, porteur et récepteur d'informations. La technologie moderne accroît la portée de ces informations, tout en facilitant leurs stockage, restitution et leur manipulation.

Un grand nombre d'applications, disposant de certaines de ces informations émanant de notre personne, demandent à des procédures automatiques de les exploiter, pour aider à nous différencier ou à nous reconnaître. Qu'il s'agisse de protéger nos données personnelles ou de rechercher notre présence en un lieu et un instant donné, ces applications attendent de la technique qu'elle prenne en charge ces tâches de discrimination des identités.

Mais l'identité est complexe et aucun caractère exclusif ne permet de l'authentifier sans ambiguïté. Les procédures automatiques de détection de l'identité produisent donc des probabilités, ou plus généralement des mesures de vraisemblance (ou similarité) d'une hypothèse. Elles peuvent être transformées en décisions conformes aux spécifications des décideurs, par exemple une réponse binaire d'acceptation ou rejet, ou un indicateur d'évaluation de risque.

Parmi l'ensemble des techniques d'authentification de l'identité, celles dites *biométriques* cherchent à dégager des caractéristiques physiques une signature, un profil de la personne, ou plus généralement des caractéristiques qui lui sont propres, en s'appuyant sur des hypothèses crédibles de différentiation.

Le terme de reconnaissance automatique du locuteur (RAL) désigne l'ensemble des techniques biométriques d'authentification basées sur le signal de parole. Les applications de la RAL sont multiples :

- sécurisation d'accès à des espaces personnels, sécurisation de transactions,
- analyse de communications téléphoniques pour le renseignement (recherche de terroristes) ou la criminalistique,

– gestion électronique de documents : indexation, recherche d'informations par navigation dans leurs contenus audio et video.

L'énoncé de parole contient plus d'informations que le message linguistique qu'il porte. Langue employée, émotions, identité sont plus ou moins perceptibles par l'être humain dans la communication orale. Nous disposons d'une certaine aptitude à distinguer les locuteurs, comme à détecter une identité familière à partir de son discours. La connaissance qu'un système de RAL doit acquérir soulève une question centrale : qu'est-ce qui fait l'identité vocale ?

Un certain nombre de variabilités interviennent dans un énoncé de parole, extrinsèques ou intrinsèques au locuteur :

– des paramètres environnementaux sont à prendre en compte dans le traitement du signal : bruit ambiant, caractéristiques du canal de transmission.
– le signal de voix n'est que l'héritage d'un processus organique de génération de sons, dont il n'est pas encore possible de coder directement les constituants. Le conduit vocal de chaque individu possède ses propres caractéristiques physiologiques, qui influent sur la production de parole. Le caractère acquis d'une grande partie de la production de parole conduit également à des distinctions d'ordre culturel et plus généralement comportemental.
– les contenus linguistique et phonologique d'un énoncé, l'état physique et moral du locuteur, contribuent encore à varier les productions orales d'une personne.

La théorie du traitement du signal permet la numérisation des énoncés analogiques. Cette étape est dite de "paramétrisation acoustique". Des approches mathématiques permettent ensuite de former des représentations vectorielles à partir des énoncés de voix, adaptées à la discrimination du locuteur. Elles s'appuient sur des hypothèses probabilistes et l'apprentissage automatique, domaine d'acquisition de connaissances par exploitation statistique de vastes corpus d'entraînement.

Ces étapes sont décrites au chapitre suivant. Nous présentons dans ce chapitre différentes ressources employées pour estimer la qualité de ces méthodes et plus généralement d'un système de RAL.

De vastes collections d'énoncés de voix et de grands jeux de tests sont utilisés par les organismes industriels et de recherche, leur permettant d'estimer et comparer l'efficacité de leurs différents systèmes et configurations. Nous décrivons brièvement dans la section 1.3 de ce chapitre les collections d'énoncés et les évaluations diffusées par les organismes américains LDC (*Linguistic Data Consortium*) et NIST (*National Institute of Standard and Technology*), que nous avons utilisées dans nos travaux. Avant cela, nous référençons dans la section 1.2 les mesures de performance employées pour quantifier l'efficacité d'un système, lorsque la réponse à fournir est de type binaire (acceptation ou rejet). En premier lieu, nous présentons les différents types de tâches auxquelles la RAL est confrontée.

1.1 Procédures de reconnaissance

Les domaines d'application de la reconnaissance du locuteur sont très variés. Leurs diverses problématiques se ramènent à trois types de tâches, qui peuvent elles-mêmes se résumer aux questions suivantes :

"- Qui parmi ces personnes a prononcé ce message ?" (Identification),

"- Comment se sépare ce signal de parole, en termes d'interventions des divers participants ?" (Segmentation, indexation),

"- Ces énoncés proviennent-ils du même locuteur ?" (Vérification, détection).

1.1.1 Identification du locuteur

L'objectif de l'identification du locuteur est de déterminer qui, dans un panel de personnes, a émis un message vocal donné. Des "modèles" de voix de chacune de ces personnes sont constitués. Etant donné un nouveau message vocal, des mesure de similarité de ce message à chacun de ces modèles sont calculées. La tâche est dite *fermée* si le locuteur du message vocal est obligatoirement dans le panel, *ouverte* si le cas d'un inconnu au système peut se produire. Ses mesures de similarité peuvent conduire à une décision associant le message à un unique locuteur.

1.1.2 Segmentation et structuration en locuteurs

Cette tâche traite les signaux de parole contenant les interventions de parole de plusieurs locuteurs. Il s'agit (avant, par exemple, une tâche de reconnaissance de la parole) de décomposer le signal en segments supposés propres à un seul intervenant et de classifier ces segments pour reconstituer le panel des participants et leurs interventions.

1.1.3 Vérification du locuteur

La vérification du locuteur (VAL) détermine si un énoncé de voix a bien été prononcé par un locuteur présumé ou "proclamé" (appelé *cible* ou *client*). Cette tâche intervient notamment dans les systèmes de sécurité d'accès et dans le cadre d'expertises judiciaires.

L'énoncé de voix à étudier, dit de *test*, est comparé à un modèle de voix du locuteur. Ce modèle est élaboré à partir des énoncés disponibles du locuteur-cible, se résumant parfois à une seule occurrence. Les tests portant sur le même locuteur sont appelés "tests-cible", dans le cas contraire "tests-imposteur".

Dans certains cas, la tâche de vérification peut imposer l'émission d'un message vocal (accès par mot de passe, enquêtes judiciaires). Le système est dit dépendant du texte.

Nous avons porté nos études sur des tâches de vérification du locuteur indépendantes du texte et les méthodes que nous décrivons dans les chapitres suivants s'adaptent à ce type de tâche.

1.2 Mesures de performance en VAL

Les systèmes de VAL produisent des mesures de vraisemblance d'une hypothèse (ou de "similarité") entre des énoncés de voix. On appelle *score* en général une telle mesure et *scoring* la méthode qui l'a générée. La traduction française du mot score ("notation statistique") rappelle le caractère estimatif de cette mesure, souvent appuyé par des hypothèses probabilistes.

Lorsqu'une décision binaire est demandée, les valeurs de score sont seuillées pour produire une réponse, d'acceptation ou rejet de l'hypothèse de même identité. La performance d'un système de VAL à décision binaire peut être évaluée à partir d'un jeu de tests composé de tests-cible et tests-imposteur. Nous décrivons dans cette partie les métriques d'évaluation usuellement employées dans ce cadre et en particulier dans celui des campagnes d'évaluation NIST, décrites à la section 1.3.

1.2.1 Fonction de coût

Une réponse négative sur un test-cible est appelée faux rejet (*False Reject FR*) ou détection manquée (*Missed detection*). Un second type d'erreur du système se produit lorsqu'une réponse positive a été renvoyée sur un test-imposteur. Il est appelé fausse alerte [1] (*False Alarm FA*) ou fausse acceptation.

La qualité d'un système est mesurée à partir des taux de fausse alerte t_{FA} et faux rejet t_{FR} obtenus sur des données de développement. Etant donné un jeu de n tests composé de n_{cible} tests-cible et n_{imp} tests-imposteur, ces taux sont définis par :

$$t_{FA} = \frac{n_{FA}}{n_{\text{imp}}} \tag{1.1}$$

$$t_{FR} = \frac{n_{FR}}{n_{\text{cible}}} \tag{1.2}$$

où n_{FA} est le nombre de fausses alertes et n_{FR} le nombre de faux rejets.

Ces deux types d'erreur n'induisent pas nécessairement les mêmes "coûts" (au sens large du terme). La mesure de qualité fera donc intervenir des valeurs C_{FA} et C_{FR} de coûts FA et FR fixées a priori. La fonction de coût de décision (*Decision Cost Function, DCF* (Martin and Przybocki, 2000)) mesure cette qualité par la formule :

1. Le terme d'alerte ne traduit pas nécessairement l'apparition d'un événement négatif. Il marque le fait qu'une action découle en général de la détection.

$$DCF = C_{FA}\frac{n_{FA}}{n} + C_{FR}\frac{n_{FR}}{n} \tag{1.3}$$

qui peut se réécrire :

$$DCF = C_{FA}\frac{n_{\text{imp}}}{n}\frac{n_{FA}}{n_{\text{imp}}} + C_{FR}\frac{n_{cible}}{n}\frac{n_{FR}}{n_{\text{cible}}} \tag{1.4}$$

$$= C_{FA}\left(1 - \frac{n_{\text{cible}}}{n}\right)t_{FA} + C_{FR}\frac{n_{\text{cible}}}{n}t_{FR} \tag{1.5}$$

En fixant une fréquence d'événements a priori π_{cible}, on obtient :

$$DCF = C_{FA}\left(1 - \pi_{cible}\right)t_{FA} + C_{FR}\pi_{cible}t_{FR} \tag{1.6}$$

Cette valeur peut être normalisée par le coût minimal des deux systèmes triviaux obtenus en déclenchant systématiquement ou jamais l'alerte. Ce coût minimal, noté $DCF_{\text{défaut}}$, est égal à :

$$DCF_{\text{défaut}} = \min\left(C_{FA}\left(1 - \pi_{cible}\right), C_{FR}\pi_{cible}\right) \tag{1.7}$$

La valeur de coût normalisée qui s'en déduit est :

$$DCF_{\text{Norm}} = \frac{DCF_{\text{défaut}}}{DCF} \tag{1.8}$$

Une valeur supérieure à 1 indique un système à coût plus élevé qu'un système trivial.

En phase de développement, un système peut être évalué en calculant sur un jeu de tests renseignés la DCF minimale suivant l'ensemble des seuils possibles de décision :

$$DCF_{\text{min}} = \min_{\theta}\left\{C_{FA}\left(1 - \pi_{cible}\right)t_{FA}\left(\theta\right) + C_{FR}\pi_{cible}t_{FR}\left(\theta\right)\right\} \tag{1.9}$$

où $t_{FA}\left(\theta\right)$ et $_{FR}\left(\theta\right)$ sont les taux de FA et FR obtenus en fixant le seuil de décision à la valeur θ.

La DCF_{min} est un indicateur (optimiste) de coût potentiel du système. Le seuil θ_{min} correspondant est un estimateur du seuil optimal, qui peut être éventuellement ajusté puis utilisé lors du déploiement de l'application.

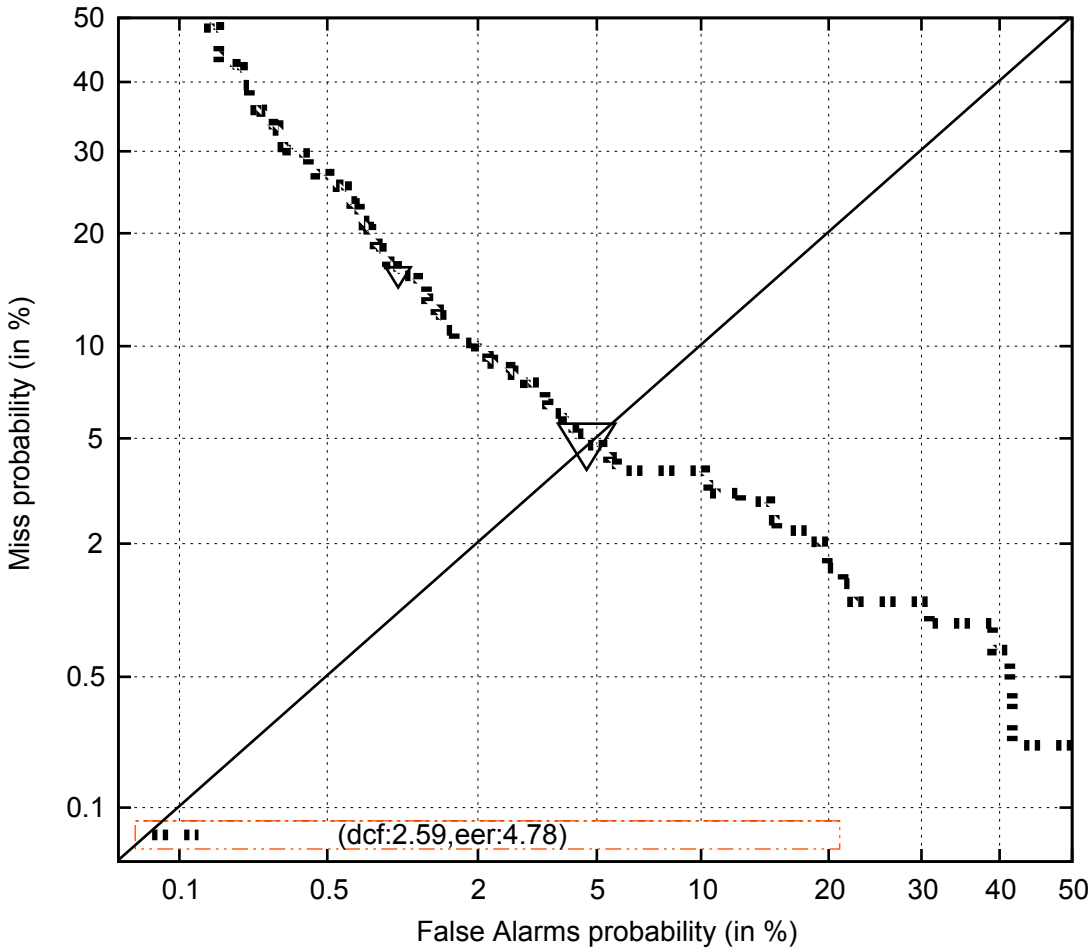

FIGURE *1.1* – *Un exemple de courbe décisionnelle DET (Detection Error Tradeoff)*

1.2.2 Mesures d'erreurs en développement

La capacité discriminante d'un système peut difficilement être évaluée lorsque les coûts FA et FR de la DCF sont trop déséquilibrés. C'est pourquoi, dans un cadre de recherche, d'autres mesures viennent compléter, voire remplacer, la DCF. Il peut s'agir du point de fonctionnement obtenu en minimisant l'HTER (*Half total Error Rate*), moyenne des taux d'erreur :

$$HTER_{\min} = \min_{\theta} \frac{t_{FA}(\theta) + t_F(\theta)}{2} \tag{1.10}$$

Il est également utile d'observer l'évolution complète de ces taux en fonction du seuil θ. Ceci est possible, l'intervalle de variation de θ étant borné aux valeurs extrêmes des scores du jeu d'essais. Les valeurs successives de ces taux sont représentées par la courbe des points $(t_{FA}(\theta), t_{FR}(\theta))$. Cette courbe, variante de la célèbre courbe décisionnelle ROC (*Receiver Operating Characteristic*), affiche ces points dans une échelle suivant l'évolution de la fonction de répartition d'une loi normale et porte le nom de courbe DET (*Detection Error Tradeoff* (Martin et al., 1997)). La figure 1.1 affiche un exemple de courbe DET. Par la nature -en général- gaussienne des scores, la courbe présente une allure linéaire grâce au changement d'échelle.

Le point de coordonnées $(1,0)$ correspond au cas d'un système à déclenchement systématique. Le point $(0,1)$ correspond à celui d'un système à abstention systématique. Entre ces points, la courbe permet d'observer l'évolution des deux taux d'erreur, inversement proportionnels. Les défauts de linéarité en bas à droite de la figure 1.1 indiquent par exemple une faiblesse de ce système, qui sous-évalue certains tests-cible.

A partir de cette courbe, une mesure basée sur un coût équilibré est proposée, nommée taux d'erreur égale (*Equal Error Rate, EER*). Etant donné le point de fonctionnement à taux égaux, c'est à dire celui obtenu pour le seuil θ_0 tel que les taux de FA et FR soient égaux, l'EER est défini par leur valeur commune $EER = t_{FA}(\theta_0) = t_{FR}(\theta_0)$.

Sur la figure 1.1, cette valeur correspond à l'intersection de la courbe avec la première bissectrice. Ce point de fonctionnement n'a pas d'interprétation directe en terme de coût, mais est très souvent utilisé comme une première indication des performances d'un système.

Notons que le point de fonctionnement minimisant la DCF, telle que spécifiée, est indiqué sur la courbe DET (second triangle sur la figure, en haut à gauche). L'EER sera utilisé le plus souvent dans ce document, où notre objectif principal est la comparaison de méthodes sans prise en compte de coûts déséquilibrés.

Période	C_{FA}	C_{FR}	π_{cible}
1997-2008	10	1	0.01
2010-2012	1	1	0.001

TABLE *1.1* – *Valeurs des coûts d'erreurs et probabilité cible a priori de NIST-SRE.*

1.3 Corpus et campagnes d'évaluation (LDC - NIST)

1.3.1 Présentation

L'organisme public américain de métrologie NIST est le maître d'ouvrage du gouvernement américain pour les campagnes d'évaluation de traitements automatiques de la parole, en particulier des systèmes de VAL [2]. Les objectifs stratégiques de ces campagnes concernent essentiellement les questions de sécurité du territoire et de renseignement.

Ces campagnes s'appuient sur les données collectées depuis plus d'une quinzaine d'année par le LDC [3]. Les sessions portent sur des personnes des deux sexes, d'âge varié, pouvant parlant en anglais natif ou pas, mais aussi dans une vingtaine d'autre langues. Les énoncés sont enregistrés à l'aide de différents microphones dans une salle peu bruitée, ou bien les locuteurs sont mis en relation téléphonique de manière aléatoire et conversent entre eux. Les durées sont variables (quelques secondes à plusieurs minutes), dans des cadres environnementaux et techniques variés (niveaux de bruit ambiant, d'effort vocal, caractéristiques du microphone et du canal de transmission, processus de codage du signal). Un important travail de nettoyage, validation, étiquetage et correction du corpus a été effectué récemment (Ferrer et al., 2011), permettant la constitution de fichiers d'apprentissage plus cohérents.

Par la taille considérable de leurs corpus, la possibilité qu'ils offrent aux organismes de recherche de comparer leurs systèmes et configurations, la relative variété de leurs conditions, les jeux d'évaluation NIST constituent l'un des moyens d'expérimentation les plus répandus dans le domaine de la VAL.

1.3.2 Evolution des campagnes

En dehors de l'élargissement du spectre des caractéristiques du corpus (cité plus haut), les campagnes NIST ont évolué par une inflation des volume de tests. A l'heure du *big data*, cette politique -coûteuse pour les participants- espère multiplier les tests de comparaison et permettre ainsi de mieux cerner certaines insuffisances particulières des systèmes.

Les paramètres de coût, pour les conditions principales, ont connu l'évolution indiquée sur la table 1.1.

2. *National Institute of Standard and Technology* http ://www.itl.nist.gov/iad/mig/tests/sre/
3. LDC The Linguistic Data Consortium http ://www.ldc.upenn.edu/

Jusqu'en 2008, le système pénalisait principalement les fausses alertes (une politique habituelle dans les procédures de détection d'événements rares, où la prudence est de mise). Les nouveaux paramètres depuis 2010 accentuent encore ce déséquilibre.

Il faut également noter que la fréquence a priori des tests-cible π_{cible} ne correspond pas à celle, effective, des jeux d'évaluation, ce qui ne facilite pas l'appréciation exacte de la qualité d'un système.

Depuis la campagne d'évaluation en reconnaissance du locuteur 2012, NIST propose une nouvelle DCF. Celle-ci tient compte du fait que l'apprentissage de modèles des locuteurs-cibles à partir de plusieurs de leurs segments de voix est autorisé (Bousquet et al., 2012c), ainsi que l'inclusion de locuteurs-cibles dans les données d'apprentissage des modèles de représentation. Ainsi, NIST distingue, parmi les tests-imposteur, ceux pour lesquels le locuteur proposé appartient à la liste des locuteurs-cible et ceux où il s'agit d'un inconnu vis à vis de cette liste. La nouvelle DCF calculée est :

$$DCF = C_{FR}\pi_{cible}t_{FR} + C_{FA}(1-\pi_{cible})\left(t_{FA/connu}\pi_{connu} + t_{FA|inconnu}(1-\pi_{connu})\right) \tag{1.11}$$

où π_{connu} est une fréquence a priori, imposée, qu'un test-imposteur porte sur un locuteur de la liste cible et $t_{FA/connu}$ (resp. $t_{FA/inconnu}$) le taux de fausse alerte sur les test-imposteurs de locuteurs connus (resp. inconnus).

Les expériences que nous avons menées pour évaluer et comparer des méthodes sont basées sur les conditions NIST-SRE (*Speaker Recognition Evaluation*) 2008 et 2010 les plus largement utilisées. La mesure de performance que nous avons privilégiée est celle par EER. Son analyse dégage en effet des enseignements plus généralisables sur les méthodes employées.

Les nombreux résultats consignés dans la littérature, sur ces conditions et en terme d'EER, ont également permis une meilleure appréciation du caractère significatif de la performance d'un système.

Chapitre 2

Systèmes de vérification automatique du locuteur

Nous décrivons dans ce chapitre les étapes constitutives d'un système de VAL. Le signal de voix n'est pas traité directement par ces systèmes. La phase de "paramétrisation acoustique", décrite à la section 2.1 convertit la version numérique de ce signal en données exploitables. Les représentations vectorielles des énoncés de voix adaptées à la discrimination du locuteur sont présentées à la section 2.2. Nous inventorions brièvement un certain nombre de méthodes de représentation et de décision, avant de détailler celles basées sur les lois probabilistes par mixture de gaussiennes, qui ont occupé nos recherches.

Dans ce cadre formel, les techniques de décomposition des représentations vectorielles en facteurs, attachés à des causes de variabilité précises, ont grandement contribué à la progression des systèmes. Nous décrivons ces techniques à la section 2.4.

La solution i-vectors est ensuite présentée, avec les différentes modélisations et métriques décisionnelles qui y sont appliquées.

Afin de mieux estimer l'impact des avancées dans le domaine, les performances de systèmes successifs basés sur les mixtures de gaussiennes sont indiquées dans la dernière section.

2.1 Paramétrisation acoustique

Le signal de parole est une onde acoustique convertie en signal électrique par un microphone. Ce dernier ne peut être traité directement. Après numérisation, la phase de paramétrisation acoustique a pour objectif de transformer ce signal, afin d'en extraire des caractéristiques morphologiques du conduit vocal du locuteur. La démarche adoptée suit une voie usuelle en traitement du signal : ce dernier est découpé en séquences de durée suffisamment courte pour considérer ses paramètres périodiques comme in-

variants (pseudo-stationnarité du signal sur un bref intervalle de temps). Chaque séquence doit être alors résumée par un jeu de mesures prédéfini, aboutissant à une représentation vectorielle de dimension fixe de cette séquence, appelée trame (*frame*, *feature*) : la théorie de Fourier sur les fonctions périodiques permet de considérer l'onde comme une combinaison linéaire de fonctions périodiques et de la résumer par des coefficients de ces fonctions.

La prise en compte du problème de reconnaissance du locuteur dès cette phase de paramétrisation est un point important, qui a fait l'objet de nombreuses études. La différenciation des locuteurs sera facilitée par une paramétrisation acoustique moins sensible à leurs variabilités extrinsèques : conditions d'enregistrement, bruit, variabilité intra-locuteur, ... A cet effet, des étapes de pré-traitement sont habituellement réalisées : pré-accentuation des hautes fréquences, filtrage passe-bande, ...

Les séquences de signal sont extraites sur une fenêtre temporelle glissante, avec recouvrement partiel, de durée variant généralement entre 20 et 50 millisecondes. Les trames sont constituées des vecteurs de coefficients cepstraux (coefficients de décomposition en série de Fourier du logarithme de la densité spectrale de puissance du signal (Bogert et al., 1963)). D'autres techniques de vectorisation, plus complémentaires que concurrentielles de l'analyse cepstrale, seront citées plus bas.

L'extraction des coefficients cepstraux est réalisée suivant deux approches ; elles ont en commun de produire des coefficients statistiquement décorrélés, cette indépendance permettant en théorie d'en limiter le volume.

La première approche utilise la transformée de Fourier rapide (FFT) sur chaque trame, puis une analyse par bancs de filtres qui est censée mieux rendre compte de la perception des fréquences par l'oreille humaine. Les fréquences centrales des coefficients d'énergie obtenus sont alors fixés suivant une échelle linéaire (*Linear Frequency Cepstral Coefficient : LFCC*) ou une échelle Mel (*Mel Frequency Cepstral Coefficient : MFCC*) (Stevens et al., 1937) (Davis and Mermelstein, 1980). Enfin, une transformée en cosinus discrète (*DCT*) appliquée au logarithme de ces coefficients d'énergie permet de les décorréler.

La seconde approche s'appuie sur la corrélation entre les échantillons successifs de paroles, attribuable à la résonance du conduit vocal. Une prédiction linéaire (*Linear Predictive Coding, LPC*) permet d'estimer des coefficients supposés significatifs de l'enveloppe spectrale. Leur transformation dans l'espace cepstral génère alors des coefficients cepstraux (*Linear Predictive Cepstral Coefficient : LPCC*) (Atal, 1974). La technique de prédiction linéaire est également utilisée pour produire des coefficients censés tenir compte du modèle de perception humaine de la parole (*Perceptual Linear Prediction, PLP*)(Hermansky, 1990), en particulier la perception non linéaire des fréquences par l'oreille. Une analyse spectrale relative (RASTA) peut alors éventuellement être effectuée, simulant l'insensibilité de l'oreille aux variations temporelles lentes (Hermansky et al., 1991) (Hermansky et al., 1992).

Citons également l'approche VTLN (*Vocal Tract Length Normalisation*) (Eide and Gish, 1996) de modélisation du conduit vocal du locuteur pour extraire des paramètres ca-

ractéristiques du locuteur. Elle est actuellement utilisée dans (Singer et al., 2012) pour un système de reconnaissance du langage.

Dans tous les cas de paramétrisations précédents, seuls les premiers coefficients cepstraux sont conservés (moins d'une vingtaine en général). La paramétrisation peut être enrichie par des informations dites "dynamiques" : il s'agit des variations immédiates du spectre. La vitesse et l'accélération de ces variations sont contenues dans les dérivées temporelles, première (Δ) et seconde ($\Delta\Delta$), formant les coefficients habituellement retenus pour tenir compte du caractère dynamique du signal (Furui, 1981). Remarquons que ces dérivées sont estimées localement par l'étude du taux d'accroissement sur des séries de trames consécutives et que des coefficients supplémentaires de mesure dynamique ont été proposés (Fredouille, 2000) (Magrin-Chagnolleau, 1997), les coefficients Δ et $\Delta\Delta$ restant toutefois les plus répandus.

Le vecteur acoustique d'une trame contient au final un nombre de coefficients variant en général entre 30 et 60. Le LIA, par exemple, utilise actuellement les 19 premiers coefficients cepstraux de fréquence linéaire (LFCC), leurs 19 Δ correspondants, 11 $\Delta\Delta$ et enfin le Δ d'énergie, soit 50 valeurs par trame.

D'autres méthodes de paramétrisation ont été proposées. Certaines visent à mieux prendre en compte certaines informations phonologiques de haut niveau : c'est le cas des paramétrisations prosodiques qui s'appuient sur des indices et paramètres tels que les formants F0, les intensités, "vallées" des éléments voisés, durées des mots, états ou pauses, ... (L. Ferrer and and et E. Shriberg, 2010) (Shriberg and Stolcke, 2008). Il existe également des paramétrisations basées sur les niveaux glottique, segmental ou lexical (Campbell et al., 2004). Ces paramétrisations sont utilisées par certains laboratoires (Scheffer et al., 2011), mais toujours en combinaison avec des systèmes basés sur les paramètres acoustiques -en général les coefficients cepstraux-, pour gagner en performance.

La paramétrisation ayant ici pour seul objectif la reconnaissance du locuteur, il est indispensable de faciliter la différenciation des locuteurs en éliminant la part la plus importante possible de données non-informatives. L'objectif de la sélection de trames (*Voice Activity Detection : VAD*) est de supprimer les parts de bruit, silence qui risquent de perturber les étapes de modélisation. Chaque trame issue du signal est soumise à un classifieur, visant à la conserver ou à l'éliminer. La décision est effectuée à partir de la quantité d'énergie de la trame : les périodes de "bruit seul" et celles de "parole dominante + bruit", définies comme celles de faible ou forte énergie, sont modélisées par des lois gaussiennes, entraînant une décision par seuil (Magrin-Chagnolleau, 1997). D'autres classifieurs, basés sur des machines à vecteur de support SVM (Vapnik, 1998) (Vapnik, 1995) (Enqing et al., 2002) (Larcher, 2009) ou des réseaux de neurones (Ikedo, 1998) ont également été proposés.

Une dernière étape de normalisation des vecteurs acoustiques est, en général, effectuée. Elle consiste en un centrage-réduction des coefficients cepstraux, par soustraction de leur moyenne (*Cepstral Mean Substraction CMS*) et éventuellement division par leur

écart-type. Combinées, ces deux opérations ramènent les paramètres de tendance centrale de l'échantillon aléatoire aux valeurs de moyenne 0 et variance 1 d'une loi normale standard. Il a été en effet remarqué (Furui, 1981) un décalage de la moyenne cepstrale induit par le type de canal de transmission. De même, des bruits peuvent occasionner une diminution de la variance spectrale. C'est donc un ensemble de biais induits par l'environnement de la session que cette étape de normalisation tente d'atténuer. Citons également la technique dite de *feature warping* (Pelecanos and Sridharan, 2001), qui centre et réduit la distribution des coefficients cepstraux suivant un principe assez courant de correspondance des fonctions de répartition.

2.2 Un bref inventaire de représentations et modèles en VAL

Une fois paramétré un segment de parole par une série de trames, le problème se pose de séparer les locuteurs à partir de ces séries. Les trames d'un segment constitue un ensemble de points, d'effectif variable, d'un même espace acoustique. La comparaison de nuages de points appartient au domaine statistique des *processus spatiaux*. Mais les méthodes employées dans ce domaine, aussi perfectionnées soient-elles, semblent encore trop complexes pour s'appliquer dans notre espace acoustique (un grand nombre d'entre elles travaillent sur des représentation de l'espace géographique réel, en 2 ou 3 dimensions). Une phase de modélisation est donc effectuée en VAL. La stratégie la plus communément adoptée consiste à produire une représentation vectorielle de chaque énoncé de voix dans un espace commun. Si la distinction naturelle entre les locuteurs est suffisamment préservée par ses représentations, des modélisations à hypothèses probabilistes permettent alors d'effectuer efficacement la reconnaissance du locuteur.

Nous présentons ici un bref inventaire de méthodes de représentation et de modélisation utilisées en VAL. La description détaillée des modélisations basées sur les mixtures de gaussiennes (dont les nouveautés de la solution i-vector) est effectuée plus loin (paragraphe 2.3), ainsi que celle de la modélisation par clés binaires du locuteur (paragraphe 2.8) à laquelle nous avons contribué et qui est aussi utilisée comme solution alternative pour les évaluations de systèmes i-vectors effectuées au chapitre 4.

La quantification vectorielle (*Vector Quantization VQ*) (Soong et al., 1985) (Mason et al., 1989) effectue, à partir des données d'enrôlement d'un locuteur donné, un partitionnement de l'espace acoustique en un nombre fini de régions et représente chacune par un vecteur centroïde. L'ensemble de ces vecteurs forme ce qu'on appelle un *dictionnaire de quantification*. La mesure de proximité entre ce dictionnaire et les trames d'un segment de test, calculée comme la moyenne des distances minimales entre chaque trame et les centroïdes, permet la comparaison de segments de parole. Cette méthode est qualifiable de déterministe, mais aussi de vectorielle et de non-paramétrique. Ses résultats dépendent très fortement de la taille fixée du dictionnaire.

Les approches à noyaux, telles que les machines à vecteur support (*Support Vector Machine SVM*) (Vapnik, 1979) (Vapnik, 1998) (Vapnik, 1995) (Wan and Campbell, 2000)

(Fine et al., 2001) appliquent aux trames une transformation non-linéaire vers un sur-espace de grande dimension. La méthode assure alors de maximiser la *marge* linéaire entre les classes. Les performances des systèmes basés sur les SVM concurrencent aujourd'hui celles de l'état de l'art par mixture de gaussiennes que nous présentons plus loin. Cette méthode est qualifiable de semi-déterministe (la transformation peut ou pas découler de lois théoriques), vectorielle et discriminante.

Dans la famille des classifieurs linéaires discriminants, (Burget et al., 2011) a introduit le classifieur par régression logistique, celui-ci ayant l'avantage de s'appuyer sur des hypothèses probabilistes et bayesiennes. Ses performances n'égalent toutefois pas celles des SVM dans les expériences de (Burget et al., 2011). Signalons également une approche récente de classifieur linéaire discriminant basé sur les SVM, *Pairwise SVMs* (Cumani et al., 2013), qui va au-delà des limites des SVM basés sur les supervectors.

Les modèles d'ancrage (Merlin et al., 1999) (Collet, 2006) représentent un locuteur relativement à un espace de modèles, sous la forme d'un vecteur de similarités vis à vis de ceux-ci. Une similarité peut être calculée par un score de comparaison (modèle d'ancrage vs données d'apprentissage du locuteur). Cette méthode est fortement dépendante de la robustesse des modèles d'ancrage. Elle dépend de leur faculté à synthétiser par un recouvrement discret les informations structurelles de l'espace acoustique.

D'autres méthodes ont été présentées depuis une vingtaine d'années, ainsi que des variantes des méthodes précédentes. Elle sont "orientées" génératives ou discriminatives : Réseaux Bayesiens Dynamiques (*DBN*) (Cooper and Herskovitz, 1992) (Heckerman, 1995), modèles de mélanges de segments (*SMM*) (Starpert and Mason, 2001) combinant les approches GMM (que nous proposons plus bas) et *Dynamic Time Warping* (*DTW*) (Furui, 1981) pour prendre en compte la structure temporelle du signal de parole. Citons également l'approche GDW (*Gaussian Dynamic Warping*) de (Bonastre et al., 2003), qui constitue un hybride exploitant la généralisation propre aux approches génératives et la modélisation structurelle du DTW.

2.3 Modélisations basées sur les mixtures de gaussiennes

La notion de modélisation basée sur les mixtures de gaussiennes (*Gaussian Mixture Model GMM*) s'inscrit dans le cadre plus général des modèles de Markov cachés (*Hidden Markov Models HMM*). Ceux-ci utilisent les notions d'état introduites par le mathématicien russe. Leur succession, affectée de probabilités de transition, permet d'élaborer un modèle stochastique du locuteur. Cette approche, initialement utilisée pour la reconnaissance de la parole (Rabiner, 1989), s'avère très efficace en mode dépendant du texte. Cette méthode effectue une structuration temporelle du signal. Elle est probabiliste, encore appelée générative car elle considère les vecteurs de paramètres comme issus d'une distribution statistique a priori. Elle est à classer dans les travaux de statistique *confirmatoire*, où il s'agit de vérifier l'adéquation des données empiriques à une distribution théorique, par opposition à la statistique *exploratoire* qui évite tout a priori.

En mode indépendant du texte, les GMM, qui correspondent à des HMM à un

état, s'adaptent bien à une succession temporelle des événements acoustiques supposée non-contrainte (Reynolds et al., 2000). Par rapport à un modèle mono-gaussien, les GMM permettent une estimation plus fine des distributions des vecteurs acoustiques. Les contraintes de carence en informations pour bâtir un GMM robuste peuvent être partiellement levées par des techniques d'adaptation à partir d'un modèle a priori. Nous décrivons dans la suite les détails de cette approche.

2.3.1 Densité d'une mixture de gaussienne

Etant donnée la collection de trames d'un segment de voix, il est possible de la modéliser suivant une démarche probabiliste en considérant que cette collection est issue d'une combinaison linéaire de distributions gaussiennes locales (*Gaussian Mixture Model, GMM*). La fonction de densité de distribution de la collection de trames $\mathcal{X}$ sous cette hypothèse s'écrit, suivant l'habitude bayesienne de notation par probabilités conditionnelles :

$$\forall x \in \mathcal{X}, \qquad P\left(x|\Theta\right) = \sum_{g=1}^{G} w_g \mathcal{N}\left(x|\mu_g, \mathbf{\Sigma}_g\right) \tag{2.1}$$

où :

- G est le nombre de gaussiennes de la mixture,
- $\left\{w_g\right\}_{g=1}^{G} \in \mathbb{R}^G$ est le vecteur de poids des gaussiennes,
- $\left\{\mu_g\right\}_{g=1}^{G} \in \left(\mathbb{R}^F\right)^G$ est la série des vecteurs de moyenne par gaussiennes (F étant la dimension de l'espace acoustique),
- $\left\{\Sigma_g\right\}_{g=1}^{G} \in \left(\mathcal{M}\left(F \times F\right)\right)^G$ est la série des matrices de covariance par gaussiennes,
- $\Theta = \left\{w_g, \mu_g, \mathbf{\Sigma}_g\right\}_{g=1}^{G}$ est le métaparamètre de la mixture de gaussienne,
- $\mathcal{N}\left(x|\mu_g, \mathbf{\Sigma}_g\right)$ est la fonction de densité de la loi gaussienne de paramètres $\left(\mu_g, \mathbf{\Sigma}_g\right)$ appliquée à la trame x.

La détermination de Θ pour une collection de trames $\mathcal{X}$ s'effectue par l'algorithme d'apprentissage EM (*Expectation Maximisation*)(Dempster et al., 1977). Cet algorithme itératif effectue à chaque étape deux phases *Expectation* et *Maximisation* destinées à augmenter la vraisemblance des données d'apprentissage au modèle de gaussiennes (d'où le suffixe *ML Maximum Likelihood* ajouté à son nom). L'algorithme garantit à chaque itération la croissance d'une fonction objective de vraisemblance des paramètres sachant les données $\mathcal{X}$. Il converge vers un maximum de vraisemblance, mais seulement local, dans le champ d'optimisation de la fonction de densité.

L'initialisation des paramètres Θ à optimiser peut être aléatoire, ou assistée comme nous l'indiquerons plus bas. L'approximation d'une fonction de densité par un GMM présente l'avantage de modéliser les distributions par des membres de la famille des lois exponentielles, ces lois autorisant beaucoup de manipulations mathématiques. Il est à noter que le GMM ne suppose pas que la densité empirique suive nécessairement

des lois gaussiennes : de même qu'une fonction réelle peut être approximée par des polynômes (sur un compact) ou une fonction périodique par des séries de Fourier, une loi peut être approximée par une mixture de Gaussiennes. L'augmentation de la taille de la mixture permet en effet, dans la plupart des cas, de faire tendre le GMM vers la distribution empirique. Mais cette remarque met en avant un obstacle à la modélisation d'un modèle locuteur par GMM : étant donnée la collection de trames d'un segment de parole d'un locuteur, l'approximation de sa distribution par mixture de gaussiennes nécessite l'estimation d'un nombre important de paramètres. Le métaparamètre $\Theta = \left\{w_g, \mu_g, \mathbf{\Sigma}_g\right\}_{g=1}^{G}$ de la mixture contient $G\left(F\left(F+1\right)/2+F+1\right)$ valeurs à estimer. Or, la collection de trames $\mathcal{X}$ pour un segment de voix de durée initiale allant de 30 secondes à quelques minutes -comme c'est le cas dans beaucoup de systèmes de reconnaissance- contient un effectif de trames de l'ordre de 5 à 20 000 trames après VAD. Pour une dimension de l'espace acoustique $F = 50$ et une mixture à $G = 64$ gaussiennes, le métaparamètre Θ contient déjà $64\left(50 \times 49/2 + 50 + 1\right) = 81664$ valeurs. Et ici, le nombre G de 64 s'avère assez loin de la quantité minimale empirique nécessaire pour façonner une mixture d'un segment de voix de vraisemblance satisfaisante. L'estimation EM-ML conduit alors, par sous-apprentissage, à un modèle médiocre.

L'alternative consiste à imposer aux matrices de covariance Σ_g la contrainte de diagonalité dans l'algorithme EM. La matrice $\mathbf{\Sigma}_g$ de chaque gaussienne est seulement remplie avec la diagonale σ_g^2 des variances. Cette contrainte peut apparaître restrictive, voire peu réaliste, mais en réduisant de $G\left(F\left(F+1\right)/2+F+1\right)$ à $G\left(2F+1\right)$ le nombre de valeurs à estimer, elle permet l'accroissement du nombre G de gaussiennes de la mixture et donc de la précision locale de l'estimation.

2.3.2 Mesures de vraisemblance

Etant donné la collection de trames $\mathcal{X}$ d'un énoncé de voix et un locuteur présumé s, le système doit déterminer la probabilité de l'hypothèse locuteur H_0 : "cet énoncé de voix a été prononcé par s". Cette probabilité s'écrit :

$$P\left(H_0|\mathcal{X}\right) = \frac{P\left(\mathcal{X}|H_0\right)P\left(H_0\right)}{P\left(\mathcal{X}\right)} \tag{2.2}$$

Etant définie une densité pour la loi du modèle de s, le facteur de vraisemblance $P\left(\mathcal{X}|H_0\right)$ est alors estimé par la valeur de cette densité pour $\mathcal{X}$. Sous l'hypothèse d'indépendance des trames x de $\mathcal{X}$, ce facteur est le produit des vraisemblances $P\left(x|H_0\right)$.

Le problème du calcul du dénominateur est levé en introduisant l'hypothèse H_1 "non-locuteur", complémentaire de H_0. La probabilité de l'équation précédente s'écrit :

$$P\left(H_0|\mathcal{X}\right) = \frac{P\left(\mathcal{X}|H_0\right)P\left(H_0\right)}{P\left(\mathcal{X}|H_0\right)P\left(H_0\right) + P\left(\mathcal{X}|H_1\right)P\left(H_1\right)} \tag{2.3}$$

Il est alors possible de calculer $P(H_0|\mathcal{X})$, à condition d'avoir défini la densité de la loi des "imposteurs" de s (tous les locuteurs hormis s) et estimé les fréquences a priori $P(H_0)$ et $P(H_1)$.

La comparaison des probabilités $P(H_0|\mathcal{X})$ et $P(H_1|\mathcal{X})$ permet de mesurer le risque associé à la décision d'acceptation. Le ratio des hypothèses complémentaires (*likelihood ratio*) est défini par :

$$LR(H_0, H_1|\mathcal{X}) = \frac{P(H_0|\mathcal{X})}{P(H_1|\mathcal{X})} = \frac{P(\mathcal{X}|H_0)\,P(H_0)}{P(\mathcal{X}|H_1)\,P(H_1)} \tag{2.4}$$

En VAL, une décision binaire d'acceptation ou rejet est obtenue en fixant un seuil de décision Ω à $LR(H_0, H_1|\mathcal{X})$. Après incorporation des probabilités a priori $P(H_0)$ et $P(H_1)$ à ce seuil, la décision dépend seulement de la valeur $\frac{P(\mathcal{X}|H_0)}{P(\mathcal{X}|H_1)}$:

$$H_0 \text{ acceptée si} \frac{P(\mathcal{X}|H_0)}{P(\mathcal{X}|H_1)} > \Omega, \text{ rejetée sinon} \tag{2.5}$$

Dans le cadre de la modélisation par GMM, la vraisemblance $P(\mathcal{X}|H_0)$ est évaluée par la densité de $\mathcal{X}$ suivant la mixture de gaussiennes du locuteur s. La vraisemblance $P(\mathcal{X}|H_1)$ nécessite l'estimation d'un modèle GMM des imposteurs de s.

2.3.3 Le modèle GMM-UBM

Les paramètres GMM des imposteurs de s, nécessaires au calcul de $P(\mathcal{X}|H_1)$, peuvent être estimés à partir des trames d'une cohorte d'imposteurs, de modèles "proches" de s(Higgins et al., 1991) (Rosenberg et al., 1992). Mais la difficulté à déterminer la liste de ses imposteurs pour un locuteur donné, ainsi qu'à gérer des énoncés "éloignés" à partir d'imposteurs proches, a conduit (Carey and Parris, 1992) (Reynolds, 1995) à élaborer un modèle GMM unique appelé modèle du monde.

La stratégie adoptée consiste à collecter un nombre considérable de trames, issus de sessions de voix différentes et de locuteurs distincts, pour entraîner une mixture de gaussiennes générique. L'initialisation des paramètres peut être aléatoire, arbitraire ou assistée (par exemple par un dictionnaire VQ). La convergence vers un maximum local et non absolu s'avère moins préoccupante, étant donné la masse considérable de données d'entraînement. Cette masse autorise pour l'instant à se passer des raffinements algorithmiques consistant à ajouter régulièrement du bruit pour sortir de la "selle" d'un maximum local et espérer relancer la convergence vers une meilleure valeur d'optimisation, ou à des démarches par sous-échantillons tirés aléatoirement (Breiman, 2001) (Celeux and Diebolt, 1985) (Celeux and Diebolt, 1986).

Le modèle de mixtures de gaussiennes obtenu porte le nom de modèle du monde (GMM-UBM *Universal Background Model*). Cette dénomination sera commentée dans le paragraphe 2.3.6.1.

Le choix des données d'apprentissage du GMM-UBM est primordial dans la qualité d'un système de reconnaissance. Quantité et variabilité des données sont des facteurs prévisibles de qualité. Les modélisations par GMM-UBM privilégient les segments bien renseignés (longue durée, grand nombre de locuteurs). L'apport d'informations plus pauvres : enregistrements bruités, durée insuffisante, ... ne contribue que de manière minime à la qualité du résultat.

2.3.4 Adaptation à posteriori (MAP)

Malgré la réduction du nombre de paramètres GMM à estimer lorsque ses matrices de covariance par composante sont diagonales, celui-ci reste élevé pour un segment de voix. Pour un nombre de gaussiennes G de 512, valeur réaliste pour espérer constituer un GMM-UBM robuste, le nombre de valeurs à estimer, dans l'exemple LIA choisi, est de $512\,(2 \times 50 + 1) = 51712$. Disposant d'une collection de trames ne dépassant pas en général un effectif de 20 000 après VAD, le métaparamètre Θ d'un énoncé de voix reste trop volumineux pour espérer une approximation par GMM satisfaisante. Dans le cas de segments courts (moins de 20 secondes), il s'avère même tout à fait inadapté.

La solution consistant à réduire la durée de la fenêtre de Hamming de découpage en trames (habituellement fixée à une durée de 20 millisecondes avec un pas de 10 millisecondes), afin d'augmenter le nombre de trames fournies à l'algorithme EM, n'est pas pertinente : une "unité élémentaire" de parole couvre une durée de 50 à 150 millisecondes pour les occlusives (Marchal, 2007), autour de 150 millisecondes pour les voyelles (Ladefoged, 2005). S'il s'agit de parole, l'émiettement du signal instantané n'apporte plus d'informations à son analyse.

Etant donné un modèle GMM-UBM, la représentation d'un segment de voix s'obtient par adaptation de ce modèle aux données contingentes du segment. La méthode utilisée s'appuie sur la notion statistique de Maximum à Posteriori (MAP) ((Gauvain and Lee, 1994) et (Reynolds et al., 2000) pour son application en reconnaissance du locuteur). Le métaparamètre $\Theta_{UBM} = \left\{w_g, \mu_g, \mathbf{\Sigma}_g\right\}_{g=1}^{G}$ du GMM-UBM est adapté aux nouvelles données contingentes aux trames du segment à représenter, formant un métaparamètre adapté $\widehat{\Theta} = \left\{\widehat{w}_g, \widehat{\mu}_g, \widehat{\mathbf{\Sigma}}_g\right\}_{g=1}^{G}$. Le critère utilisé cherche à maximiser sa vraisemblance à posteriori du métaparamètre Θ_{UBM} de l'UBM et des nouvelles observations $\mathcal{X}$ du segment considéré :

$$\widehat{\Theta} = \arg\max_{\Theta} g\left(\Theta|\mathcal{X}\right) = \arg\max_{\Theta} f\left(\mathcal{X}|\Theta\right) g\left(\Theta\right) \tag{2.6}$$

où $f\left(\mathcal{X}|\Theta\right) = \prod_{x\in\mathcal{X}} \sum_{g=1}^{G} w_g \mathcal{N}\left(x|\mu_g, \mathbf{\Sigma}_g\right)$ est la densité de probabilité jointe des observations suivant l'UBM et g la densité de probabilité a priori de Θ. S'appuyant sur l'hypothèse d'indépendance des paramètres entre gaussiennes, (Gauvain and Lee, 1994) définit cette dernière fonction comme le produit d'une distribution de Dirichlet (pour

traiter les poids w) et d'une distribution Normal-Wishart inverse (loi conjointe a priori des moyennes et covariances, comme le propose DeGroot (DeGroot, 1970)).

L'adaptation selon le critère MAP s'effectue par algorithme EM. A chaque itération t, la phase d'*expectation* met à jour la variable latente y par son espérance :

$$\forall x \in \mathcal{X} \qquad y_g^{(t)} = \frac{w_g^{(t)} f\left(x|\mu_g^{(t)}, \mathbf{\Sigma}_g^{(t)}\right)}{\sum_{g'=1}^{G} w_{g'}^{(t)} f\left(x|\mu_{g'}^{(t)}, \mathbf{\Sigma}_{g'}^{(t)}\right)} \tag{2.7}$$

et celle de *maximization* évalue alors (itération $t+1$) les paramètres $\widehat{w}_g, \widehat{\mu}_g, \widehat{\Sigma}_g$ par maximum de vraisemblance :

$$w_g^{(t+1)} = \frac{n_g}{n_g + \tau_g^{[w]}} \left(\frac{1}{n} \sum y_g^{(t)}\right) + \frac{\tau_g^{[w]}}{n_g + \tau_g^{[w]}} w_g \tag{2.8}$$

où n taille de la collection $\mathcal{X}$ et $n_g = \sum y_g^{(t)}$

$$\mu_g^{(t+1)} = \frac{n_g}{n_g + \tau_g^{[\mu]}} \left(\frac{\sum y_g^{(t)} x}{\sum y_g^{(t)}}\right) + \frac{\tau_g^{[\mu]}}{n_g + \tau_g^{[\mu]}} \mu_g \tag{2.9}$$

$$\mathbf{\Sigma}_g^{(t+1)} = \frac{n_g}{n_g + \tau_g^{[\Sigma]}} \left(\frac{\sum y_g^{(t)\ t} \left(x - \mu_g^{(t+1)}\right) \cdot \left(x - \mu_g^{(t+1)}\right)}{\sum y_g^{(t)}}\right) + \frac{\tau_g^{[\Sigma]}}{n_g + \tau_g^{[\Sigma]}} \mathbf{\Sigma}_g \tag{2.10}$$

Nous détaillons ici ces formules pour y souligner le rôle des facteurs τ. Qualifiables de facteurs de confiance (*relevant factors*) ou de pertinence, ils contrôlent le degré d'adaptation de chaque paramètre UBM aux données. En théorie, ils dépendent du paramètre à estimer. En pratique, ils sont confondus, qu'il s'agisse de poids, moyenne ou matrice de covariance, en un seul vecteur $\left(\tau_g\right)_{g=1}^{G}$. Lorsque le nombre τ_g tend vers l'infini, les estimations $w_g^{(t+1)}$, $\mu_g^{(t+1)}$ et $\Sigma_g^{(t+1)}$ tendent vers les paramètres w_g, μ_g, $\mathbf{\Sigma}_g$ de l'UBM. Lorsque ce nombre tend vers 0, ces estimations sont seulement déterminées par la collection de trames du segment à représenter (algorithme EM-ML). Le problème du manque de données dans la seule collection de trames d'un segment de voix est donc évité, par balance entre un modèle du monde robuste et les nouvelles observations.

Comme le montrent les formules précédentes, un facteur τ_g indique une quantité d'informations associée à chaque composante du GMM-UBM. Le prolongement empirique de l'adaptation MAP consiste à ramener ces valeurs τ_g à des échelles qui les rendent comparables aux effectifs n_g des segments utilisés et à conserver les valeurs

fournissant les meilleures performances. L'alchimie expérimentale a même conduit à figer ces G valeurs à une seule constante τ (indépendance du facteur de confiance à la composante gaussienne), fixée par chaque laboratoire au chiffre magique semblant le mieux s'accorder à sa configuration.

Il s'est avéré que la seule adaptation des moyennes suffisait à atteindre les performances optimales et cette technique s'est généralisée dans les systèmes état-de-l'art. L'adaptation MAP fournit alors en sortie une batterie de vecteurs de moyennes adaptées $\widehat{\mu}_g \in \mathbb{R}^F$ (un vecteur par composante gaussienne du GMM). Ces G vecteurs peuvent être alors exprimés, par concaténation, sous forme d'un *supervecteur* s de dimension GF :

$$s = [\widehat{\mu}_1 \widehat{\mu}_2 ... \widehat{\mu}_G] \in \mathbb{R}^{GF} \tag{2.11}$$

Ce supervecteur constitue la statistique d'ordre 1 issue du GMM-UBM. Les procédures succédant à la production de modèles adaptés nécessitent d'ajouter à la représentation la statistique d'ordre 0 des trames du segment considéré. Il s'agit du vecteur n des effectifs de trames associées (probabilistiquement) à chaque gaussienne :

$$n = \left(n_g\right)_{g=1}^{G} \in \mathbb{R}^G \tag{2.12}$$

où n_g, effectif de trames affectées à la gaussienne g, est égal à la somme sur toutes les trames de la collection des probabilité d'occupation de cette gaussienne.

La représentation par adaptation du GMM-UBM fournit donc en sortie, dans sa version la plus courante, une numérisation du signal de parole sous forme vectorielle de dimension $G\,(F+1)$.

2.3.5 Calcul de score

Une fois que des objets à comparer ont été numérisés, il est toujours possible de mesurer leur proximité deux à deux et d'en déduire une décision, lorsqu'une variable latente à déterminer est supposée maintenir proche ces productions. Dans le cas de la reconnaissance du locuteur, l'identité de la personne constitue la variable cachée. La mesure de similarité entre deux représentations (*scoring*) gagne à s'appuyer sur des modèles préliminaires dans lesquels la machine a été renseignée sur l'identité du locuteur et des hypothèses a priori ont été avancées.

Le champ d'investigation des mesures de similarité n'est malheureusement pas borné et nous développerons seulement, dans cette section et par la suite, autour des modèles et scorings actuellement appliqués aux représentations issues du GMM-UBM.

Plusieurs alternatives sont possibles pour comparer segment de test et modèle-locuteur :

– un modèle peut être réalisé à partir du segment de test, puis ce modèle est comparé au modèle du locuteur-cible. La mesure s'effectue alors directement entre les représentations vectorielles issues des collections de trames.
– les trames du segment de test sont présentées une à une au modèle du locuteur-cible et ses comparaisons synthétisées dans une valeur unique. La comparaison peut s'effectuer par mesure de vraisemblance de la trame sur le modèle a priori du locuteur-cible et la moyenne de ces mesures renvoyée comme score. Plutôt qu'une vraisemblance d'une moyenne des trames, c'est la moyenne de leur vraisemblance qui est produite.

Cette seconde alternative, jugée plus précise, a été le plus souvent élue. Dans la logique d'une compétition affrontant modèle-cible et modèle-imposteur, le ratio de vraisemblance de la trame entre ces deux modèles (*Log Likelihood Ratio by Frame, LLR-by-frame*) retournera une mesure plus homogène. Etant donnée la collection $\mathcal{X}$ de trames d'un segment de test et s le locuteur-cible, la mesure de vraisemblance de $\mathcal{X}$ sous l'hypothèse que ses trames sont issues du modèle s s'écrit :

$$P\left(\mathcal{X}|s\right) = \prod_{x \in \mathcal{X}} P\left(x|s\right) \tag{2.13}$$

les trames x de $\mathcal{X}$ étant supposées indépendantes.

Sous l'hypothèse inverse (notée $\overline{s}$), le GMM-UBM est utilisé comme modèle imposteurs et l'on peut écrire :

$$P\left(\mathcal{X}|\overline{s}\right) = P\left(\mathcal{X}|UBM\right) = \prod_{x \in \mathcal{X}} P\left(x|UBM\right) \tag{2.14}$$

Notant n le nombre de trames de la collection $\mathcal{X}$, ces vraisemblances sont normalisées pour obtenir un score homogène, en calculant leur moyenne (géométrique, pour tenir compte d'un produit et non d'une somme). Le score LLR-by-frame est défini par :

$$\begin{aligned} score\left(\mathcal{X}, s\right) &= \log \frac{P\left(\mathcal{X}|s\right)^{\frac{1}{n}}}{P\left(\mathcal{X}|UBM\right)^{\frac{1}{n}}} \\ &= \frac{1}{n}\left\{\sum_{x \in \mathcal{X}}\left(\log \sum_{g=1}^{G} w_g^{(s)} \mathcal{N}\left(x|\mu_g^{(s)}, \mathbf{\Sigma}_g^{(s)}\right) - \log \sum_{g=1}^{G} w_g \mathcal{N}\left(x|\mu_g, \mathbf{\Sigma}_g\right)\right)\right\} \end{aligned} \tag{2.15}$$

où les paramètres $w_g^{(s)}, \mu_g^{(s)}, \Sigma_g^{(s)}$ (resp. w_g, μ_g, Σ_g) sont issus de l'adaptation du GMM-UBM à s (resp. sont ceux du GMM-UBM). La représentation pratique la plus usuelle n'adaptant que la moyenne, ce score se réduit à :

$$score\left(\mathcal{X}, s\right) = \frac{1}{n}\left\{\sum_{x\in\mathcal{X}}\left(\log\sum_{g=1}^{G} w_g\mathcal{N}\left(x|\mu_g^{(s)}, \mathbf{\Sigma}_g\right) - \log\sum_{g=1}^{G} w_g\mathcal{N}\left(x|\mu_g, \mathbf{\Sigma}_g\right)\right)\right\} \quad (2.16)$$

2.3.6 Observations diverses autour des GMM

L'étude détaillée des modèles par mélanges de gaussiennes en reconnaissance du locuteur nous a conduits à plusieurs observations, sur leur nature, leurs propriétés et sur les caractéristiques de leurs métriques induites, que nous présentons dans cette partie.

2.3.6.1 Structuration de l'espace acoustique par le GMM-UBM

Le GMM-UBM ne fait pas que modéliser une hypothèse "non-locuteur". Il structure l'espace acoustique (Scheffer and Bonastre, 2006) en régions de densité élevée, donc en classes probabilistes, par une technique d'apprentissage, hors toute considération phonologique ou linguistique. Le résultat obtenu peut donc être considéré, si la collection est suffisamment vaste et variée, comme une classification probabiliste de l'espace acoustique de la voix humaine enregistrée, d'où son nom d'*Universal Background Model*.

2.3.6.2 Pertinence de la représentation par supervecteur

La représentation par supervecteur relève un point par gaussienne, la moyenne des trames affectées à cette gaussienne, avec relativisation de sa position par maximum à posteriori. La moyenne est choisie parce que ce paramètre est en fait le mode (valeur de plus grande densité) d'une distribution lorsque celle-ci est supposée symétrique. Sous hypothèses gaussiennes, sa manipulation en tant que paramètre de tendance centrale des observations est donc pertinente : le sous-vecteur s_g de la gaussienne g d'un supervecteur s et l'effectif n_g de trames affectées à cette gaussienne résument la distribution de ces trames sur cette gaussienne, assimilée à une loi symétrique de moyenne s_g, de covariance celle du monde $\mathbf{\Sigma}_g$, avec une fiabilité n_g, la quantité d'informations qui a permis de l'estimer.

2.3.6.3 Facteur de confiance

Les études réalisées plus loin (5.2.3) montrent un lien, peu surprenant en fait, entre la quantité d'informations contenue dans chaque dimension de chaque gaussienne (c'est à dire une valeur ponctuelle $\tau_{g,k}, g \in [1, G], k \in [1, F]$) et la dispersion des données pour cette dimension et cette gaussienne (la variance, égale à $\left(\Sigma_g\right)_{k,k}$). La pondération des effectifs entre n_g et le facteur de confiance peut alors s'interpréter comme une forme précoce de *standardisation* des données dans la chaîne de traitement d'un système. Avant

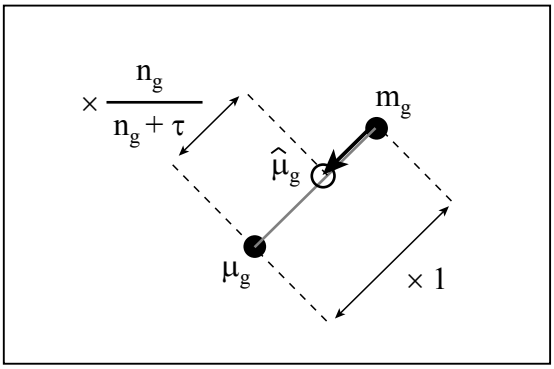

FIGURE 2.1 – *L'adaptation à posteriori MAP relativise l'écart à la moyenne monde, mais préserve la direction.*

les processus de normalisation que nous introduisons dans la suite, ce fait est à remarquer.

2.3.6.4 Adaptation des moyennes

Comme nous l'avons indiqué précédemment, il s'est avéré que la seule adaptation des moyennes suffisait à atteindre les performances optimales. Dans la partie ultérieure 5.3.1, la comparaison des vecteurs d'accumulation de divers modèles nous a amené à produire des systèmes basés sur des adaptations des vecteurs de poids seuls. Les performances, bridées par la taille modeste des vecteurs (G coordonnées au lieu de GF, où F est la dimension de l'espace acoustique), restent tout de même très intéressantes, eu égard au faible volume d'informations fournies.

L'utilisation du facteur de confiance met alors en évidence un aspect, important pour la suite, de l'adaptation de modèles. La formule 2.9 d'adaptation de moyenne conduit à l'expression :

$$\widehat{\mu}_g = \frac{n_g}{n_g + \tau} m_g + \frac{\tau}{n_g + \tau} \mu_g \tag{2.17}$$

Cette expression montre que la moyenne adaptée sur chaque gaussienne $\widehat{\mu}_g$est une combinaison barycentrique de la moyenne μ_g (moyenne *du monde*) et de la moyenne empirique m_g issue des seules trames du segment considéré. Elle peut se reformuler :

$$\widehat{\mu}_g = \mu_g + \frac{n_g}{n_g + \tau} \left(m_g - \mu_g \right) \tag{2.18}$$

qui montre, comme l'indique la figure 2.1, que la statistique d'ordre 1 empirique est relativisée sur chaque gaussienne suivant la quantité des informations qui l'a façonnée.

L'adaptation MAP peut alors se voir comme un ensemble de translations par gaussienne du modèle du monde dans la direction des données du locuteur. Elle fait apparaître l'importance de la direction des données par rapport au paradigme UBM, c'est à dire d'une localisation qui n'est pas prioritairement absolue (euclidienne) mais radiale : MAP se permet de déplacer les données empiriques par rapport au monde, mais conserve intacte leur direction par rapport à ce dernier. Ce point mérite d'être souligné, car nous retrouverons plusieurs fois par la suite la prédominance de la localisation *radiale* sur l'*absolue*, notamment dans la chaîne de traitement des systèmes i-vectors décrite à la section 2.5 de ce chapitre. Cette prédominance se retrouve dans le calcul de score *LLR-by-frame*, comme le montre le paragraphe suivant.

2.3.6.5 Le score LLR-by-frame : proximités absolue et directionnelle

L'analyse de la mesure de proximité du LLR-by-frame n'est pas simple, du fait des sommes de logarithmes intervenant dans la formule. On peut toutefois tirer quelques enseignements du cas simplifié où chaque trame est alignée binairement à une gaussienne. La part associée à cette trame x, associée à la gaussienne g, est :

$$llr = \log\left(w_g \mathcal{N}\left(x|\mu_g^{(s)}, \mathbf{\Sigma}_g\right)\right) - \log\left(w_g \mathcal{N}\left(x|\mu_g, \mathbf{\Sigma}_g\right)\right) \tag{2.19}$$

qui se développe en :

$$llr = -\frac{1}{2}\left(x-\mu_g^{(s)}\right)^t \mathbf{\Sigma}_g^{-1}\left(x-\mu_g^{(s)}\right) + \frac{1}{2}\left(x-\mu_g\right)^t \mathbf{\Sigma}_g^{-1}\left(x-\mu_g\right) + C \tag{2.20}$$

où C est une constante indépendante du test réalisé. La part du score quadratique associée à cette gaussienne se résume à un score linéaire en x :

$$\left(x-\mu_g\right)^t \mathbf{\Sigma}_g^{-1}\left(\mu_g^s-\mu_g\right) - \frac{1}{2}\left(\mu_g^s-\mu_g\right)^t \mathbf{\Sigma}_g^{-1}\left(\mu_g^s-\mu_g\right) \tag{2.21}$$

ou plus simplement

$$= \left\langle x-\mu_g, \mu_g^s-\mu_g\right\rangle_{\Sigma_g^{-1}} - \frac{1}{2}\left\|\mu_g^s-\mu_g\right\|_{\mathbf{\Sigma}_g^{-1}} \tag{2.22}$$

où les notations $\langle\rangle_{\Sigma_g^{-1}}$ et $\|\|_{\Sigma_g^{-1}}$ définissent le produit scalaire et la norme au sens de la matrice de précision $\mathbf{\Sigma}_g^{-1}$.

Pour un locuteur-cible fixé, la décision dépend du produit scalaire entre les deux vecteurs $x-\mu_g$ et $\mu_g^s-\mu_g$. La figure 2.2 illustre ce propos. L'estimation initiale de la moyenne du locuteur avant adaptation MAP est également représentée.

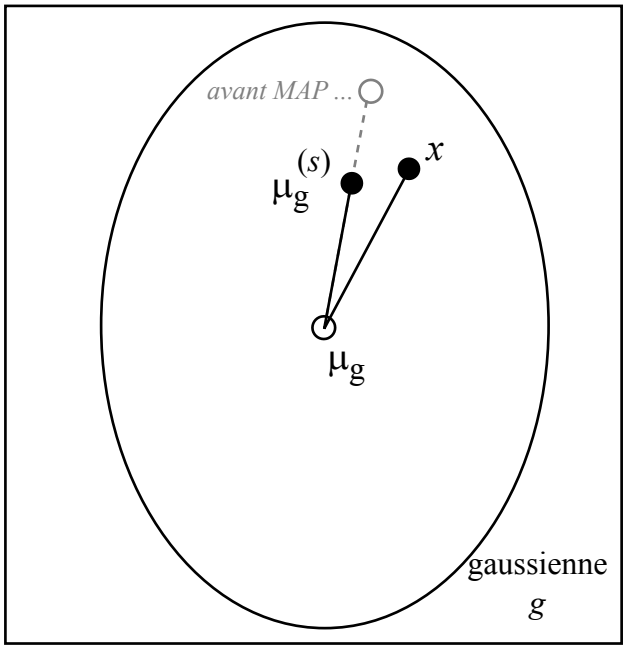

FIGURE 2.2 – *Sur une gaussienne, le score LLR-by-frame doit mesurer la proximité entre la trame de test x et le modèle du locuteur $\mu_g^{(s)}$ adapté par MAP.*

En notant $p_{(\mu_g^s - \mu_g)}(x)$ la projection oblique (suivant Σ_g^{-1}) de la trame x sur l'axe porté par le vecteur $\mu_g^s - \mu_g$, la mesure précédente se ramène à celle de l'abscisse de $p_{(\mu_g^s - \mu_g)}(x) - \mu_g$ sur cet axe.

La figure 2.3 affiche les courbes de niveaux de la mesure de l'équation 2.22. En grisé, sont affichées celles de la même mesure, mais effectuée sur l'estimation de moyenne de s avant adaptation MAP. On voit que :

- les courbes de niveau sont bien les droites correspondant à l'image inverse d'un projeté sur l'axe $\left(\mu_g^s - \mu_g\right)$ suivant la direction oblique de $\mathbf{\Sigma}_g^{-1}$,
- la valeur 0 de non-décidabilité est celle correspondant au milieu de μ_g^s et μ_g. Une valeur positive est obtenue au-delà, dans la direction du modèle de s,
- la valeur tend vers l'infini avec x, mais ce cas ne peut se produire, la gaussienne g n'étant pas alors la plus vraisemblable.

Ce constat est obtenu sur le cas simplifié d'un alignement de trame uni-gaussien. Mais, d'une part, les expériences s'appuyant sur cette configuration approchent les meilleures performances et, d'autre part, l'alignement par gaussienne est, comme nous l'avons indiqué, un pré-supposé du développement mathématique rigoureux de la Factor Analysis, présentée plus bas.

L'adaptation MAP des moyennes ne peut que les rapprocher de la moyenne monde. Ceci entraîne, comme l'indique la comparaison avec les droites grisées, un basculement plus rapide de la mesure dans les positifs (voir courbes 0), mais aussi son tassement en

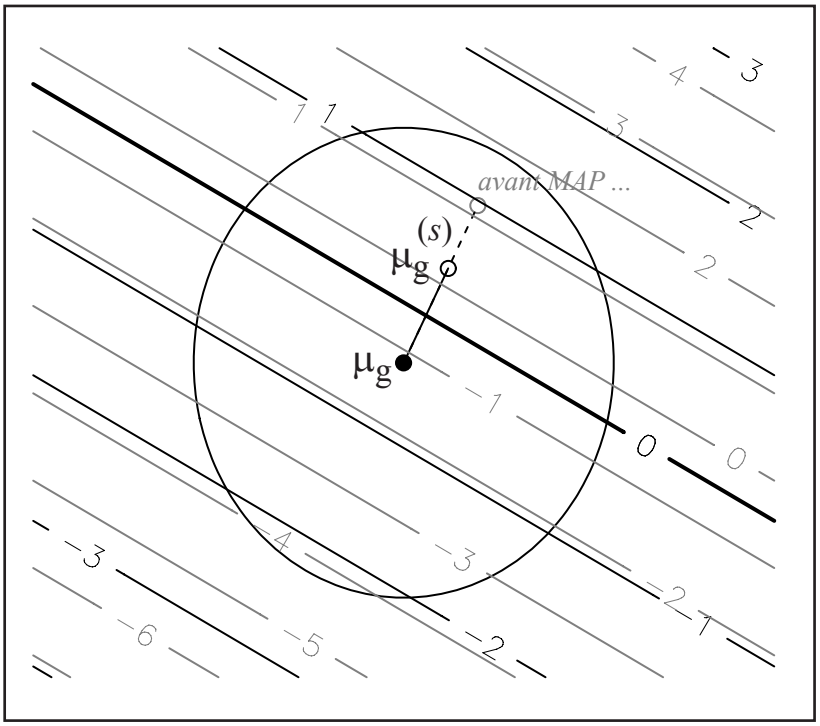

FIGURE 2.3 – *Lignes de niveaux du LLR-by-frame dans le cas simplifié d'alignement binaire de la trame à une gaussienne. En trait fin grisé les valeurs avant adaptation MAP, en trait épais après cette procédure.*

s'éloignant de 0. La mesure est ainsi relativisée.

Ces considérations montrent que la proximité trames / modèle locuteur-cible doit s'entendre moins comme absolue que comme relative à une direction suivant l'axe de la translation $\mu_g^s - \mu_g$. L'adaptation MAP, qui modifie la position absolue, préserve soigneusement cette direction.

Cette importante de la proximité directionnelle doit rester à l'esprit, car elle reste sous-jacente lors de l'analyse, non plus de l'espace des représentations en haute dimension (supervecteurs, ...), mais de sa "version" compressée (i-vectors).

2.4 Décomposition en facteurs et réduction de dimensionnalité

Une fois obtenue la représentation en statistiques d'ordre 0 et 1 issues du GMM-UBM, plusieurs techniques ont été mises en oeuvre pour exploiter ce vecteur dans l'objectif de discrimination des locuteurs.

2.4.1 Les diverses décompositions

La représentation des segments de voix par supervecteur issu de l'adaptation MAP peut s'écrire vectoriellement :

$$s = m + \mathbf{D}z \tag{2.23}$$

où le supervecteur s est l'adaptation du supervecteur des moyennes du monde m par addition d'un terme vectoriel $\mathbf{D}z$ et dans lequel la matrice diagonale $(FG \times FG)$ $\mathbf{D}$ vérifie (Matrouf et al., 2007) :

$$I = \tau \mathbf{D}^t \mathbf{\Sigma}^{-1} \mathbf{D} \tag{2.24}$$

$\mathbf{I}$ étant la matrice-identité et $\mathbf{\Sigma}$ la matrice de covariance dont les blocs diagonaux sont les matrices de covariance diagonales par gaussienne $\mathbf{\Sigma}_g$. Le vecteur z de $\mathbb{R}^{FG}$ constitue le facteur de l'adaptation.

Plus généralement, une représentation vectorielle peut faire l'objet d'une décomposition additive en *facteurs*. Ceux-ci sont des vecteurs censés représenter une partie précise de la variabilité générale. Dans le cas de l'adaptation MAP, il s'agit d'une simple addition à un terme universel, le supervecteur du monde, d'un facteur $\mathbf{D}z$ propre au locuteur (ou au moins à son segment de voix considéré).

Cette notion de décomposition vectorielle additive en facteurs a fait l'objet de nombreuses études, destinées à améliorer la qualité des systèmes s'appuyant sur la représentation en supervecteurs issue du GMM-UBM. La principale avancée a consisté à émettre des hypothèses contraignant la liberté des facteurs, supposés propres à certains

types de variabilité. Ainsi, il a été proposé (Kuhn et al., 1998) de décomposer le supervecteur moyen d'un locuteur, c'est à dire l'espérance de l'ensemble des supervecteurs des segments de ce locuteur qu'on notera $E[s]$, de la manière suivante :

$$E[s] = m + \mathbf{V}y \tag{2.25}$$

où la matrice $\mathbf{V}$ est $FG \times r$, son rang r étant largement inférieur à FG et le facteur y propre au locuteur se restreignant à un vecteur de dimension r réduite.

Cette approche, inspirée par les notions d'*eigenfaces* en reconnaissance faciale (Turk and Pentland, 1991a) (Turk and Pentland, 1991b), conduit la reconnaissance du locuteur vers les techniques de réduction de dimensionnalité (*Dimensionality Reduction Techniques, DRT*). Les colonnes de $\mathbf{V}$ sont les vecteurs propres de la voix (les *eigenvoices)* et le vecteur y le facteur locuteur (*speaker factor*).

L'estimation du supervecteur locuteur moyen $E[s]$ ne pouvant s'appuyer, dans beaucoup de protocoles, que sur un nombre limité d'occurrences de voix (voire une seule), il a paru naturel de proposer une décomposition en facteurs d'une occurrence unique, c'est à dire d'un supervecteur s unique "modèle" de représentation du locuteur. Pour tenir compte de cette variabilité, la décomposition s'écrit :

$$s = m + \mathbf{U}x + \mathbf{D}z \tag{2.26}$$

où apparaît le terme $m + \mathbf{D}z$, modélisant cette fois la part propre au locuteur et un facteur $\mathbf{U}x$ prenant en compte la variabilité de ce locuteur sur la session de voix représentée. Notons que le terme $m + \mathbf{D}z$ s'étend sans contrainte dans l'espace vectoriel de représentation.

Le facteur $\mathbf{U}x$ a été considéré (Teunen et al., 2000) comme attribuable au canal spécifique à l'enregistrement. D'où le nom de canaux propres (*eigenchannels*) pour la décomposition du type précédent (Kenny et al., 2003) (Burget et al., 2007), dans laquelle la matrice $\mathbf{U}$ est rectangulaire de rang faible et le facteur x (*channel factor*) un vecteur de faible dimension.

Les approches précédentes supposent les variabilités locuteur et canal linéairement contraignables. Combinant ces approches, la décomposition additive en facteur devient :

$$s = m + \mathbf{V}y + \mathbf{U}x \tag{2.27}$$

où y (resp. x) est le facteur locuteur (resp. canal) du segment de voix représenté, de dimension réduite.

Enfin, l'expérimentation a montré que cette décomposition idéale pouvait être relativisée en :

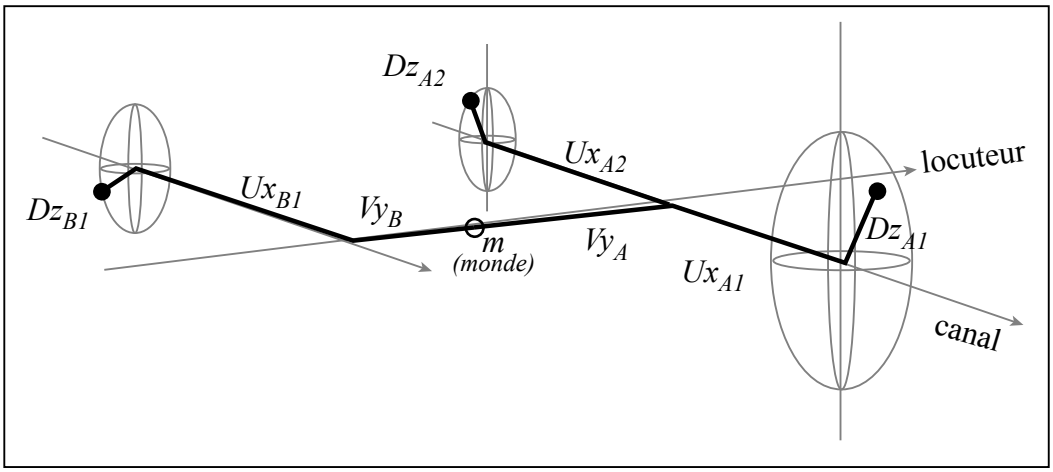

FIGURE 2.4 – *Les hypothèses de la décomposition Joint Factor Analysis en 3D.*

$$s = m + \mathbf{V}y + \mathbf{U}x + \mathbf{D}z \tag{2.28}$$

dans laquelle le supervecteur d'un segment d'un locuteur est considéré comme décomposable en un décalage par rapport au supervecteur du monde, formé d'une composante spécifique au locuteur $\mathbf{V}z$, d'une composante spécifique au canal $\mathbf{U}x$ et enfin d'un résidu $\mathbf{D}z$, ces deux derniers formant la variabilité dite *session*. Le terme $\mathbf{D}z$ respecte deux conditions d'un bruit aléatoire : indépendance entre les dimensions et non compressibilité ($\mathbf{D}$ diagonale et de plein rang).

La figure 2.4 illustre les hypothèses de cette décomposition. Trois supervecteurs s_{A1}, s_{A2}, s_{B1} de deux locuteurs distinct A et B sont représentés en dimension 3. Le facteur locuteur s'étend sur une droite vectorielle. Le facteur canal s'étend sur une seconde droite. Avec les résidus ($\mathbf{D}z_{A1}, \mathbf{D}z_{A2}, \mathbf{D}z_{B1}$), il différencie les segments de voix de ces locuteurs.

Cette formule de décomposition fait clairement apparaître les défis statistiques de la reconnaissance du locuteur : l'objet "signal de voix" contient tout autant des variabilités propres aux caractéristiques physiologiques et psychologiques des locuteurs qu'au contenu phonétique et linguistique de la parole, à la nature du support audiophonique et aux conditions environnementales. Nous verrons dans la section sur les modèles et scorings des i-vectors qu'une telle décomposition réapparaît naturellement dans l'espace de ces nouvelles représentations.

2.4.2 Réduction de dimensionnalité

Les décompositions présentées ci-dessus constituent des réductions de dimension de tout ou partie des variabilités en présence (locuteur, canal). Il faut donc faire appel aux techniques de réduction de dimension pour calculer leurs facteurs. La plus

commune d'entre elles dans le champ des variables numériques est l'analyse en composantes principales (*Principal Component Analysis PCA*), initiée par Pearson et formalisée par Hotelling dans le cadre de l'analyse canonique. La PCA réduit un vecteur par projection orthogonale sur le sous-espace, de dimension fixée a priori, qui maximise la variance des projetés. Dans le cas où la variabilité traitée est la variabilité totale, sa solution exacte est le sous-espace engendré par les premiers vecteurs propres de la matrice de covariance, dans l'ordre décroissant des valeurs propres. Cette méthode garantit une erreur minimale (*reconstruction error*) entre vecteurs initiaux et projetés au sens euclidien du terme (principe de "moindre inertie"). Il s'agit donc d'une méthode géométrique. La PCA peut s'écrire :

$$s = \mathbf{P}x + \varepsilon \tag{2.29}$$

où un vecteur s est réduit à un vecteur x de rang r avec une erreur ε. La matrice rectangulaire $\mathbf{P}$ est orthogonale, contenant les r premiers vecteurs propres de la matrice de covariance.

Le caractère de la représentation par GMM a conduit naturellement à employer des méthodes de réduction probabilistes. La *Factor Analysis* (Bartholomew, 1987) (Tipping and Bishop, 1999a) est une forme d'analyse en composantes principales probabiliste (*Probabilistic Principal Component Analysis PPCA*). Elle postule une décomposition équivalente à celle de la formule précédente, mais avec les contraintes que x suive une loi probabiliste (conventionnellement la loi normale standard $\mathcal{N}(0, \mathbf{I})$) et que l'erreur ε suive une loi probabiliste compatible avec un bruit aléatoire : indépendance et gaussianité des dimensions. Cette dernière hypothèse réduit le risque d'éliminer par compression une part d'informations non aléatoires, donc potentiellement explicatives. De plus, elle s'adapte au caractère inférentiel de notre domaine de reconnaissance où la matrice de projection, apprise sur un échantillon d'apprentissage, sera appliquée sur de nouvelles données. Enfin, ces deux hypothèses séparent les rôles des deux facteurs x et ε : le premier gère les corrélations seules, le second les fluctuations d'échantillonnage. Notons que la matrice de projection $\mathbf{P}$ obtenue n'est plus nécessairement constituée de vecteurs-colonnes orthogonaux (comme c'était le cas en PCA, par la nature symétrique de la matrice de covariance).

L'obtention de la matrice de projection s'effectue par apprentissage EM-ML. Mais dans le cadre de la reconnaissance du locuteur avec représentation par GMM, le supervecteur est en fait une batterie de vecteurs d'un espace commun, pondérés par gaussienne. A partir des travaux de Tipping et Bishop (Tipping and Bishop, 1999b), deux algorithmes spécifiques à ce type de représentation ont été proposés pour déterminer ces facteurs.

Le premier algorithme estime une décomposition à un facteur (nommée par la communauté *Factor Analysis FA)* suivant la formule 2.26 (Kenny et al., 2005) . Le second gère le cas plus complexe de la formule 2.28 (*Joint Factor Analysis JFA*) où deux facteurs hors résidu doivent être estimés par un processus itératif sur critère de maximum de vraisemblance (Kenny et al., 2007) .

Dans les deux cas, l'estimation est réalisée à l'aide d'un vaste fichier d'apprentissage multi-sessions et multi-locuteurs. Celui-ci peut ou non différer de celui du GMM-UBM. Son effectif, bien évidemment, mais aussi les types de canaux qu'il inclut, le nombre minimal d'occurrences par locuteur peuvent influer sur la qualité du résultat.

Dans les deux cas, en estimant à chaque itération les différentes parties successivement et non simultanément, une version simplifiée a été implémentée dans la plupart des laboratoires (Matrouf et al., 2007). De plus, des obstacles mathématiques dans l'expression de la fonction de vraisemblance obligent à un alignement de chaque trame par gaussienne (affectation de la trame à la gaussienne suivant différentes méthodes, comme celle de Viterbi). La non-prise en compte de cette obligation n'a pas altéré les performances des systèmes résultants et cette précaution théorique n'est pas respectée en général.

2.4.3 Scores après décomposition en facteurs FA et JFA

Le choix de la représentation adoptée comme modèle dépend ensuite de la décomposition en facteurs. Il pourra s'agir de soustraire au supervecteur la composante canal $\mathbf{U}x$, dans le cas de la Factor Analysis (Matrouf et al., 2007), ou bien les composantes "session" $\mathbf{U}x$ et $\mathbf{D}z$, dans le cas de la Joint Factor Analysis.

Diverses formules directes, obtenues par intégration d'une fonction objective, ont été proposées dans le cas de la Joint Factor Analysis. Une synthèse de ces variantes de scoring peut être trouvée dans (Glembek et al., 2009) .

2.5 Le concept de i-vectors

Si les décompositions précédentes apparaissaient comme exhaustives (sous l'hypothèse de possible compression des variabilités), elles ont été remises en question ((Dehak, 2009), (Matrouf et al., 2008)). Le modèle suivant a été proposé par (Dehak et al., 2011) :

$$s = m + \mathbf{T}w \tag{2.30}$$

dans lequel la matrice T, rectangulaire et de faible rang, permet de dégager un seul facteur w de basse dimension. Extraire ce facteur réduit, issu d'une compression de toutes les variabilités en jeu (*total variability factor*), repousse le problème de la discrimination du locuteur dans l'espace de ces facteurs, où pourront alors être envisagées, par exemple, de nouvelles décompositions.

Comme il a été effectué et proposé initialement dans (Matrouf et al., 2008), l'extraction d'i-vectors reprend exactement l'algorithme de Factor Analysis (*FA* formule 2.26) mais en considérant le jeu d'apprentissage de n modèles comme issu de n locuteurs différents (la variable locuteur est ignorée).

Les modèles de décompositions en facteurs précédents préservaient soigneusement la représentation du segment de voix en supervecteurs, issue du GMM-UBM. Cette représentation, considérée comme un équivalent mathématique du signal initial, était seulement soumise à des propositions de décomposition vectorielle additives. Dans le cas des i-vectors, le supervecteur s est transformé en un vecteur réduit w, qui plus est par une réduction de dimension drastique, la projection transportant la représentation d'une dimension $G\,(F+1)$ dépassant le plus souvent 25000 à une dimension résultante de 400 à 600. Une part considérable d'informations est éliminée et le facteur w obtenu est envisagé comme une représentation vectorielle de l'ensemble de l'énoncé de voix initial.

Une nouvelle représentation vectorielle du signal vocal méritait mieux qu'un attribut de "facteur" et le nom de *i-vector* a été finalement choisi (pour *intermediate* ou *identity* [1]). Dans (Dehak et al., 2011) (Dehak, 2009) , cette approche est justifiée par les expériences menées sur les facteurs canal et locuteur. Les auteurs remarquent que le premier, censé ne contenir qu'un effet-canal, capture également une partie des informations locuteur. Si la remarque est exacte, elle ne justifie pas pleinement les capacités discriminantes de la décomposition en facteurs de variabilité totale, qui surpasse en performances toutes ses concurrentes. La réduction de dimensionnalité opérée sur les supervecteurs est aveugle à la variable-cible locuteur : c'est une part considérable des informations qui est éliminée, sans égard à leur potentielle valeur explicative. Nous n'avons pas constaté, après réduction de s en w, une augmentation significative de la proportion d'informations attribuables au locuteur : celle-ci peut être évaluée par les parts de variance inter-locuteur dans la variance totale. La justification théorique du pouvoir de ces facteurs reste à faire. Nous avancerons seulement quelques voies d'investigation :

- le supervecteur est en fait la concaténation de G vecteurs de moyenne adaptés par translation des moyennes du monde. Si ces translations par gaussienne présentent un degré suffisant de corrélation (donc si les caractéristiques acoustiques de l'énoncé vocal induisent la non-indépendance entre gaussiennes), elles peuvent être utilement résumées.
- les variabilités secondaires éliminées lors de réduction de dimensionnalité contiennent des informations peu "maîtrisables" par les lois théoriques usuelles (écart trop important aux lois gaussiennes employées en modélisation),
- une fois projetées dans un espace de dimension 400 à 600, les 1000 à 2000 classes-locuteur d'apprentissage (ces nombres sont ceux actuellement disponibles dans les laboratoires) forment un recouvrement moins "clairsemé" qu'en haute dimension, permettant une estimation plus robuste de métaparamètres de variabilité.

1. ou *i-quelque chose* comme nos nouveaux gadgets électroniques ...

2.6 Modèles décisionnels et scorings avec les i-vectors

2.6.1 Outils mathématiques

Nous décrivons en premier lieu quelques objets mathématiques utilisés durant les différentes phases de modélisation et scoring basés sur les i-vectors.

2.6.1.1 Score "log-ratio d'hypothèses complémentaires"

La formule de scoring appliquée aux i-vectors, si elle est dépendante du modèle considéré, a pris ces dernières années une allure (semi-)standardisée. Le score par "log-ratio d'hypothèses complémentaires" dans le cas des i-vectors est présenté au paragraphe suivant.

Etant donnés deux i-vectors w_1,w_2 à comparer dans un problème de discrimination du locuteur, deux hypothèses peuvent être définies :

$$\theta_{\text{tar}} : "w_1 \text{ et } w_2 \text{ appartiennent au même locuteur}" \tag{2.31}$$

$$\theta_{\text{non}} : "w_1 \text{ et } w_2 \text{ n'appartiennent pas au même locuteur}" \tag{2.32}$$

Le scoring par log-ratio de vraisemblance des hypothèses complémentaires s'écrit :

$$score\,(w_1, w_2) = \log \frac{P\,(w_1, w_2|\theta_{\text{tar}})}{P\,(w_1, w_2|\theta_{\text{non}})} \tag{2.33}$$

où $P\,()$ est une probabilité, au sens le plus large du terme. Un tel type de scoring peut être appliqué en général, quelles que soient les hypothèses probabilistes. Notamment la probabilité $P\,(.)$ peut s'appuyer sur différentes lois théoriques, mais aussi sur des estimations non-paramétriques de densité : locales, par voisinage, etc...

Lorsque w_1 et w_2 sont deux observations indépendantes d'un locuteur s connu, il est possible d'écrire :

$$P\,(w_1, w_2|\theta_{\text{tar}}, s) = P\,(w_1|s)\,P\,(w_2|s) \tag{2.34}$$

En vérification du locuteur, la classe est inconnue. La probabilité conjointe peut alors être calculée par :

$$P\,(w_1, w_2|\theta_{\text{tar}}) = \int P\,(w_1, w_2|s)\,dP\,(s) \tag{2.35}$$

De même, la probabilité conjointe de w_1 et w_2 sous l'hypothèse θ_{non} peut être décomposée en :

$$P\left(w_1, w_2|\theta_{\text{non}}\right) = \int P\left(w_1|s\right) dP\left(s\right) \int P\left(w_2|s\right) dP\left(s\right) \tag{2.36}$$

et le score peut s'écrire :

$$score\left(w_1, w_2\right) = \log \frac{\int P\left(w_1|s\right) P\left(w_2|s\right) dP\left(s\right)}{\int P\left(w_1|s\right) dP\left(s\right) \int P\left(w_2|s\right) dP\left(s\right)} \tag{2.37}$$

L'expression finale de ce score dépendra donc des modélisations des variabilités intra- et inter-locuteur qui ont été proposées.

2.6.1.2 Matrices de covariance

Soit $\{w_i\}_{i=1}^{n}$ un jeu de n i-vectors de moyenne μ, appartenant à un ensemble de S locuteurs distincts. On notera par commodité $\{w_i^s\}_{i=1}^{n_s}$ le sous-ensemble des n_s i-vectors d'un locuteur s.

La matrice de covariance totale $\mathbf{\Sigma}$ de ce jeu peut s'écrire :

$$\mathbf{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} \left(w_i - \mu\right) \left(w_i - \mu\right)^t \tag{2.38}$$

Pour tout locuteur s, la matrice de covariance $\mathbf{W}_s$ de ses observations est estimée par :

$$\mathbf{W}_s = \frac{1}{n_s} \sum_{i=1}^{n_s} \left(w_i^s - y_s\right) \left(w_i^s - y_s\right)^t \tag{2.39}$$

où y_s est la moyenne des i-vectors de s.

La variabilité moyenne à l'intérieur des classes-locuteur peut être modélisée par la matrice de covariance intra-locuteur $\mathbf{W}$ (*within-class covariance matrix*). Elle est définie comme la moyenne des matrices de covariance des différentes classes-locuteur, pondérées par leurs effectifs :

$$\mathbf{W} = \sum_{s=1}^{S} \frac{n_s}{n} \mathbf{W}_s \tag{2.40}$$

d'où la formule :

$$\mathbf{W} = \frac{1}{n} \sum_{s=1}^{S} \sum_{i=1}^{n_s} \left(w_i^s - y_s\right) \left(w_i^s - y_s\right)^t \tag{2.41}$$

La variabilité entre les classes-locuteur peut être modélisée par la matrice de covariance inter-locuteur B (*between-class covariance matrix*). Elle est définie comme la matrice de covariance des moyennes (centroïdes) des différentes classes-locuteur pondérées par leur effectif :

$$\mathbf{B} = \sum_{s=1}^{S} \frac{n_s}{n} (y_s - \mu) (y_s - \mu)^t \tag{2.42}$$

Ces deux matrices modélisent les variabilités imputables à la dispersion interne des classes-locuteur et à la dispersion entre ces classes. La décomposition des variances de Huyghens (ou des inerties en fait) assure que :

$$\mathbf{\Sigma} = \mathbf{B} + \mathbf{W} \tag{2.43}$$

Deux autres matrices sont utilisées dans le domaine. Les matrices de dispersion (*scatter matrix*) intra-locuteur $S_{\mathbf{W}}$ et inter-locuteur $S_{\mathbf{B}}$ sont définies par :

$$S_{\mathbf{W}} = \sum_{s=1}^{S} \frac{1}{n_s} \sum_{i=1}^{n_s} (w_i^s - y_s) (w_i^s - y_s)^t \tag{2.44}$$

$$S_{\mathbf{B}} = \sum_{s=1}^{S} (y_s - \mu) (y_s - \mu)^t \tag{2.45}$$

L'analogie avec les matrices précédentes est évidente. La différence réside dans le fait que les matrices ne sont pas divisées par un effectif (elles mesurent une dispersion totale et pas moyenne) et surtout qu'elles mettent "à égalité" les classes-locuteur en terme d'effectifs. Ainsi, le poids d'une classe-locuteur sera égalisé à la valeur $\frac{1}{S}$, indépendamment de son nombre de sessions d'apprentissage disponibles, traduisant l'hypothèse d'équiprobabilité a priori d'apparition d'un locuteur dans un système de RAL.

Nous verrons que ces matrices de dispersion se retrouvent dans différents cadres du champ d'application. Leur comparaison avec **B** et **W** sera notamment réalisée dans le cadre précis du modèle "LDA-Two covariance".

La matrice de dispersion intra-classes $S_{\mathbf{W}}$, qui ne prend pas en compte les fréquences a priori des classes-locuteur, peut être moyennée pour former une matrice de covariance intra-locuteur. Pour ne pas la confondre avec **W**, elle est notée $\mathbf{W}_{[\text{WCCN}]}$ et vaut :

$$\mathbf{W}_{[\text{WCCN}]} = \frac{1}{S} S_{\mathbf{w}} = \frac{1}{S} \sum_{s=1}^{S} \frac{1}{n_s} \sum_{i=1}^{n_s} (w_i^s - y_s) (w_i^s - y_s)^t \tag{2.46}$$

Introduite par Andrew Hatch (Hatch et al., 2006) dans le contexte des classifieurs SVM (qui l'utilisent dans le cadre d'une technique dite de *Within Class Covariance Normalization WCCN*), elle est reprise dans les scorings initiaux des i-vectors mis en place par N. Dehak (Dehak et al., 2011) que nous présentons plus bas.

Les indicateurs de dispersion globale d'un partitionnement en classes sont les variances totale, intra- ou inter-locuteur $\sigma^2_{\mathbf{\Sigma}}$, $\sigma^2_{\mathbf{B}}$ et $\sigma^2_{\mathbf{W}}$ égales aux traces des matrices respectives $\mathbf{\Sigma}$, $\mathbf{W}$ et $\mathbf{B}$. Elles vérifient :

$$\sigma^2_{\mathbf{\Sigma}} = \sigma^2_{\mathbf{B}} + \sigma^2_{\mathbf{W}} \tag{2.47}$$

Cette relation est utile pour estimer la proportion de variabilité imputable aux écarts entre classes locuteurs (inter-locuteur). Cette proportion, appelée R^2 en classification, est égale à :

$$R^2 = \frac{\sigma^2_{\mathbf{B}}}{\sigma^2_{\mathbf{\Sigma}}} \tag{2.48}$$

La part imputable aux fluctuations internes des classes (intra-locuteur) est égale à :

$$1 - R^2 = \frac{\sigma^2_{\mathbf{W}}}{\sigma^2_{\mathbf{\Sigma}}} \tag{2.49}$$

Le R^2, compris dans $[0, 1]$, est un indicateur du degré de séparation entre les classes. La valeur $1 - R^2$ mesure le degré de compacité des classes. Une valeur élevée du R^2 indique une bonne disposition des vecteurs à être discriminés suivant leur classe locuteur.

2.6.1.3 Analyse discriminante linéaire (Linear Discriminant Analysis LDA)

Cette méthode de réduction de dimension projette les données dans un sous-espace exaltant la variabilité inter-locuteur et limitant la variabilité intra-locuteur (Rao, 1948). Notons avant tout que le terme de LDA s'emploie également pour une méthode de scoring d'appartenance à une classe, qui rejoint en fait l'analyse discriminante géométrique de Fisher. Nous réserverons dans la suite le terme de LDA à la méthode de réduction de dimension.

La LDA procède d'une manière comparable à la PCA : les données sont projetées sur un sous-espace qui optimise de manière déterministe un critère de variance des projetés. Cette fois, le critère doit prendre en compte un double objectif : maximiser la variance inter-classes des projetés et minimiser leur variance intra-classes.

Or, si ces variances sont initialement liées par la loi de décomposition des variances de Huyghens, deux sous-espaces de projection distincts optimisent séparément les deux

critères. Il est proposé de lever ce problème d'optimisation à deux contraintes en maximisant le critère de Rayleigh :

$$\arg\max_{v} J(v) = \frac{v^t \mathbf{B} v}{v^t \mathbf{W} v} \tag{2.50}$$

Les valeurs $v^t \mathbf{B} v$ et $v^t \mathbf{W} v$ mesurent les dispersions inter- et intra-locuteur des projetés. La valeur maximale de ce critère fournit le vecteur v portant le meilleur axe de projection.

Le sous-espace de rang r optimisant le critère peut être déterminé par la matrice $\mathbf{A}^{[opt]}$ des r vecteurs-colonnes de sa base, qui vérifie :

$$\mathbf{A}^{[opt]} = \arg\max_{\mathbf{A}} \frac{\left|\mathbf{A}^t \mathbf{B} \mathbf{A}\right|}{\left|\mathbf{A}^t \mathbf{W} \mathbf{A}\right|} \tag{2.51}$$

où $|...|$ dénote le déterminant.

La solution exacte de ce problème est obtenue par le zéro de la différentielle du critère de Rayleigh. Il s'agit du sous-espace porté par les premiers vecteurs propres de $\mathbf{W}^{-1}\mathbf{B}$, dans l'ordre décroissant des valeurs propres. La qualité de la projection peut être évaluée par l'évolution du critère de Rayleigh entre l'espace initial et le sous-espace optimal.

Dans le domaine de la discrimination du locuteur, comme dans un certain nombre d'autres, l'optimisation du critère est souvent effectuée sur les matrices de dispersion $S_{\mathbf{B}}$ et $S_{\mathbf{W}}$ plutôt que $\mathbf{B}$ et W, ce pour les raisons invoquées plus haut (homogènéisation des fréquences d'apparition a priori des locuteurs). Nous noterons par la suite $\text{LDA}_{\mathbf{B},\mathbf{W}}$ et $\text{LDA}_{S_{\mathbf{B}},S_{\mathbf{W}}}$ les LDA pratiquées sur l'un ou l'autre de ces couples de matrices.

Un grand nombre de variantes de la LDA ont été proposées, par exemple, la technique de *Heteroscedastic Discriminant Analysis* qui permet de prendre en compte des matrices de covariance distinctes par classe (Kumar and Andreou, 1998) . Dans notre domaine et dans le cas des i-vectors, le lecteur pourra se référer à (McLaren and Leeuwen, 2011) (Kanagasundaram et al., 2012) . La technique de LDA que nous faisons suivre par le modèle two-covariance est parfois aussi utilisée comme pré-processus à la PLDA décrite plus bas.

2.6.2 Modèle LDA + WCCN-cosine-scoring

Dans (Dehak et al., 2011), N. Dehak propose un certain nombre de méthodes de scoring. La plus efficace est une méthode enchaînant une réduction de dimension par LDA et un scoring de type cosinus entre les i-vectors issu de travaux initiaux autour des SVM (Dehak et al., 2009)(Dehak et al., 2010). Etant donnés deux i-vectors w_1 et w_2 (ou leur projection par LDA), ce score s'écrit :

$$score\left(w_1, w_2\right) = \cos_{W_{WCCN}^{-1}}\left(w_1, w_2\right) \tag{2.52}$$

Le cosinus s'entend ici au sens de la métrique de $\mathbf{W}_{WCCN}^{-1}$. Ce score s'écrit analytiquement :

$$score\left(w_1, w_2\right) = \frac{w_1^t \mathbf{W}_{WCCN}^{-1} w_2}{\sqrt{w_1^t \mathbf{W}_{WCCN}^{-1} w_1}\sqrt{w_2^t \mathbf{W}_{WCCN}^{-1} w_2}} = \frac{\left(\mathbf{W}_{WCCN}^{-\frac{1}{2}} w_1\right)^t \mathbf{W}_{WCCN}^{-\frac{1}{2}} w_2}{\left\|\mathbf{W}_{WCCN}^{-\frac{1}{2}} w_1\right\| \left\|\mathbf{W}_{WCCN}^{-\frac{1}{2}} w_2\right\|} \tag{2.53}$$

ou plus brièvement :

$$score\left(w_1, w_2\right) = \frac{\left(\mathbf{L}w_1\right)^t \mathbf{L}w_2}{\|\mathbf{L}w_1\| \, \|\mathbf{L}w_2\|} \tag{2.54}$$

La matrice L est obtenue par décomposition de Cholesky de la matrice $\mathbf{W}_{WCCN}^{-1} = \mathbf{L}\mathbf{L}^t$. Cette opération justifie le terme *WCCN* (le *N* pour *Normalization*). Partant d'un cosinus euclidien, celui-ci prend en compte la nécessité de réduire la variabilité intra-session. Après avoir réduit la dimension par LDA, l'application de la matrice $\mathbf{W}_{[\text{WCCN}]}$ dans la formule de score permet d'estimer une proximité angulaire, suivant l'orientation de la covariance intra-classes.

2.6.3 Modèle et scoring de Mahalanobis

Nous avons introduit dans la reconnaissance du locuteur par i-vector le score suivant (Bousquet et al., 2011b) : en l'absence de toute information paramétrique sur la constitution des classes, la moyenne-classe la plus vraisemblable à posteriori de deux observations w_1 et w_2 d'un même locuteur est le vecteur qui minimise la somme des écarts aux observations. Il s'agit de leur milieu $\frac{1}{2}\left(w_1 + w_2\right)$. Sous l'hypothèse de gaussianité de la variabilité intra-locuteur, l'estimation ponctuelle de la distribution de classe la plus probable est alors la loi normale $\mathcal{N}\left(\frac{1}{2}\left(w_1 + w_2\right), \mathbf{W}\right)$ où $\mathbf{W}$ est la matrice de covariance intra-locuteur. Ceci peut s'écrire :

$$P\left(w_1, w_2 | \theta_{\text{tar}}\right) = P\left(w_1, w_2 | \mathcal{N}\left(\frac{w_1 + w_2}{2}, \mathbf{W}\right)\right) \tag{2.55}$$

Les vecteurs w_1 et w_2 étant supposés indépendants sachant leur classe, on obtient :

$$P\left(w_1, w_2 | \theta_{\text{tar}}\right) = \mathcal{N}\left(w_1 | \frac{w_1 + w_2}{2}, \mathbf{W}\right) \mathcal{N}\left(w_2 | \frac{w_1 + w_2}{2}, \mathbf{W}\right) \tag{2.56}$$

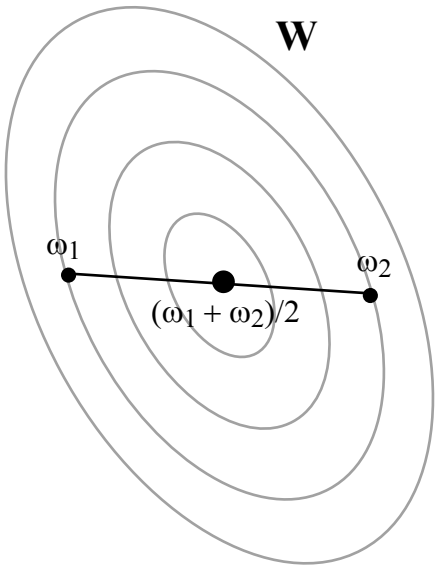

FIGURE 2.5 – *Illustration du modèle de Mahalanobis. En l'absence d'information sur la variabilité locuteur, le milieu des deux vecteurs w_1 et w_2 est la moyenne-classe la plus vraisemblable à posteriori de deux observations.*

Les deux facteurs sont égaux. La log-probabilité est égale à :

$$\begin{aligned}\log P\left(w_1, w_2|\theta_{\text{tar}}\right) &= C_1 - 2\left(w_1 - \frac{w_1+w_2}{2}\right)^t \mathbf{W}^{-1}\left(w_1 - \frac{w_1+w_2}{2}\right) \\ &= C_1 - C_2\frac{1}{2}\left(w_1 - w_2\right)^t \mathbf{W}^{-1}\left(w_1 - w_2\right) \end{aligned} \tag{2.57}$$

où C_1,C_2 sont des constantes indépendantes de w_1 et w_2. Le score proposé est issu de la probabilité précédente :

$$score_{\text{Maha}}\left(w_1, w_2\right) = -\frac{1}{2}\left(w_1 - w_2\right)^t \mathbf{W}^{-1}\left(w_1 - w_2\right) \tag{2.58}$$

La matrice inverse de **W** est souvent appelée matrice de Mahalanobis, du nom de son initiateur(Mahalanobis, 1936). Il s'agit d'une matrice de précision intra-locuteur.

La figure 2.5 illustre cette modélisation. Les courbes de densité de **W** correspondent pour les deux vecteurs, si la moyenne du locuteur présumé est leur milieu (point le plus vraisemblable en l'absence d'information sur la variabilité des facteurs-locuteurs).

Dans le cas d'un classifieur, lorsque les vecteurs w_1,w_2 sont supposés appartenir à une classe d'un locuteur d'apprentissage, la métrique de Mahalanobis s'avère souvent très efficace (McLachlan, 1999). Nous avons montré son optimalité sous certaines hypothèses (Bousquet et al., 2011a). Ses applications sont nombreuses, notamment en

traitement d'image, reconnaissance des caractères (Kato et al., 1999) mais aussi de la parole (Plannerer, 2005) .

Appuyé sur des hypothèses gaussiennes, le score de Mahalanobis rejoint la solution géométrique, en se lisant comme une mesure de distance entre deux vecteurs au sens de la métrique de Mahalanobis :

$$score_{\text{Maha}}(w_1, w_2) = -\frac{1}{2} \|w_1 - w_2\|^2_{\mathbf{W}^{-1}} \tag{2.59}$$

Il peut être aussi considéré comme une mesure d'adéquation de la différence de deux vecteurs à une distribution intra-classe estimée a priori par $\mathbf{W}$.

Remarques :

- le score obtenu ne paraît pas être un log-ratio d'hypothèses complémentaires. Il l'est pourtant, comme nous le montrons au paragraphe suivant.
- le modèle de Mahalanobis suppose qu'on ne dispose d'aucune information sur la distribution probabiliste des classes-locuteur. C'est pourquoi il ne fait pas intervenir la matrice $\mathbf{B}$ inter-classes. Ceci se traduit mathématiquement par une matrice de précision inter-locuteur $\mathbf{B}^{-1}$ égale à 0.

2.6.4 Modèle et scoring LDA-Two-covariance

Le modèle et scoring LDA-Two-covariance combine :

- une réduction de dimension dans l'espace des i-vectors par LDA,
- une modélisation des variabilités intra- et inter-locuteur et un scoring résultant, introduits dans le domaine par (Brummer and de Villiers, 2010) à partir des travaux de (Bishop, 2006) .

Le modèle à double-covariance (*Two-Covariance Model*) est un modèle génératif simple, linéaire et Gaussien, basé sur la décomposition présumée de tout i-vector w d'un locuteur s en :

$$w = y_s + \varepsilon \tag{2.60}$$

où le modèle locuteur y_s, de même dimension que w, est la part propre au locuteur de w et ε est un bruit gaussien, avec les hypothèses (en dénotant par commodité $\mathcal{N}(|)$ la densité d'une loi normale multivariée) :

$$P(y_s) = \mathcal{N}(\mu, \mathbf{B}) \tag{2.61}$$

$$P(w|y_s) = \mathcal{N}(y_s, \mathbf{W}) \tag{2.62}$$

La densité de y_s suit une loi normale de moyenne μ et de covariance la matrice inter-locuteur B et la densité conjointe de w sachant y_s suit une loi normale de moyenne y_s et de covariance la matrice intra-locuteur W.

Les deux hypothèses précédentes peuvent s'écrire, pour tout i-vector w d'un locuteur s :

$$y_s = \mu + \mathbf{B}^{\frac{1}{2}}\eta \text{ où } \eta \sim \mathcal{N}(0, \mathbf{I}) \tag{2.63}$$

$$w = y_s + \mathbf{W}^{\frac{1}{2}}\varepsilon \text{ où } \varepsilon \sim \mathcal{N}(0, \mathbf{I}) \tag{2.64}$$

Le vecteur w se décompose en :

$$w = \mu + \mathbf{B}^{\frac{1}{2}}\eta + \mathbf{W}^{\frac{1}{2}}\varepsilon \tag{2.65}$$

où η et ε suivent des lois normales standards. Ce modèle constitue donc une décomposition en facteurs locuteur et session dans l'espace des i-vectors.

Le score log-ratio d'hypothèses complémentaires, sous sa formulation de l'équation 2.37, devient :

$$score_{2Cov}(w_1, w_2) = \log \frac{\int_y \prod_{i=1,2} \mathcal{N}(w_i|y, \mathbf{W})\, P(y|\mu, \mathbf{B})\, dy}{\prod_{i=1,2} \int_y \mathcal{N}(w_i|y, \mathbf{W})\, P(y|\mu, \mathbf{B})\, dy} \tag{2.66}$$

L'expression explicite de ce score est donnée dans (Brummer and de Villiers, 2010). Nous détaillons son calcul à l'annexe B. Après centrage des données d'apprentissage, la moyenne μ est égale à 0 et son écriture devient :

$$score(w_1, w_2) = -\frac{1}{2}\left\{w_1^t \mathcal{P} w_1 + w_2^t \mathcal{P} w_2 + 2 w_1^t \mathcal{Q} w_2\right\} \tag{2.67}$$

où

$$\mathcal{P} = \mathbf{W}^{-1}\left\{\left(\mathbf{B}^{-1} + 2\mathbf{W}^{-1}\right)^{-1} - \left(\mathbf{B}^{-1} + \mathbf{W}^{-1}\right)^{-1}\right\}\mathbf{W}^{-1} \tag{2.68}$$

$$\mathcal{Q} = \mathbf{W}^{-1}\left(\mathbf{B}^{-1} + 2\mathbf{W}^{-1}\right)^{-1}\mathbf{W}^{-1} \tag{2.69}$$

La figure 2.6 illustre ce calcul de score. Pour deux facteurs-locuteur y et y', dotés d'une vraisemblance normale de covariance $\mathbf{B}$, les vraisemblances des deux observations vis à vis de $\mathbf{W}$ sont chaque fois calculées. Ces vraisemblances sont sommées sur l'ensemble des facteurs-locuteur pondérés par leur vraisemblance suivant $\mathcal{N}(\mu, \mathbf{B})$.

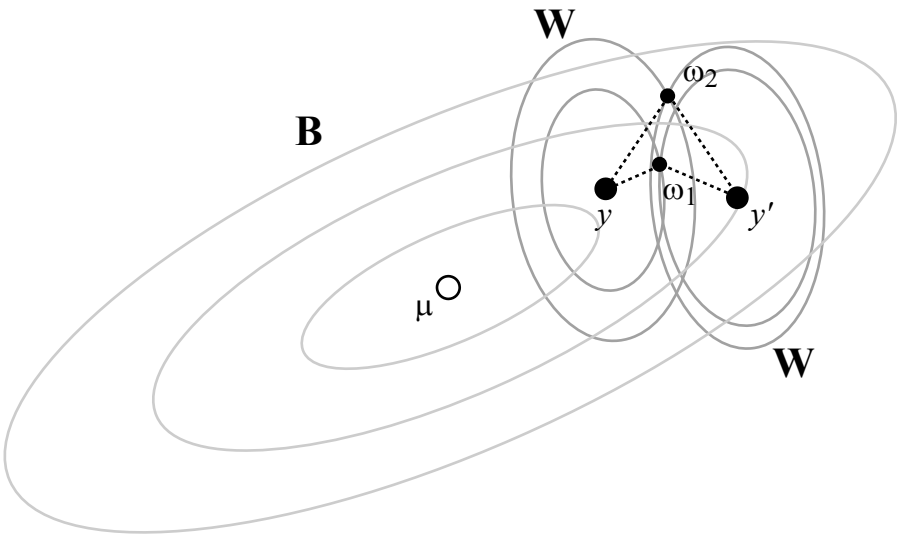

FIGURE 2.6 – *Illustration du modèle two-covariance, pour deux observations w_1 et w_2 et deux exemples de facteurs-locuteur y et y' présumés.*

Alors que le score de Mahalanobis n'émet aucune hypothèse sur la distribution statistique des modèles locuteurs y_s centroïdes des classes (estimant donc le centroïde de la classe commune aux deux i-vecteurs par leur milieu), le score two-covariance parcourt l'espace statistiquement distribué des centroïdes pour accumuler les vraisemblances des vecteurs.

Cette démarche est évidemment plus précise, mais à la condition de disposer d'une estimation de cette distribution suffisamment robuste, donc de $\mathbf{B}$. Son rang est nécessairement inférieur au nombre S de locuteurs d'apprentissage. En dimension 400 à 600 de l'espace des i-vectors, une valeur S de 1000, voire plus de 2000, et éventuellement un nombre minimal de 5 ou 6 observations par locuteur, sont les conditions habituellement employées pour tenter de satisfaire aux exigences de robustesse.

Lien avec le modèle de Mahalanobis

Comme indiqué précédemment, l'absence d'hypothèses sur la distribution statistique des centres de classes équivaut à une matrice de précision inter-locuteur $\mathbf{B}^{-1}$ égale à 0. L'écriture des matrices $\mathcal{P}$ et $\mathcal{Q}$ de l'équation 2.69 précédente devient :

$$\mathcal{P} = \mathbf{W}^{-1}\left\{\frac{1}{2}\mathbf{W} - \mathbf{W}\right\}\mathbf{W}^{-1} = -\frac{1}{2}\mathbf{W}^{-1} \tag{2.70}$$

$$\mathcal{Q} = \frac{1}{2}\mathbf{W}^{-1} \tag{2.71}$$

et le score entre deux i-vectors w_1 et w_2 s'écrit :

$$score\,(w_1, w_2) = \frac{1}{2}\left\{-\frac{1}{2}w_1^t\mathbf{W}^{-1}w_1 - \frac{1}{2}w_2^t\mathbf{W}^{-1}w_2 + w_1^t\mathbf{W}^{-1}w_2\right\} \tag{2.72}$$

$$= -\frac{1}{4}\left(w_1 - w_2\right)^t\mathbf{W}^{-1}\left(w_1 - w_2\right) \tag{2.73}$$

où l'on retrouve, à un facteur multiplicatif près, l'expression de l'équation 2.58.

LDA et modèles précédents

Les modèles de Mahalanobis et Two-covariance peuvent être précédés d'une LDA. Ces deux modèles et cette technique de réduction de dimensionnalité s'appuient en effet sur les mêmes estimations matricielles déterministes de variabilité $\mathbf{B}$ et $\mathbf{W}$. Les scores des deux modèles mesurent des distances suivant les métriques de $\mathbf{W}^{-1}$ et $\mathbf{B}^{-1}$, favorisant les observations proches au sens de la première (similarité intra-locuteur) et pénalisant les observations lointaines au sens de la seconde (dissimilarité inter-locuteur). La LDA ne conserve que les axes prédominants de ces variabilités et concentre ainsi la mesure de similarité sur ses axes les plus significatifs.

2.6.5 Modèles PLDA

2.6.5.1 PLDA gaussienne (G-PLDA)

Introduite par S. Prince dans le cadre de la reconnaissance faciale (Prince and Elder, 2007), l'analyse discriminante linéaire probabiliste (*Probabilistic Linear Discriminant Analysis, PLDA*) élargit le modèle double-covariance, qui en est un cas particulier, au cas probabiliste et à celui de la décomposition en facteurs. Proposée par P. Kenny en reconnaissance du locuteur (Kenny, 2010), elle peut être vue comme un équivalent de la Joint Factor Analysis dans l'espace des i-vectors, c'est à dire dans un espace monogaussien.

Ce modèle est génératif. Il ignore le mécanisme d'extraction des i-vectors pour les considérer comme des observations d'un modèle génératif probabiliste. Il fait l'hypothèse que tout i-vector w de dimension p peut être décomposé comme :

$$w = \mu + \mathbf{\Phi}y_s + \mathbf{\Gamma}z + \varepsilon \tag{2.74}$$

Ce modèle est formé de deux parties :

(i) la composante $\mu + \mathbf{\Phi}y_s$ ne dépend que du locuteur,

(ii) la composante $\mathbf{\Gamma}z + \varepsilon$ diffère pour chaque session de voix et représente la "variabilité" intra-locuteur.

La matrice $\mathbf{\Phi}$ est rectangulaire, à r_{voix} colonnes ($r_{\text{voix}} < p$) fournissant une base pour le sous-espace locuteur des voix propres (*eigenvoices*). De même, Γ est rectangulaire, à

r_{canal} colonnes ($r_{\text{canal}} < p$) fournissant une base pour le sous-espace "canal". des canaux propres (*eigenchannels*). Les facteurs y_s et z suivent des lois normales standards. Enfin, le terme résiduel ε est supposé être gaussien avec une moyenne de 0 et une matrice de covariance diagonale Σ.

Le cas particulier $r_{\text{canal}} = p$ ($\mathbf{\Gamma}$carrée de plein rang) est équivalent à la version de la PLDA proposée par (Kenny, 2010). Les *eigenchannels* sont retirées de l'équation 2.74 et le bruit résiduel σ est supposé avoir une matrice de covariance pleine. Le modèle PLDA devient :

$$w = \mu + \mathbf{\Phi} y_s + \varepsilon \tag{2.75}$$

où $\mathbf{\Phi}$ est une matrice $p \times r$ ($r < p$) et ε est un vecteur de dimension p avec une matrice $\mathbf{\Sigma}$ de covariance pleine.

Après estimation des métaparamètres de la PLDA, le score log-ratio de vraisemblance appliqué sur ce modèle est le résultat d'une proposition de S. Prince, qui considère le sur-vecteur concaténé de deux vecteurs w_1 et w_2 à comparer. L'équation générative 2.75 sous l'hypothèse θ_{tar} s'écrit :

$$\begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = \begin{bmatrix} \mu \\ \mu \end{bmatrix} + \begin{bmatrix} \mathbf{\Phi} & \mathbf{\Gamma} & \mathbf{0} \\ \mathbf{\Phi} & \mathbf{0} & \mathbf{\Gamma} \end{bmatrix} \begin{bmatrix} y \\ z_1 \\ z_2 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \end{bmatrix} \tag{2.76}$$

et sous l'hypothèse θ_{non} :

$$\begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = \begin{bmatrix} \mu \\ \mu \end{bmatrix} + \begin{bmatrix} \mathbf{\Phi} & \mathbf{0} & \mathbf{\Gamma} & \mathbf{0} \\ \mathbf{0} & \mathbf{\Phi} & \mathbf{0} & \mathbf{\Gamma} \end{bmatrix} \begin{bmatrix} y \\ z_1 \\ z_2 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \end{bmatrix} \tag{2.77}$$

Le modèle PLDA est illustré sur la figure 2.7. Le sous-espace locuteur (une droite vectorielle engendrée par $\mathbf{\Phi}$ sur la figure) est distribué suivant une loi de moyenne μ et covariance $\mathbf{\Phi}\mathbf{\Phi}^t$. Pour un facteur-locuteur y, les vraisemblances de deux vecteurs w_1, w_2, sous l'hypothèse qu'ils sont issus de ce locuteur, sont évaluées par la loi de moyenne y et covariance $\mathbf{\Gamma}\mathbf{\Gamma}^t$.

Dans le cas de la PLDA gaussienne (*Gaussian-PLDA*), les vraisemblances marginales sont supposées gaussiennes et le score peut être évalué analytiquement (Prince and Elder, 2007). L'élégance de l'expression obtenue tient au fait qu'elle n'inclut pas les variables cachées y_s et z, qui n'ont donc pas besoin d'être calculées. Le numérateur du ratio de l'équation 2.33 est égal à :

$$P\left(w_1, w_2 | \theta_{\text{tar}}\right) = \mathcal{N}\left(\begin{bmatrix} w_1 \\ w_2 \end{bmatrix} \mid \begin{bmatrix} \mu \\ \mu \end{bmatrix}, \begin{bmatrix} \mathbf{\Phi}\mathbf{\Phi}^t + \mathbf{\Gamma}\mathbf{\Gamma}^t + \mathbf{\Sigma} & \mathbf{\Phi}\mathbf{\Phi}^t \\ \mathbf{\Phi}\mathbf{\Phi}^t & \mathbf{\Phi}\mathbf{\Phi}^t + \mathbf{\Gamma}\mathbf{\Gamma}^t + \mathbf{\Sigma} \end{bmatrix} \right) \tag{2.78}$$

et son dénominateur à :

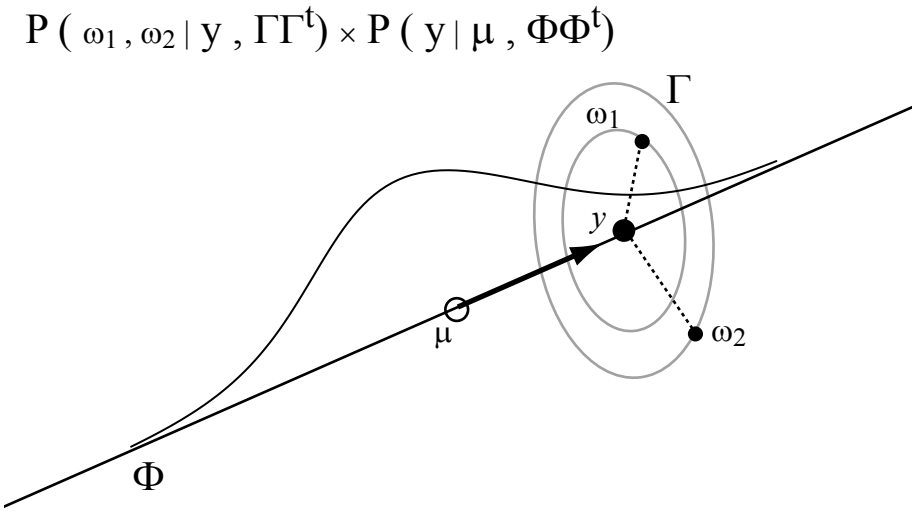

FIGURE 2.7 – *Illustration du modèle PLDA. Le sous-espace locuteur est distribué suivant une loi* $(\mu, \mathbf{\Phi\Phi}^t)$ *et la distribution intra-locuteur suivant.une loi* $(y, \mathbf{\Gamma\Gamma}^t)$.

$$P\left(w_1, w_2|\theta_{\text{non}}\right) = \mathcal{N}\left(\left[\begin{array}{c} w_1 \\ w_2 \end{array}\right] \mid \left[\begin{array}{c} \mu \\ \mu \end{array}\right], \left[\begin{array}{cc} \mathbf{\Phi\Phi}^t + \mathbf{\Gamma\Gamma}^t + \mathbf{\Sigma} & \mathbf{0} \\ \mathbf{0} & \mathbf{\Phi\Phi}^t + \mathbf{\Gamma\Gamma}^t + \mathbf{\Sigma} \end{array}\right]\right) \tag{2.79}$$

Le score log-ratio de vraisemblance s'écrit donc (en retirant une constante) :

$$score\left(w_1, w_2\right) = -\left[\begin{array}{c} w_1 - \mu \\ w_2 - \mu \end{array}\right]^t \left(\mathbf{M}_{\text{tar}}^{-1} - \mathbf{M}_{\text{non}}^{-1}\right) \left[\begin{array}{c} w_1 - \mu \\ w_2 - \mu \end{array}\right] \tag{2.80}$$

où

$$\mathbf{M}_{\text{tar}} = \begin{bmatrix} \mathbf{\Phi\Phi}^t + \mathbf{\Gamma\Gamma}^t + \mathbf{\Sigma} & \mathbf{\Phi\Phi}^t \\ \mathbf{\Phi\Phi}^t & \mathbf{\Phi\Phi}^t + \mathbf{\Gamma\Gamma}^t + \Sigma \end{bmatrix} \tag{2.81}$$

$$\mathbf{M}_{\text{non}} = \begin{bmatrix} \mathbf{\Phi\Phi}^t + \mathbf{\Gamma\Gamma}^t + \mathbf{\Sigma} & \mathbf{0} \\ \mathbf{0} & \mathbf{\Phi\Phi}^t + \mathbf{\Gamma\Gamma}^t + \Sigma \end{bmatrix} \tag{2.82}$$

Dans le cas simplifié de l'équation 2.75, le score s'écrit

$$score\left(w_1, w_2\right) = -\left[\begin{array}{c} w_1 - \mu \\ w_2 - \mu \end{array}\right]^t \left(\mathbf{N}_{\text{tar}}^{-1} - \mathbf{N}_{\text{non}}^{-1}\right) \left[\begin{array}{c} w_1 - \mu \\ w_2 - \mu \end{array}\right] \tag{2.83}$$

où

$$\mathbf{N}_{\text{tar}} = \begin{bmatrix} \mathbf{\Phi}\mathbf{\Phi}^t + \mathbf{\Sigma} & \mathbf{\Phi}\mathbf{\Phi}^t \\ \mathbf{\Phi}\mathbf{\Phi}^t & \mathbf{\Phi}\mathbf{\Phi}^t + \mathbf{\Sigma} \end{bmatrix} \tag{2.84}$$

$$\mathbf{N}_{\text{non}} = \begin{bmatrix} \mathbf{\Phi}\mathbf{\Phi}^t + \mathbf{\Sigma} & \mathbf{0} \\ \mathbf{0} & \mathbf{\Phi}\mathbf{\Phi}^t + \mathbf{\Sigma} \end{bmatrix} \tag{2.85}$$

où $\mathbf{\Phi}$ est une matrice rectangulaire et la matrice carrée $\mathbf{\Sigma}$ contient la variabilité résiduelle non-informative.

Des écritures simplifiées de cette formule ont été proposées (Burget et al., 2006) (Garcia-Romero and Espy-Wilson, 2011). Mais notons que la nature non inversible, en général, de la matrice $\mathbf{\Phi}\mathbf{\Phi}^t$ (si $\mathbf{\Phi}$ n'est pas carrée) n'est pas prise en compte dans les calculs de (Garcia-Romero and Espy-Wilson, 2011).

Lien avec le modèle two-covariance

Comme l'indique l'équation 2.65, le modèle two-covariance est un cas particulier de la PLDA gaussienne, dans lequel :

- les matrices locuteur et résidu sont estimées de manière déterministe,

- la matrice $\mathbf{\Phi}\mathbf{\Phi}^t$ est égale à $\mathbf{B}$ et de plein rang,

- la matrice $\mathbf{\Gamma}\mathbf{\Gamma}^t$ est égale à $\mathbf{W}$ et de plein rang,

- la matrice $\mathbf{\Sigma}$ est nulle, la variabilité nuisible étant entièrement estimée par $\mathbf{W}$.

Il est à noter que le score issu du modèle PLDA peut également être obtenu par une voie mathématique analogue à celle de l'équation 2.66 du modèle two-covariance :

$$score_{PLDA}\left(w_1, w_2\right) = \log \frac{\int_y \prod_{i=1,2} P\left(w_i | \mu + \mathbf{\Phi} y, \mathbf{\Lambda}\right) P\left(y | \mathbf{0}, \mathbf{I}\right) dy}{\prod_{i=1,2} \int_y P\left(w_i | \mu + \mathbf{\Phi} y, \mathbf{\Lambda}\right) P\left(y | \mathbf{0}, \mathbf{I}\right) dy} \tag{2.86}$$

Ce score, apparemment distinct de celui de l'équation 2.83, aboutit strictement au même résultat, les intégrations pouvant être interprétées comme des convolutions de gaussiennes.

2.6.5.2 PLDA "heavy-tailed" (HT-PLDA)

Kenny a proposé (Kenny, 2010) une nouvelle approche de la PLDA en reconnaissance du locuteur par i-vectors. Il constate un certain manque d'adéquation de ces vecteurs au postulat gaussien et tente de le prendre en compte dans la modélisation. La distribution des i-vectors d'apprentissage présente une queue de distribution plus importante que sous hypothèse gaussienne (excès de valeurs extrêmes *outliers*). L'alternative à la PLDA gaussienne consiste à supposer les facteurs locuteur et résiduels comme

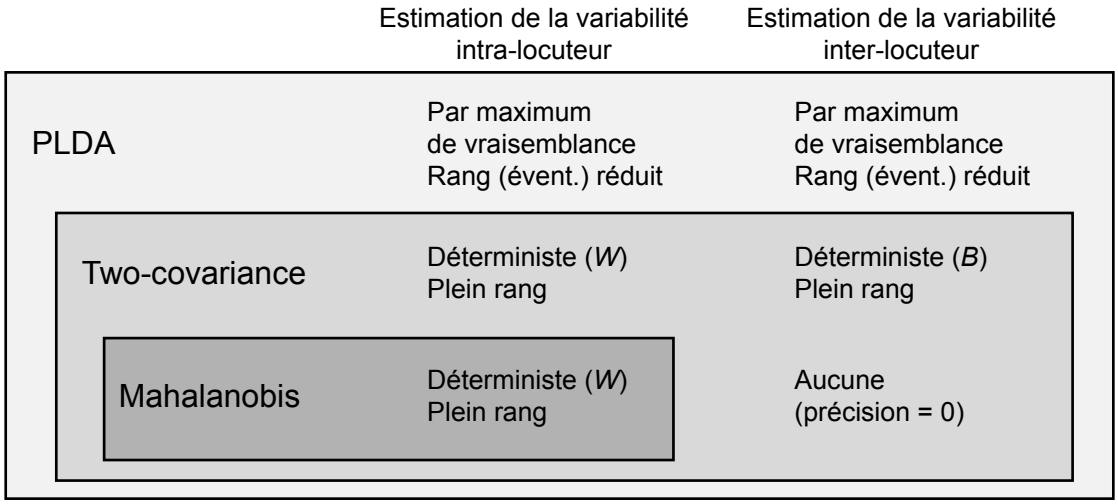

FIGURE 2.8 – *Hiérarchie des modèles suivant leur niveau de sophistication probabiliste.*

distribués suivant une loi du t de Student. Ces lois présentent une queue de distribution plus forte et permettent une meilleure représentation des valeurs extrêmes. Les facteurs y_s et z de l'équation 2.74 sont représentés par :

$$\begin{aligned} y_s &\sim \mathcal{N}\left(0, u_1^{-1}\mathbf{I}\right) \text{ où } u_1 \sim G\left(\frac{n_1}{2}, \frac{n_1}{2}\right) \\ z &\sim \mathcal{N}\left(0, v_r^{-1}\mathbf{\Lambda}^{-1}\right) \text{ où } v_r \sim G\left(\frac{\nu}{2}, \frac{\nu}{2}\right) \end{aligned} \tag{2.87}$$

où n_1 et ν sont les degrés de liberté, u_1 et v_r sont les facteurs de la distribution Gamma et $G(a, b)$ est la distribution Gamma de paramètres de forme a et d'échelle b.

Il est à noter que l'estimation des paramètres de la HT-PLDA est beaucoup plus gourmande en complexité que celle des paramètres de la G-PLDA.

2.6.6 Commentaires et détails sur les méthodes

Les différents modèles et scorings que nous avons décrits présentent une organisation hiérarchique dans le traitement de la problématique de discrimination locuteurs. Comme l'illustre la figure 2.8, les modèles successifs sont, dans l'ordre croissant de sophistication probabiliste :

- le modèle de Mahalanobis, qui s'appuie sur une estimation déterministe et de plein rang de la seule variabilité résiduelle (la variabilité inter-locuteur étant de précision nulle),
- le modèle two-covariance, qui s'appuie sur les estimations déterministes et de plein rang des variabilités résiduelle et locuteur,

– le modèle PLDA gaussienne, qui estime ces métaparamètres en tentant de maximiser leur vraisemblance à posteriori des observations, sous hypothèse de normalité standard des facteurs.

Ce classement hiérarchique n'est pas forcément celui des performances croissantes : l'estimation globale de la variabilité inter-locuteur pose parfois des problèmes de manque de données. Contrairement à la variabilité session pour laquelle on dispose d'échantillons de très grande taille, l'effectif de classes locuteur (1000 à 2000 actuellement), donc de centres de classes pour estimer la variabilité inter-locuteur, reste certes supérieur à la dimension de l'espace, fournissant ainsi une matrice **B** non singulière, mais relativement modeste. Dans certains cas -notamment celui de la segmentation en locuteurs d'un document audio- le faible nombre de locuteurs peut faire du score de Mahalanobis la seule alternative.

Les expressions détaillées des scores LDA-Mahalanobis et LDA-two-covariance, lorsque des techniques de normalisation ont été préalablement appliquées aux i-vectors, sont présentées et analysées à l'annexe G.

Apprentissage des modèles

L'apprentissage des métaparamètres des différents modèles précédents gagne à être effectuée sur un vaste jeu de données, éventuellement dépendant du genre, contenant ou pas les données d'apprentissage du GMM-UBM et comprenant un nombre suffisant de sessions mais aussi de locuteurs distincts (en gardant à l'esprit la dimension 400 à 600 des i-vectors). Le dernier fichier du LIA produit pour l'évaluation NIST-SRE 2012 (Bousquet et al., 2012c) utilise des sessions de NIST-SRE 2004, 2005, 2006, de Switchboard II Phases 2 et 3, Cellular Parts 1 et 2, téléphone seulement, de longueur nominale supérieure à 180 secondes, avec 6 occurrences au moins par locuteur. Les effectifs sont de 21475 sessions de 1575 locuteurs pour les hommes, de 27155 sessions de 2012 locuteurs pour les femmes.

Ces effectifs de sessions permettent, en dimension 400 de l'espace des i-vectors, de produire des estimations relativement fiables de la variabilité intra-locuteur. Par contre, le nombre de classes-locuteur, de l'ordre de 1500 à 2000, constitue un recouvrement limité de l'espace de représentation, pour l'estimation précise de la variabilité extrinsèque au locuteur. Cette limitation se retrouve dans l'ensemble des configurations actuelles des laboratoires, étant inhérente aux données disponibles.

Autres modèles et scorings

D'autres modèles et scorings sur les i-vectors ont été mis en place. Nous citerons les scorings reposant sur le classifieur Machine à Vecteurs Support (SVM) : (Scheffer et al., 2011) où les i-vectors sont projetés par le SVM, la distance étant estimée à l'aide d'un noyau-cosinus (*cosine kernel*) et l'approche récente *Pairwise SVM* (Cumani et al., 2013).

2.7 Performances successives de systèmes basés sur le GMM-UBM

Nous présentons ici, à titre de comparaison, les performances de systèmes basés sur les représentations issues des mixtures de gaussienne avec GMM-UBM, qui ont chronologiquement pris le titre d'état-de-l'art dans cette gamme de systèmes de VAL. Elles sont données pour une configuration du LIA qui est détaillée dans le paragraphe suivant.

- Le premier système utilise un GMM-UBM, la représentation par statistiques du GMM (supervecteurs) adaptées par MAP et un scoring LLR-by-frame dans lequel le modèle imposteur est approximé par le GMM-UBM. Il est noté **MAP** en abrégé.
- Le second est basé sur les avancées de la Factor Analysis. Il soustrait le terme attribué au canal $\mathbf{U}x$ au supervecteur et applique le scoring LLR-by-frame. Il est noté **FA**.
- Le troisième est la meilleure fusion obtenue au LIA, en terme de performance, d'un système du type du précédent et d'un système par classifieur SVM (Larcher et al., 2010). Une normalisation des scores par les techniques z- ou zt-norm (Auckenthaler et al., 2000) a été également réalisée. Il constitue donc la meilleure solution dans notre laboratoire avant l'apparition des i-vectors. Il est noté **FA+SVM**.
- Le quatrième utilise la solution i-vector proposée par son initiateur N. Dehak : une réduction de dimension par LDA suivie du scoring WCCN-cosinus. Il est noté **IVECT+COS**.

Pour évaluer plus complètement cette dernière solution, les résultats obtenus par N. Dehak lors de sa présentation originelle des i-vectors sont également présentés et commentés.

Protocole expérimental

Le GMM-UBM utilisé est celui du LIA codé GMM-UBM-LIA dont le détail est donné en annexe A.

L'expérience de vérification du locuteur est conduite sur l'évaluation NIST-SRE 2008 détaillée est en annexe A. Les résultats sont donnés en terme d'EER et de DCF minimale, suivant les coûts de fausse alerte et faux rejet imposés par NIST durant cette campagne 2008.

Seuls les résultats des expériences sur les locuteurs mâles sont présentés ici. Ceux obtenus avec les femmes conduisent au même conclusion. De même, ces conclusions sont généralisables, ayant été constatées dans l'ensemble des laboratoires.

Pour le système basé sur un SVM, la matrice de variabilité intra-locuteur est apprise sur la base NIST-SRE-2004 de 2938 exemples de 124 locuteurs (20 itérations pour converger). De la même base, 200 locuteurs imposteurs sont utilisés pour la normalisation des scores et comme exemples "négatifs" pour le classifieur SVM. Dans le troisième système, qui fusionne deux approches (GMM-UBM FA et SVM), les scores sont normalisés avant calibration, par zt-norm (Auckenthaler et al., 2000) pour le premier

	EER %	DCF min
MAP	**7.74**	0.0354
FA	**3.87**	0.0189
FA+SVM	**2.72**	0.0154
IVECT+COS	**3.26**	0.0187

TABLE 2.1 – *Performances, en terme d'EER, de différents systèmes état-de-l'art successifs évalués sur la condition "téléphone- téléphone anglais natifs det 7" de NIST-SRE 2008 short2-short3.*

et z-norm pour le second. La calibration est effectuée par régression logistique (*Linear Logistic Regression LLR*) à l'aide du toolkit FoCal de Niko Brümmer [2].

La matrice de variabilité totale **T** du système i-vector est celle codée T15660-LIA dont le détail est en annexe A.

Comparaison de performance

La table 2.1 affiche les performances des quatre systèmes. Le premier (**MAP**) a été grandement amélioré par les hypothèses de décomposition en facteurs réduits de P. Kenny (**FA**). La fusion d'un tel système et d'un classifieur SVM atteint la meilleure performance.

Le dernier système, basé sur les i-vectors, n'améliore pas le troisième, mais dépasse le second. C'est à dire que la stratégie de facteurs *total variability* s'avère plus judicieuse que celle des facteurs séparés. Qui plus est, cette stratégie ramène entièrement les calculs de décision dans un espace de faible dimension, laissant espérer l'application de méthodes complexes, impraticables dans la haute dimension d'origine.

Le troisième système reste pour l'instant meilleur que la solution i-vector mais il s'agit d'une fusion de systèmes et, de plus, les scores y ont été normalisés. L'EER de 3.26% obtenu au LIA avec un mono-système par i-vectors, sur des scores bruts, laisse augurer une progression des performances avec cette solution. L'amélioration de ces performances passe par l'élaboration de nouvelles modélisations et formules de scores dans l'espace des facteurs de variabilité totale.

Commentaires sur les résultats originels

Dans (Dehak et al., 2009), les auteurs comparent les performances de la JFA du laboratoire CRIM de Montréal avec le système i-vector qu'ils ont mis en place. Nous présentons dans la table 2.2 ces résultats, mais également ceux du LIA et ceux du laboratoire I2R Singapour avec qui nous avons travaillé.

Les résultats concernent les deux conditions téléphone-téléphone de NIST-SRE 2008 :

- locuteurs d'anglais natif seulement (codée det 7 par NIST, il s'agit de l'expérience précédente),

- tous locuteurs (codée det 6 par NIST, 12511 tests dont 874 tests-cible).

2. http ://www.dsp.sun.ac.za/~nbrummer/focal/index.htm

		téléphone anglais natif (det 7)		téléphone tous (det 6)	
		EER %	DCF min	EER %	DCF min
LIA	best	2.72	0.0154	6.29	0.0357
LIA	ivect	3.26	0.0187	6.29	0.0335
I2R	best	2.93	0.0116	5.96	0.0319
I2R	ivect	3.00	0.0162	5.23	0.0284
CRIM	JFA	2.64	0.0111	5.15	0.0273
BUT	best w	2.28	0.0104	5.11	0.0267
CRIM	ivect	1.12	0.0094	4.48	0.0247

TABLE 2.2 – *Performances, en terme d'EER, obtenues par différents systèmes de CRIM. La condition est "téléphone- téléphone anglais natifs det 7" NIST-SRE 2008 short2-short3.*

Pour LIA et I2R, **best** indique les meilleures performances obtenues avant i-vectors. Il s'agit dans les deux cas de fusions de systèmes. Puis les performances obtenues par un système unique basé sur les i-vectors, avec le score précédent et sans normalisation des scores, sont indiquées (**ivect**). Pour le laboratoire CRIM, le système noté **JFA** est une *Joint Factor Analysis* détaillée dans (Dehak et al., 2009). Au passage sont insérés pour comparaison les résultats obtenus par le laboratoire tchèque BUT, en fusionnant des systèmes normalisés de type FA ou SVM (Burget et al., 2009) (**best w**). Ces derniers résultats sont les meilleurs de la littérature, avant la solution i-vectors.

Pour le LIA et I2R, le mono-système i-vector approche les meilleurs multi-systèmes (notamment I2R les dépassent sur la condition "téléphone tous det 6"). Pour la configuration CRIM, les résultats i-vectors outrepassent ceux d'une JFA, mais également ceux des meilleurs multi-systèmes de l'époque (**best w** de BUT), démontrant définitivement la prédominance du concept i-vectors sur les décompositions en facteurs antérieures. Mais ce résultat doit être relativisé : le gain exceptionnel de performance procuré par le scoring WCCN-cosinus n'a été reconstitué ni au LIA ni à I2R. Le même phénomène a été plus ou moins constaté dans la plupart des laboratoires qui se sont penchés sur le nouveau concept d'i-vectors. Ce n'est pas au volume ou à la qualité des bases d'entraînement mais bien au manque de robustesse de la méthode de scoring que cette faiblesse est attribuable.

Cette double conclusion : "- potentiel discriminant incontestable des facteurs i-vectors, - gain de performance exceptionnel ... mais non complètement reproductible" a conduit aux différents modèles et scorings que nous avons décrits précédemment. Mais l'analyse a montré l'aspect primordial de certaines transformations, que nous introduisons et décrivons en détail dans le chapitre 3, pour assurer la réussite de ces méthodes dans le champ des i-vectors.

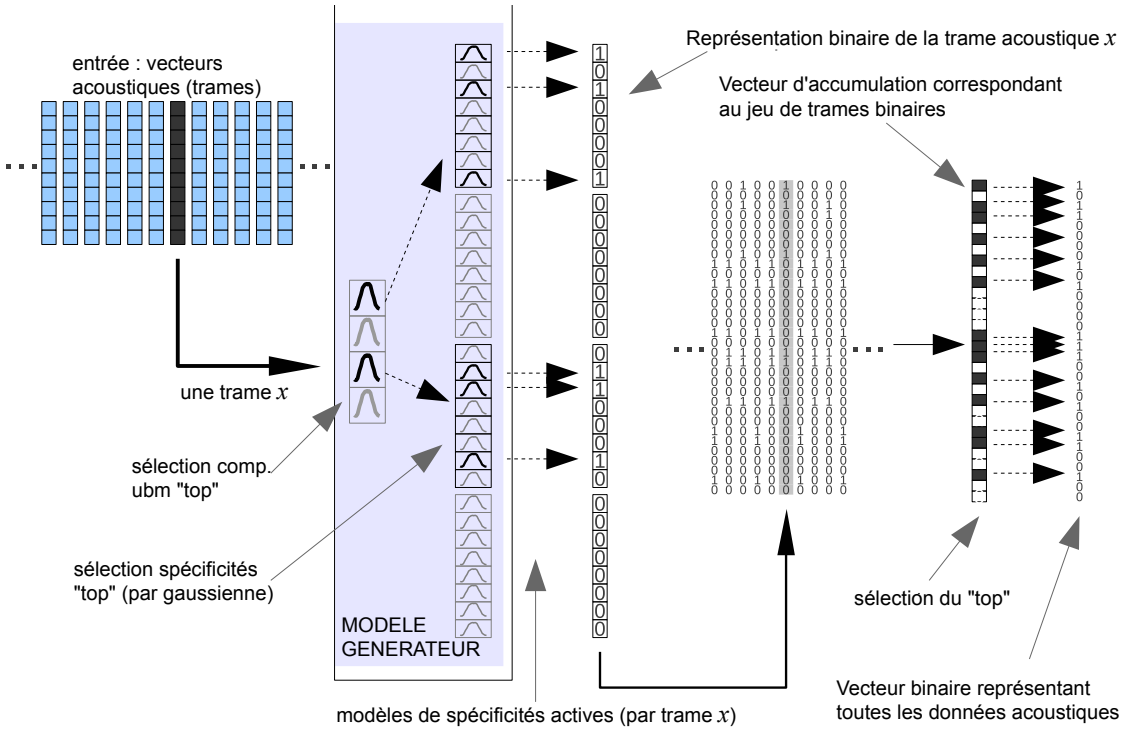

FIGURE 2.9 – *Stratégie générale du modèle de clés binaires du locuteur.*

2.8 Modélisations par clés binaires

Introduit par J.F. Bonastre (Bonastre et al., 2011b) (Bonastre et al., 2011a), le modèle de représentation de locuteurs par clés binaires (*Speaker Binary Keys*) propose une alternative semi-paramétrique aux adaptations de paramètres MAP issues d'un GMM-UBM.

Le modèle s'appuie sur la robustesse du GMM-UBM, mais la stratégie de représentation diffère de celle par supervecteurs. Nous présentons d'abord une description de ce modèle, puis développons sur la "philosophie" et l'originalité de la démarche.

2.8.1 Extraction de clés binaires

La figure 2.9 illustre les étapes de la méthode. Etant donné un GMM-UBM à G composantes et une collection de trames $\mathcal{X}$ à représenter (en haut à gauche de la figure 2.9), chaque trame x de $\mathcal{X}$ est traitée de la manière suivante :

– les vraisemblances de la trame vis à vis de chaque composante sont évaluées, d'une manière similaire à la construction d'un supervecteur. Il s'agit des densités normales $\mathcal{N}\left(x|\mu_g, \mathbf{\Sigma}_g\right)$ de la trame par rapport aux paramètres $\mu_g, \mathbf{\Sigma}_g$ du monde. Les gaussiennes les plus vraisemblables sont sélectionnées. Le nombre de gaussiennes retenues dans ce document est un paramètre fixé.

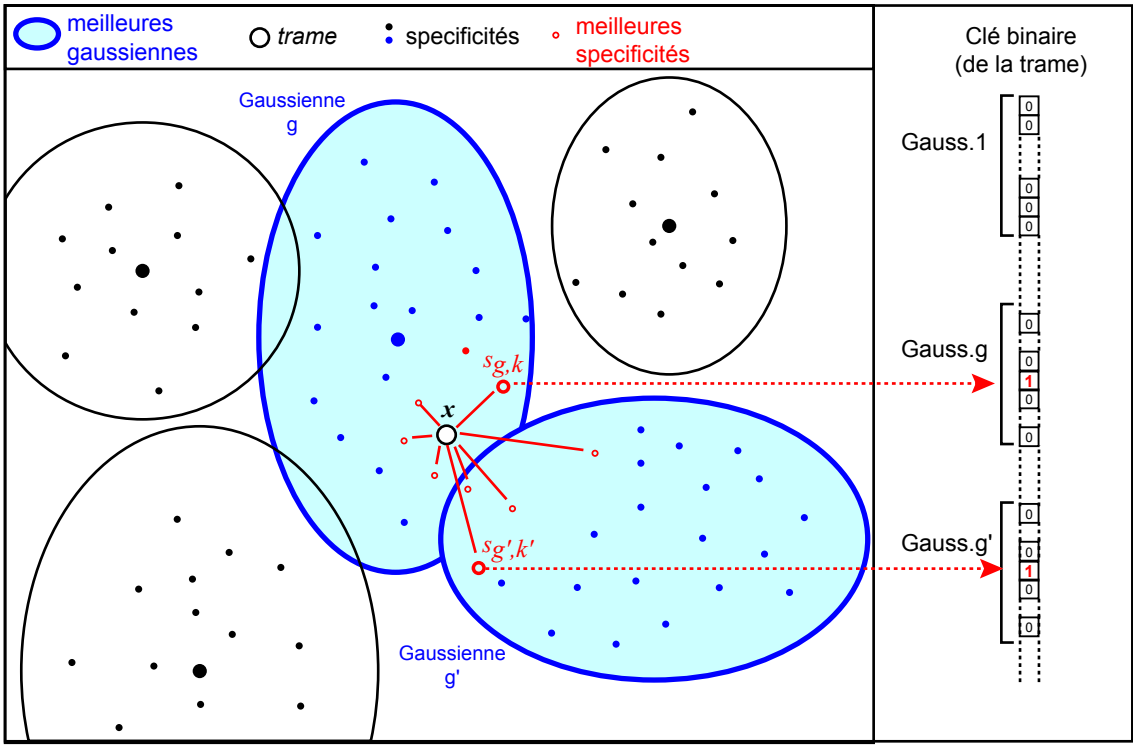

FIGURE 2.10 – *Un exemple en 2D de sélection de spécificités.*

– sur chacune d'entre elles, un "recouvrement" a été préalablement mis en place, constitué de q points dans l'espace acoustique que nous appelons **spécificités**. Ils forment un ensemble noté $\{s_{g,k}\}_{k=1}^{q}$ sur chaque gaussienne g. Dans cette thèse, le nombre q est également un paramètre prédéfini.
– les vraisemblances de la trame x vis à vis de chaque spécificité sont calculées, suivant les densités normales $\mathcal{N}\left(x|s_{g,k}, \mathbf{\Sigma}_g\right)$. Ces spécificités sont les modes de lois normales, de même covariance que celle du monde.

– les spécificités les plus vraisemblables sont sélectionnées. Le nombre de spécificités retenues par gaussienne est également un paramètre fixé a priori. La figure 2.10 détaille cette opération en dimension 2 : sur les 2 gaussiennes g et g' les plus vraisemblables, les 4 spécificités les plus vraisemblables de x sont retenues.

Un vecteur de clés binaires de la trame x, dénoté $b\left(x\right)$ et de dimension $N = Gq$, est initialisé à 0. Ce vecteur est affiché à droite de la figure 2.10. Les clés correspondants aux spécificités sélectionnées sont alors "allumées" (leur valeur devient 1).

Le passage de la trame dans le modèle génère donc un modèle de spécificités actives, sous forme d'un vecteur binaire de dimension N contenant un nombre fixé a priori de valeurs à 1.

Pour aboutir à une représentation de dimension fixe de la collection de trames de taille variable, les vecteurs binaires de l'ensemble des trames du segment sont additionnés terme à terme, produisant un vecteur d'accumulation c de dimension Gq :

$$c_{g,k} = \sum_{x \in \mathcal{X}} b\left(x\right)_{g,k} \tag{2.88}$$

Ce vecteur d'accumulation est à coefficients entiers. Chacun de ses coefficients est un décompte des appels à une spécificité (on parlera de vecteur de *comptes*). Il constitue une représentation dans un espace discret du segment de voix initial.

Binarisation du vecteur de comptes : une représentation sous forme de vecteur binaire est produite à partir de ce vecteur de comptes : pour les n valeurs maximales du vecteur de comptes (n étant un paramètre fixé a priori), la valeur du vecteur binaire est passée de 0 à 1. La représentation du segment de voix initial prend donc l'allure mathématique d'un vecteur binaire à N valeurs, dont n non nulles, obtenu par seuillage d'un vecteur de comptes discret.

Notons que la valeur d'un coefficient du vecteur binaire doit se lire non pas comme un nombre, mais comme une modalité d'une variable qualitative : "*Cette spécificité est sélectionnée pour représenter ce segment de voix*".

2.8.2 Recouvrement

L'ensemble des spécificités des composantes gaussiennes constitue un *recouvrement* de l'espace. Les spécificités choisies sur chaque composante doivent résumer la variabilité possible des trames acoustiques sur sa région de densité. Ayant contribué à l'élaboration d'algorithmes de recouvrement, cette partie est détaillée dans la section 8.

2.8.3 Scoring

Le scoring naturel proposé est un opérateur ET logique entre deux vecteurs binaires à comparer. Ceci correspond, sur des modalités de spécificité "sélectionnée" ou "non-sélectionnée" codées 1 et 0, à un produit scalaire entre les vecteurs binaires. Etant données deux clés binaires b et b', le score s'écrit :

$$score\left(b, b'\right) = b\ ET\ b' = \sum_{g=1}^{G} \sum_{k=1}^{N} b_{g,k}.b'_{g,k} \tag{2.89}$$

Ce score peut être normalisé en divisant sa valeur par le nombre total de spécificités N, ou par son maximum, égal à n.

2.8.4 Objectifs du modèle de clés binaires : le concept d'exceptions

Le modèle de clés binaires localise d'abord chaque trame suivant le régionnement du GMM-UBM. Il relève ensuite les points de densité les plus vraisemblables par gaussienne : il peut s'agir de sous-régions de la gaussienne à forte densité, mais aussi de zones plus isolées. La méthode suit une philosophie en partie inverse de celle utilisée par les statistiques d'ordre 0 et 1 : elle retient des pics de haute densité, mais surtout de haute énergie, c'est à dire dont la dérivée première de la densité apparaît significativement élevée. Plusieurs points peuvent être retenus par composante et non plus un seul, comme c'est le cas dans le modèle par supervecteur. D'autre part, la notion de point de "haute densité" doit se comprendre localement, c'est à dire au voisinage de ce point.

Alors que les méthodes statistiques relèvent par accumulation les éléments les plus représentés, aboutissant le plus souvent à un *mode* de densité maximale, la méthode de sélection de clés met également en évidence des *exceptions*. Celles-ci sont susceptibles de caractériser un segment de voix et de là un locuteur. Ce concept sera développé plus complètement dans la section 8 de nos contributions au modèle de clés binaires, en particulier une opération d'égalisation des comptes qui met en avant les exceptions d'un énoncé.

Deuxième partie

Contributions

Chapitre 3

Analyse statistique de la représentation

Aussitôt proposé le concept i-vectors de facteur réduit de variabilité totale, il nous est apparu indispensable d'effectuer une analyse statistique de cette nouvelle représentation, pour tenter d'en expliquer les capacités comme d'en améliorer l'efficacité. Le concept i-vector mérite en effet, autant qu'il le nécessite, une plus grande maîtrise de ses propriétés théoriques, de ses hypothèses et de mieux cerner ses comportements, en particulier vis à vis de la variable-cible locuteur, mais aussi ses limites.

Partant d'un constat d'insuffisance de la décomposition en facteurs des supervecteurs (le facteur locuteur de la JFA contenant une part non négligeable de variabilité session (Dehak, 2009)), leur réduction suivant la variabilité totale s'est avérée plus à même de discriminer les locuteurs, sans toutefois que cette faculté ne soit justifiée par d'autres arguments que l'efficacité de certaines techniques (LDA, cosine-scoring). De même, la décomposition en facteurs telle que la PLDA dans ce nouvel espace compact de représentation ne s'est avérée pertinente qu'une fois redressées des anomalies au postulat gaussien, par la modélisation HT-PLDA (Kenny, 2010). La question du cadre probabiliste à proposer pour cette nouvelle représentation ne nous a pas semblé réglée par cette avancée. Les points suivants ont plus particulièrement attiré notre attention, motivant et guidant l'analyse que nous avons menée :

- la meilleure efficacité de la loi de Student par rapport à la loi gaussienne pour la PLDA montre une non-conformité partielle des données au modèle théorique gaussien. Si la loi de Student intervient dans des cas statistiques de carence en informations (échantillon insuffisant) et plus ou moins dans des cas de sur-effectifs d'observations atypiques, la reconnaissance du locuteur, avec maintenant ses vastes jeux d'apprentissage en terme de segments et de locuteurs et malgré l'existence de quelques productions vocales "extrêmes", ne semble pas concernée par ce genre de lois.
- le scoring par cosinus revient à un produit scalaire sur des vecteurs dont l'information de longueur a été ignorée (au sens de la métrique euclidienne ou de la

matrice WCCN). Ce fait est surprenant : la collection complète de trames d'un énoncé de voix, et donc l'énoncé lui-même, est entièrement représenté par 400 à 600 coefficient réels seulement, ou spatialement par un unique point dans un espace de dimension réduite. Annuler l'information de distance à l'origine pour ne conserver que l'information directionnelle est une opération audacieuse. Les i-vectors sont issus des supervecteurs : s'il n'est pas possible de décrire exactement l'opération réciproque dans l'espace de grande dimension (à quoi correspond la division par la norme d'un i-vector sur son supervecteur d'origine ?), nous remarquons que la représentation par la moyenne est une translation et que l'adaptation MAP relativise l'intensité de ce mouvement mais préserve sa direction.
– Considérant les i-vectors comme une transformation des représentations issues du GMM-UBM, la question se pose de leur nature : peut-on, ignorant leur processus d'extraction, les considérer comme des observations issues d'un modèle génératif probabiliste ? Ou bien faut-il les transformer à nouveau pour les rendre compatibles aux hypothèses probabilistes d'un modèle de décomposition des variances ?

Nous présentons ici les résultats de l'analyse que nous avons menée sur les i-vectors. Pour conduire l'analyse, nous avons d'abord mis en place un outil de visualisation des variabilités, que nous emploierons régulièrement pour appuyer nos propos.

3.1 Un outil visuel d'analyse : le graphe spectral

Dans le paradigme i-vector, un énoncé de voix est représenté numériquement par un unique vecteur de dimension réduite. Les modélisations utilisent alors des fichiers d'apprentissage pour décomposer les vecteurs en une somme de facteurs associés aux parts de variabilité intrinsèque et extrinsèque au locuteur. Elles calculent également les paramètres de leurs distributions qui sont employés dans les formules de scoring. Ces modélisations déterminent donc des axes principaux de ces variabilités.

Dans le cadre d'une analyse statistique et spatiale des i-vectors, il est nécessaire d'étudier en premier lieu les variabilités propres à chaque axe d'une base de l'espace de représentation. Chaque dimension de l'espace initial contribue aux diverses variabilités et les coordonnées des vecteurs sur les axes principaux y contribuent en maximisant l'une des variabilités qui nous intéressent. D'autre part, l'emploi éventuel de transformations modifie les coordonnées des vecteurs et donc les intensités de leurs variabilités par dimension.

Les variabilités sur un axe sont estimées par les variances de la série unidimensionnelle des coordonnées sur cet axe issues d'un vaste fichier d'apprentissage. Trois types de variances doivent être prises en compte :

– les variances totales par dimension,
– les variances selon la variable latente locuteur, c'est à dire entre les classes-locuteur,
– les variances résiduelles, combinant ce que l'on nomme généralement l'effet-session et le bruit. Nous parlerons ici de variance "session".

Les séries de ces valeurs sur une base de l'espace ont pour longueur p, où p est la dimension de l'espace i-vectors (de l'ordre de 400 à 600). L'emploi d'un outil de visualisation apparaît indispensable pour les analyser de manière simple et claire. La série des variances totales des dimensions initiales est obtenue par la diagonale de la matrice de covariance totale $\mathbf{\Sigma}$. Les séries des variances locuteur et session sont obtenues par les diagonales des matrices inter-locuteur $\mathbf{B}$ et intra-locuteur $\mathbf{W}$ des équations 2.42 et 2.41. Par décomposition des variances, la somme de ces deux dernières séries (locuteur + session) est nécessairement égale à la première (totale).

Nous appelons *graphe spectral* (par analogie au "spectre" des valeurs propres d'une matrice qui indique ces variances dans sa base de vecteurs propres) le graphe affichant simultanément les trois courbes de ces diagonales dans une base de l'espace.

La visibilité du graphique sera réduite dans la base initiale (immédiatement après extraction) : en effet, ces séries ne sont alors en aucun cas assurées d'être croissantes ou décroissantes, ni proportionnelles par dimension. Il est possible d'exprimer les données dans une base préalablement choisie, ce qui permet d'afficher les trois variances suivant un ordre plus lisible. Pour ce faire, les données sont pivotées suivant les axes de variabilité décroissante d'une des trois matrices $\mathbf{\Sigma}$, $\mathbf{B}$ ou $\mathbf{W}$. L'opération s'effectue par projection des observations sur ses vecteurs propres rangés par ordre décroissant des valeurs propres. Les trois matrices sont alors recalculées dans cette nouvelle base et leurs diagonales affichées. Quelques points doivent être précisés :

- lorsqu'une telle opération est réalisée (par exemple suivant $\mathbf{\Sigma}$), la nouvelle matrice $\mathbf{\Sigma}$ après rotation est diagonale et sa diagonale est constituée par le spectre décroissant de ses valeurs propres. La courbe du graphe spectral qui lui correspond sera donc nécessairement décroissante.
- par contre, les deux autres matrices ne seront pas nécessairement diagonales et les courbes de leurs diagonales non nécessairement décroissantes.
- la variable latente étant le locuteur, la rotation suivant $\mathbf{B}$ sera la plus couramment utilisée dans notre étude.

Nous noterons "Graphe Spectral-$\mathbf{\Sigma}$" (resp. -$\mathbf{B}$, -$\mathbf{W}$) le graphe spectral dans la base des vecteurs propres de $\mathbf{\Sigma}$ (resp. $\mathbf{B}$, $\mathbf{W}$).

La figure 3.1 présente un exemple de graphe spectral-$\mathbf{\Sigma}$, sur un jeu de données artificiel. Dans la base des vecteurs propres de $\mathbf{\Sigma}$, la courbe des variances totales des 600 dimensions est décroissante. Sur chaque dimension, cette variance totale se décompose exactement en somme des variances locuteur et session. Sur cet exemple, les variances locuteur sont majoritairement plus faibles que les variances session, dans des proportions que le graphe spectral permet d'apprécier. D'autre part, aucune corrélation entre ces trois variances n'est perceptible sur ce graphe.

Les graphes spectraux sont autant utiles à effectuer une "radioscopie" des i-vectors initiaux, tels que fournis par l'extracteur FA-total-var, qu'à observer l'effet de transformations sur les vecteurs. Ils participent également à justifier des méthodes, notamment celles mises en place pour améliorer la qualité des modélisations. En effet, la séparation des parts de variabilité locuteur et session dans une représentation vectorielle constitue

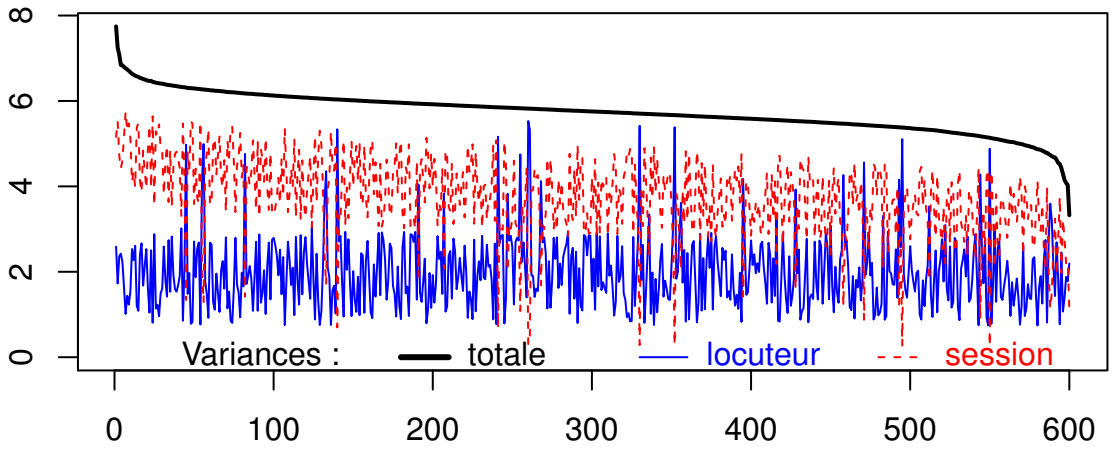

FIGURE 3.1 – *Un exemple de graphe spectral-Σ. Les abscisses sont les* 600 *dimensions de l'espace i-vector.*

un enjeu important de la discrimination du locuteur et le dépouillement visuel permet d'apprécier clairement et rapidement la justesse d'hypothèses ou les effets de transformations appliquées à ces représentations.

3.2 Transformations des i-vectors

Comme nous l'avons indiqué dans l'introduction de ce chapitre, deux voies peuvent être envisagées une fois effectuée l'extraction des i-vectors :

- ou bien ceux-ci sont considérés comme des observations issues d'un modèle génératif probabiliste. Ils sont alors soumis à une décomposition en facteurs et utilisés dans un scoring basé sur des paramètres estimés de distribution des variances,
- ou bien ceux-ci sont préalablement transformés pour les rendre compatibles aux hypothèses d'un modèle probabiliste.

Plusieurs points ont été pris en compte pour guider notre analyse :

- l'extraction des i-vectors s'appuie sur des hypothèses probabilistes : ceux-ci sont supposés suivre a priori une loi normale standard. La conformité des données issues de l'extracteur à ces hypothèses est un sujet d'étude important.
- Les i-vectors sont décomposés et scorés suivant des modèles et similarités basés sur de nouvelles hypothèses. Leur conformité à ces hypothèses est également à étudier.
- Si des défauts supposés de modélisation sont relevés, des transformations doivent être proposées, capables d'atténuer -voire éliminer- ces insuffisances de la représentation. De pus, l'efficacité de ces transformations doit être à chaque fois quantifiable.

Nous analysons ici une série de caractéristiques statistiques et spatiales qui tiennent un rôle essentiel dans la qualité des modélisations par i-vector.

3.2.1 Standardité

La Factor Analysis Total Var. réalise une réduction de dimension des supervecteurs du GMM-UBM sous la contrainte d'hypothèses probabilistes et gaussiennes : la normalité des résidus et la normalité standard des i-vectors obtenus.

Elle sous-entend donc implicitement que les écarts entre l'échantillon des vecteurs obtenus et cette loi théorique sont seulement imputables aux fluctuations d'échantillonnage. Ces fluctuations sont inhérentes à toute méthode itérative basée sur la contingence d'un jeu de données fini. En pratique, la question se pose de savoir si ces écarts doivent être conservés ou éliminés, mais également par quelle méthode. L'opération de centrage-réduction (*standardisation*) s'écrit

$$w \leftarrow \mathbf{\Sigma}^{-\frac{1}{2}} (w - \mu) \tag{3.1}$$

ou, ce qui revient au même, $w \leftarrow \mathbf{L}(w - \mu)$ où μ est la moyenne globale du fichier d'apprentissage et $\mathbf{L}$ vérifie $\mathbf{\Sigma}^{-1} = \mathbf{L}\mathbf{L}^t$, matrice obtenue par la décomposition de Cholesky de la matrice de précision. Elle transforme les vecteurs du jeu de données d'apprentissage de sorte que leurs nouvelles moyenne et matrice de covariance soient 0 et la matrice identité $\mathbf{I}$. Rien n'indique pour autant qu'ils ne suivent une loi normale, comme supposé a priori. Mais cette opération permet de les rapprocher des hypothèses, en entraînant l'indépendance statistique de leurs dimensions et en homogénéïsant leurs variabilités.

Tout vecteur de test w doit être alors standardisé selon les paramètres de tendance centrale du fichier d'apprentissage :

$w \leftarrow \mathbf{\Sigma}_{\text{appr}}^{-\frac{1}{2}} \left(w - \mu_{\text{appr}}\right)$

Cette transformation joue un rôle de *conditionnement* des nouvelles observations aux paramètres appris : celles-ci sont rapprochées de la distribution empirique de l'apprentissage. Nous parlerons ici d'*adéquation* ou de *mise à conformité* des données d'évaluation aux données d'apprentissage. Nous étudierons dans la suite la capacité de la standardisation à améliorer cette adéquation.

Concernant l'opportunité de cette transformation dans le cadre de vecteurs issus de la FA-Total Var, elle peut être justifiée de la manière suivante : tout i-vector est obtenu à partir des formules

$$\begin{aligned} w &= \left(\mathbf{I} + \mathbf{T}^t\mathbf{\Sigma}^{-1}N\mathbf{T}\right)^{-1} \mathbf{T}^t\mathbf{\Sigma}^{-1}S_{\mathcal{X}} \\ &= \left(\mathbf{I} + \mathbf{T}^t\mathbf{\Sigma}^{-1}N\mathbf{T}\right)^{-1} \mathbf{T}^t\mathbf{\Sigma}^{-1}N\,(s - \mu) \end{aligned} \tag{3.2}$$

où S_X est la statistique d'ordre 1 de la collection de trames $\mathcal{X}$, s est le supervecteur correspondant, N la matrice diagonale des statistiques d'ordre 0, μ la moyenne de l'UBM. En notant $\widetilde{\mathbf{T}} = \mathbf{\Sigma}^{-\frac{1}{2}}\mathbf{T}$ et $\widetilde{s} = \mathbf{\Sigma}^{-\frac{1}{2}}(s-\mu)$, le i-vector w se réécrit :

$$w = \left(\mathbf{I} + \widetilde{\mathbf{T}}^t N \widetilde{\mathbf{T}}\right)^{-1} \widetilde{\mathbf{T}}^t N \widetilde{s} \tag{3.3}$$

Le calcul de w ne fait intervenir que la version standardisée $\widetilde{s}$ de s. Cela signifie que les variabilités spécifiques des composantes gaussiennes du GMM sont ignorées (perdues) dans le processus d'extraction. Ce fait peut surprendre : ces informations sont indispensables à un modèle de discrimination de locuteur comme la JFA, qui travaille dans l'espace GMM. Mais l'équation précédente montre que, dans le cadre d'une réduction de dimension, les corrélations existantes entre les vecteurs de chaque gaussienne ne peuvent être compressées que si elles sont comparables entre gaussiennes, donc si leurs distributions ont été normalisées. En ce sens, les écarts à la loi théorique sont bien des anomalies.

La figure 3.2 montre le graphe spectral du jeu de données d'apprentissage BUT-hommes (détaillé en annexe A) immédiatement après son extraction par FA-Total Var, c'est à dire avant toute transformation et, donc, dans sa base canonique d'origine. Les 600 variances (600 est la dimension de l'espace i-vector) pour les trois types de variabilités sont affichées. Les i-vectors suivent en théorie une loi normale standard. Nous nous intéressons ici à leurs seules variances spectrales, i.e. à leur matrice de covariance (l'étude de leur gaussianité est effectuée plus loin). La matrice de covariance étant théoriquement égale à la matrice identité, on observe sur la figure 3.2 que les variances des dimensions initiales sont effectivement proches de 1. Les parts de variance intra- et inter-locuteur mesurées par les diagonales de $\mathbf{B}$ et $\mathbf{W}$ semblent également indépendantes entre les dimensions. La contrainte probabiliste de standardité -au moins en terme de covariance- semble apparemment respectée.

Nous procédons alors à un changement de base, par rotation suivant la base de vecteurs propres de la covariance totale $\mathbf{\Sigma}$. Notant P cette matrice de vecteurs propres de $\mathbf{\Sigma}$, les i-vectors deviennent :

$$w' = \mathbf{P}^t w \tag{3.4}$$

La matrice $\mathbf{P}$ étant orthogonale, la covariance totale serait invariante par cette transformation si $\mathbf{\Sigma}$ était exactement égale à la matrice identité.

La figure 3.3 affiche le graphe spectral-$\mathbf{\Sigma}$ (obtenu après rotation par $\mathbf{P}^t$ dans la base de $\mathbf{\Sigma}$). Le spectre de variance totale par dimension présente clairement une énergie : des axes principaux de variabilité se dégagent. L'opération de rotation n'a pourtant en rien transformé les données initiales, les proximités entre points ayant été strictement conservées. Cet état de fait constitue en soi un sujet de réflexion. La standardité des i-vectors en sortie de l'extraction n'est qu'apparente : l'algorithme FA-Total Var a plus ou moins égalisé les variances dans la base canonique de réduction, mais n'a en rien assuré le lissage uniforme de celles-ci. Comme l'indique également la figure 3.3, les trois courbes de variance présentent une corrélation élevée : les variabilités locuteur

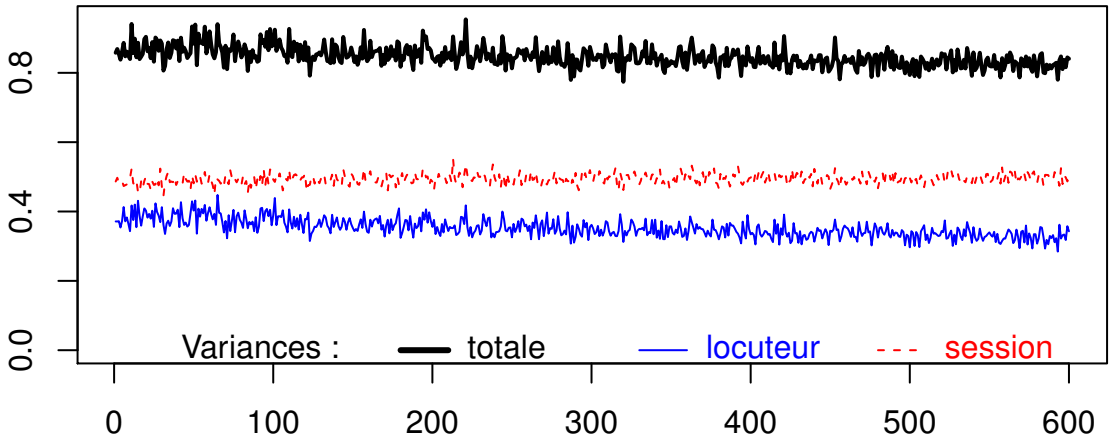

FIGURE 3.2 – *Graphe spectral du jeu de données d'apprentissage BUT-hommes, immédiatement après son extraction par FA-Total Var (dans la base canonique).*

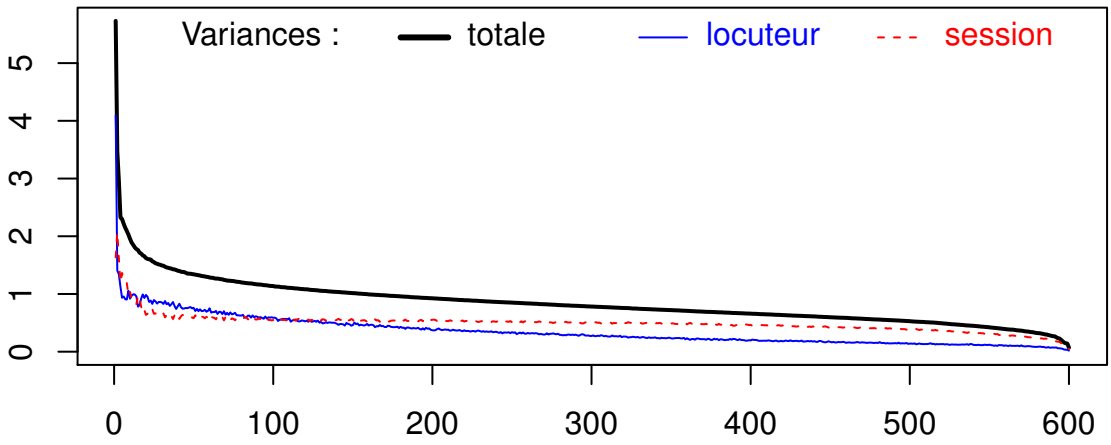

FIGURE 3.3 – *Graphe Spectral-Σ des mêmes données qu'à la figure précédente.*

et session sont significativement corrélées à la variabilité totale. Les deux premières courbes décroissent même strictement avec cette dernière.

Les allures non-aléatoires des courbes locuteur et session, en particulier la corrélation entre variances totale et locuteur ne sont pas explicables par un défaut de l'algorithme FA-total-var : celui-ci ignore la variable locuteur. La non-conformité des données à une loi standard, en sortie d'extracteur, est incontestablement liée à la présence de la variable latente locuteur. La standardisation forcée des données n'a pas pour seul but de rapprocher les i-vectors de leur modèle théorique. Comme le montre la figure 3.3, la part de variance session dans la variance totale augmente en même temps que cette dernière diminue (les derniers axes à droite contiennent la plus forte part de variance session). La standardisation va ainsi faciliter la tâche de séparation des variabilités explicative et résiduelle par un modèle génératif.

3.2.2 Gaussianité

Les modèles génératifs et méthodes de scoring utilisés dans le cadre des i-vectors s'appuient sur des hypothèses de gaussianité, qu'il s'agisse de configurer la variabilité totale, locuteur ou session.

La modélisation par mixture de gaussienne de l'espace acoustique ne traduit pas nécessairement une gaussianité des données : il est toujours possible d'approximer une distribution quelconque par une combinaison linéaire de lois normales, de même qu'on interpole une fonction par des polynômes, ou des séries de Fourier lorsqu'elle est périodique. Une fois transférés les segments de parole dans l'espace compact des i-vectors, les hypothèses sur la nature de leur distribution dans un cadre de modélisation confirmatoire (basée sur des a priori que l'empirique doit vérifier et ainsi valider) constitue le principal obstacle à lever. Des stratégies basées sur la loi normale, bien entendu, mais aussi sur la loi de Student ("Heavy-tailed" HT-PLDA) ont été mises en place et la porte reste ouverte à d'autres propositions. Mais la souplesse des lois basées sur la famille exponentielle, dont les formulations mathématiques autorisent la constitution de distributions conjuguées [1] en forme close et des calculs de vraisemblance exacts et rapides à mettre en oeuvre, a beaucoup contribué à maintenir les investigations dans leur domaine.

Le défaut de gaussiannité qui a conduit à introduire la HT-PLDA doit être constaté. Pour cela, une mesure de gaussianité sur des observations multidimensionnelles est à déterminer, dont les résultats seront commentés sur les jeux de données à disposition. D'autre part, si des transformations sont mises en place sur les i-vectors, comme nous l'avons envisagé, la mesure doit présenter un caractère absolu pour comparer les jeux initiaux et après transformation. Les techniques de gaussianisation des données s'incluent dans celles dites de "blanchiment" des données (*whitening*), destinées à éliminer, redresser ou au moins atténuer des anomalies par rapport aux hypothèses probabilistes émises.

1. la distribution a posteriori issue de la fonction de vraisemblance étant de même type que la distribution a priori

Mesures de gaussianité : remarquons d'abord que, lorsqu'un vecteur aléatoire v de $\mathbb{R}^p$ suit une loi multinormale standard, sa norme $\|v\| = \sum_{j=1}^{p} v_j^2$ est une somme de p carrés de loi normales unidimensionnelles centrées-réduites et suit donc une loi du χ^2 à p degrés de liberté. Mesurer la gaussianité de données multidimensionnelles nécessite quelques précautions. Il s'agit d'étudier l'adéquation d'un jeu de données multidimensionnel à un modèle gaussien. La fonction de répartition est incalculable, mais il est possible de procéder par mesure de vraisemblance, sous l'hypothèse que la distribution soit normale. Le critère de mesure de la vraisemblance est alors la densité gaussienne. La vraisemblance d'une observation w de $\mathbb{R}^p$ sous l'hypothèse que w suit une loi normale $\mathcal{N}(\mu, \mathbf{\Sigma})$ est :

$$\mathcal{L}(w|\mathcal{N}(\mu, \mathbf{\Sigma})) = P(w|\mathcal{N}(\mu, \mathbf{\Sigma})) = (2\pi)^{-\frac{p}{2}} |\mathbf{\Sigma}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(w-\mu)^t \mathbf{\Sigma}^{-1}(w-\mu)\right) \tag{3.5}$$

Par dérivation du logarithme de $\mathcal{L}$ suivant μ et $\mathbf{\Sigma}$, la loi normale la plus vraisemblable pour un ensemble de données $\{w_i\}_{i=1,\ldots,n}$ est celle de moyenne :

$$\mu = \frac{1}{n}\sum_{i=1}^{n} w_i \tag{3.6}$$

et de matrice de covariance

$$\mathbf{\Sigma} = \frac{1}{n-1}\sum_{i=1}^{n} (w_i - \mu)^t (w_i - \mu) \tag{3.7}$$

La vraisemblance totale de $\{w_i\}_{i=1,\ldots,n}$ est alors :

$$\mathcal{L}(\{w_i\}|\mathcal{N}(\mu, \mathbf{\Sigma})) = \left(\prod_{i=1}^{n} P(w_i|\mathcal{N}(\mu, \mathbf{\Sigma}))\right)^{\frac{1}{n}} \tag{3.8}$$

qui se simplifie en

$$\begin{aligned}
\mathcal{L}(\{w_i\}|\mathcal{N}(\mu, \mathbf{\Sigma})) &= (2\pi)^{-\frac{p}{2}} |\mathbf{\Sigma}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2} Tr\left(\mathbf{\Sigma}^{-1}\frac{1}{n}\sum_{i=1}^{n}(w_i-\mu)^t(w_i-\mu)\right)\right) \\
&= (2\pi)^{-\frac{p}{2}} |\mathbf{\Sigma}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\frac{n-1}{n} Tr\left(\mathbf{\Sigma}^{-1}\mathbf{\Sigma}\right)\right) \\
&= (2\pi)^{-\frac{p}{2}} |\mathbf{\Sigma}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\frac{(n-1)\,p}{n}\right)
\end{aligned} \tag{3.9}$$

où $Tr(.)$ est l'opérateur trace

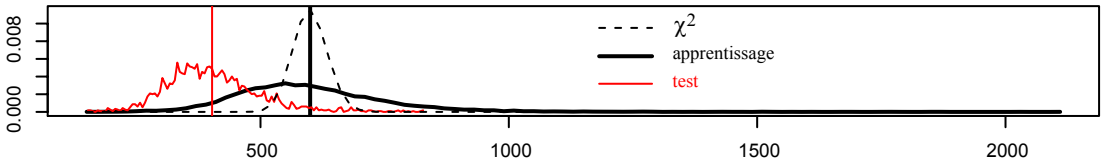

FIGURE 3.4 – *Histogrammes des normes carrées des jeux de données standardisés d'apprentissage et d'évaluation BUT-hommes, et densité du χ^2 à p degrés de liberté ($p = 600$).*

La vraisemblance n'est dépendante que de la précision de la variabilité (de plus, elle est maximale si tous les points sont confondus !). Etudier la vraisemblance gaussienne globale d'un jeu de données avant et après une transformation ne procure aucune information sur l'évolution de sa gaussianité. Elle ne peut s'entendre qu'en comparaison de différents jeux de données sur un espace commun. Par contre, la distribution des termes exponentiels de la vraisemblance précédente :

$$S = \left\{ (w_i - \mu)^t \, \mathbf{\Sigma}^{-1} \, (w_i - \mu) \right\}_{i=1,\dots n} \tag{3.10}$$

peut se réécrire :

$$S = \left\{ \left\| \mathbf{\Sigma}^{-\frac{1}{2}} \, (w_i - \mu) \right\|^2 \right\}_{i=1,\dots n} \tag{3.11}$$

soit comme une norme carrée -autrement dit la somme de carrés de lois standards-. Si celles-ci sont normales, alors S suit une loi du χ^2 à p degrés de liberté. La comparaison de l'histogramme empirique de la série S et de la courbe de densité du $\chi^2(p)$ permet donc d'estimer la gaussianité du jeu de données.

La figure 3.4 présente l'histogramme des valeurs de S calculées sur les jeux de données d'apprentissage et d'évaluation BUT-hommes, ainsi que la densité du χ^2 à p degrés de liberté (ici $p = 600$).

- la courbe en pointillé est celle de la distribution théorique du χ^2 à p degrés de liberté. On rappelle que sa moyenne est p (ici 600) et son écart-type $\sqrt{2p}$ (ici $\approx$ 34.64).
- la courbe noire en gras est la distribution empirique des normes carrées des données d'apprentissage standardisées (la distribution S de l'équation 3.11).
- la troisième courbe est la distribution empirique des normes carrées des données d'évaluation standardisées. La standardisation a été effectuée sur les paramètres de tendance centrale du fichier d'apprentissage. Nous avons donc calculé et affiché :

$$S^{\text{test}} = \left\{ \left\| \mathbf{\Sigma}^{-\frac{1}{2}} \left(w_k^{\text{test}} - \mu \right) \right\|^2 \right\}_{k=1,\dots,N_{\text{test}}} \tag{3.12}$$

avec les paramètres $\mathbf{\Sigma}$ et μ de l'apprentissage.

Comme le montre la figure, les données présentent plusieurs anomalies :

– la distribution empirique est asymétrique. Notons que la moyenne de ces valeurs, indiquée par la droite verticale noire en gras, est proche de $p = 600$, mais ce par sa nature même. En effet, l'équation 3.11 ci-dessus montre qu'elle vaut exactement :

$$\begin{aligned} E[S] &= \frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{\Sigma}^{-\frac{1}{2}}(w_i-\mu)\right)^t\left(\mathbf{\Sigma}^{-\frac{1}{2}}(w_i-\mu)\right) \\ &= Tr\left(\mathbf{\Sigma}^{-1}\frac{1}{n}\sum_{i=1}^{n}(w_i-\mu)(w_i-\mu)^t\right) \\ &= Tr\left(\mathbf{\Sigma}^{-1}\frac{n-1}{n}\mathbf{\Sigma}\right) = \frac{n-1}{n}Tr(I) \\ &= \frac{(n-1)\,p}{n} \end{aligned} \tag{3.13}$$

Le caractère asymétrique de la distribution est une preuve de faible gaussianité.

– la densité empirique de l'apprentissage est plus "plate" qu'un $\chi^2(p)$ c'est à dire que la proportion de valeurs extrêmes est plus importante qu'en situation gaussienne. La distribution est dite *heavy-tailed* (à queue de distribution importante) et l'on comprend mieux l'intérêt de la PLDA adaptée à ce type de distribution, telle qu'introduite dans (Kenny, 2010) en reconnaissance du locuteur.
– la densité empirique de l'évaluation est plus proche en courbure d'un $\chi^2(p)$, mais présente un décalage net de valeurs : la moyenne, indiquée par la droite verticale en rouge, est de l'ordre de 400, très inférieure aux 600 théoriques. Cela tient essentiellement à un biais entre les moyenne globales μ et μ^{test} des jeux de données.

En annexe C sont présentés les histogrammes correspondants pour les jeux de données BUT-femmes et LIA-hommes (détaillés en annexe A). Le premier permet notamment de vérifier le comportement similaire des distributions pour les fichiers féminins et donc l'indépendance au genre des constats précédents.

Cette étude confirme le défaut de gaussianité de la représentation i-vectors, telle que fournie par l'extracteur FA-total-var. L'alternative d'une autre loi de la famille exponentielle (Student), comme distribution a priori des vecteurs, a été proposée avec la HT-PLDA. D'autres lois restent envisageables, mais nous proposons au paragraphe suivant une autre stratégie pour pallier à ces défauts de gaussianité, basée sur la transformation des vecteurs.

3.2.3 Unitarité

La division de vecteurs par leur norme permet d'obtenir de nouveaux vecteurs de norme 1, dits *normés* ou *unitaires*. En mathématique, le terme d'*unitarisation* est alors uti-

lisé pour définir spécifiquement ce type de *normalisation de longueur*. Nous emploierons par la suite ces deux termes, dans la même acception.

Une des premières tentatives de scoring sur des i-vectors, présentée dans (Dehak et al., 2011) (Dehak et al., 2009), utilisait un simple cosinus sur des i-vectors tel qu'initialement proposé comme noyau pour une expérience basée sur un SVM dans l'espace des i-vectors. Etant donnés deux i-vectors w_1, w_2 à comparer, le score proposé est [2] :

$$k(w_1, w_2) = \cos(w_1, w_2) = \frac{w_1.w_2}{\|w_1\| \, \|w_2\|} \tag{3.14}$$

Le scoring associé, issu de ce noyau-cosinus, prend en compte la variabilité intra-locuteur, qu'il cherche à normaliser. Il s'agit du scoring présenté dans le chapitre précédent :

$$k(w_1, w_2) = \frac{\left(\mathbf{W}_{[\text{WCCN}]}^{-\frac{1}{2}} w_1\right)^t \mathbf{W}_{[\text{WCCN}]}^{-\frac{1}{2}} w_2}{\left\|\mathbf{W}_{[\text{WCCN}]}^{-\frac{1}{2}} w_1\right\| \left\|\mathbf{W}_{[\text{WCCN}]}^{-\frac{1}{2}} w_2\right\|} \tag{3.15}$$

Comme le montrent ces formules, ce scoring est également un cosinus, mais appliqué sur des vecteurs préalablement multipliés par la racine carré $\mathbf{W}_{[\text{WCCN}]}^{-\frac{1}{2}}$ de la matrice de précision intra-locuteur. Ces vecteurs ont donc une nouvelle matrice de covariance $\mathbf{W}_{[\text{WCCN}]}$ égale à l'identité : aucune dimension ne possède plus de variance qu'une autre. On parlera alors d'*isotropie*.

Les performances obtenues par ce scoring-cosinus surpassent celles obtenues par un produit scalaire entre les vecteurs ou par leur distance (euclidienne, comme suivant d'autres métriques). Un cosinus est un simple produit scalaire sur des vecteurs unitaires. Ce fait a attiré notre attention : en effet, il implique que la longueur initiale d'un i-vector ne contient pas d'information utile en discrimination du locuteur, ou, pour être plus prudent, que ces informations ne sont pas "maîtrisables" par les mesures de proximité usuelles. Dans (Dehak et al., 2011), il est avancé l'hypothèse que l'information non-locuteur (telle que session et canal) affecte la norme des i-vectors et que l'ignorer augmente la robustesse du i-vector.

Un tel cas n'est pas nouveau dans le domaine de l'analyse de données (il se retrouve, par exemple, dans certaines problématiques des génomique, morphométrie ou reconnaissance d'images (Hamsici and Martinez, 2007)) mais reste peu courant, en particulier pour les méthodes prédictives et, surtout, nécessite une adaptation des hypothèses et modélisations aux objets représentés. La **normalisation de longueur** envoie les vecteurs sur la surface d'une hypersphère de l'espace de représentation. Elle modifie de façon non-linéaire les proximités entre ces vecteurs et par conséquent l'ensemble des paramètres de leur distribution statistique.

2. Même si rien ne le précise dans les articles de référence, cette opération est effectuée sur des vecteurs préalablement centrés, suivant une moyenne estimée à partir d'un jeu d'apprentissage.

Les i-vectors étant supposés suivre une loi normale standard, il est utile de s'interroger sur l'effet d'une telle transformation qui conduit les i-vectors vers une distribution inconnue, non recensée dans la littérature. Les distributions gaussiennes radiales ne correspondent pas exactement à ce cas de configuration.

Le constat d'efficacité de la normalisation de longueur découle donc des performances obtenues par les premiers scorings-cosinus effectués sur les i-vectors. Ce constat a motivé notre travail avec trois objectifs :

(i) expliquer l'effet positif (au sens de l'objectif de discrimination du locuteur) de cette transformation,

(ii) exploiter plus complètement et plus finement son potentiel,

(iii) quantifier l'apport de cette transformation dans la qualité du système, par rapport à ceux des autres constituants.

Concernant le premier point, il sera justifié dans la suite -au moins partiellement- par un certain nombre de propriétés. Mais une remarque peut être effectuée dès maintenant : si le rôle de cette opération dans la discrimination du locuteur est supposé positif, alors ses conséquences peuvent être étudiées en amont, c'est à dire sur la représentation initialement pourvue par le modèle GMM-UBM. Supposons nulle la moyenne globale des i-vectors d'apprentissage (ce qui est conforme à l'a priori de standardité des i-vectors) et soient deux i-vectors w et w' liés par la relation homothétique :

$$w' = \alpha w (\alpha > 0) \tag{3.16}$$

Ces deux vecteurs se retrouvent confondus après normalisation de leurs longueurs. Les supervecteurs s et s' dont ils découlent vérifient :

$$s' - \mu = \mathbf{T}w' = \alpha \mathbf{T}w = \alpha \left(s - \mu\right) \tag{3.17}$$

ce qui correspond à la situation triviale illustrée par la figure 3.5 : sur chacune des trois gaussiennes g_i, le sous-vecteur de moyenne centré $\left(s'_{g_i} - \mu_{g_i}\right)$ est égal à $\alpha \left(s_{g_i} - \mu_{g_i}\right)$.

Les deux sessions représentées par ces deux supervecteurs s et s' se retrouveront confondues dans leur représentation en i-vectors, donc à score maximal pour tout scoring. Ce fait traduit l'importance de la proximité directionnelle (angulaire) et pas seulement absolue des représentations GMM-UBM. Une telle proximité d'angle est d'ailleurs bien notée par les scorings *llr-by-frame* (log-ratio de vraisemblance des moyenne adaptée / trames) de la JFA.

Concernant le second point "(ii) exploiter plus complètement et plus finement son potentiel", il est à remarquer que la normalisation de longueur n'intervient pour Dehak et al. qu'en phase finale de traitement, dans le scoring, et comme composant d'une mesure angulaire de proximité par le cosinus. Une application de cette normalisation en amont du scoring, comme transformation préparatoire à un modèle génératif ou discriminant, nous a paru mériter notre attention.

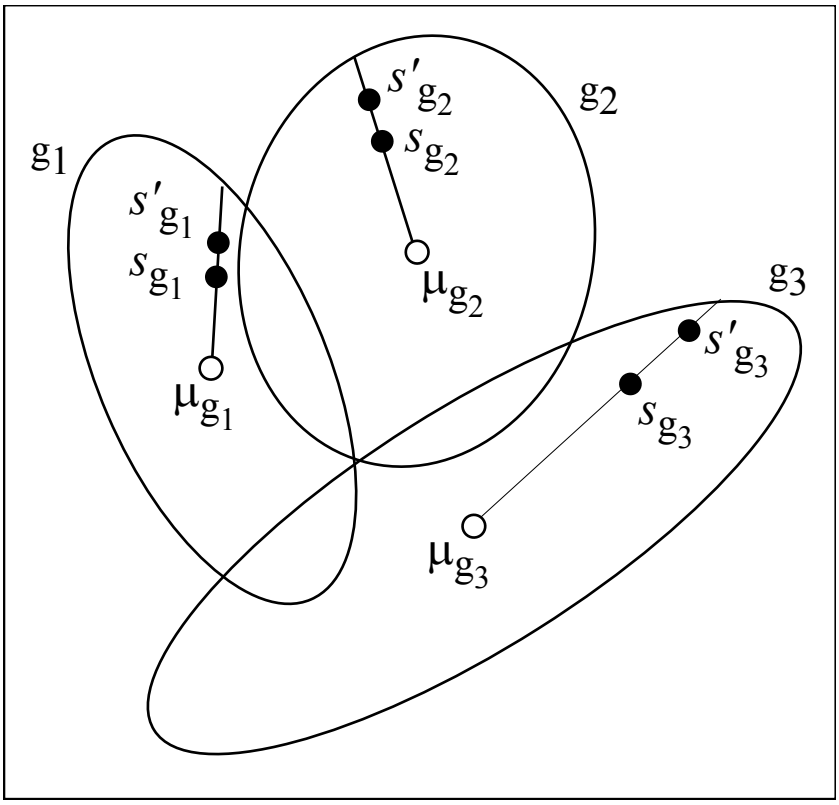

FIGURE 3.5 – *Cas de deux segments de voix aux représentations "proportionnelles" ...*

Le troisième point "(iii) quantifier l'apport de cette transformation ..." fera l'objet d'une mesure comparée de performance dans le chapitre 5.

Une étrange propriété

Nous relevons également dans la communauté [3] un résultat étrange dans les espaces vectoriels de grande dimension. Considérons la densité d'une loi normale standard multidimensionnelle. Elle est maximale en son mode, qui, la loi de Gauss étant symétrique, se trouve être sa moyenne. C'est donc autour de 0 que se trouve la plus grande concentration de densité. Mais il peut être intéressant d'étudier la densité des coquilles gaussiennes de l'espace : dans le cas d'une loi standard, une coquille gaussienne est la surface d'une sphère centrée en 0 et de rayon r. Nous nous intéressons alors à la comparaison des densités de chaque coquille. Intuitivement, c'est en se rapprochant de 0 que cette densité sera maximale. Mais l'intuition s'avère parfois erronée en grande dimension. Considérons la somme des densités d'une loi normale standard p-dimensionnelle sur la coquille de centre 0 et rayon r. Cette valeur, notée $\Delta(r)$, est égale à :

$$\Delta(r) = \int_{\|x\|=r} \mathcal{N}(x|0, \mathbf{I})\, dx \tag{3.18}$$

où $\mathcal{N}(x|0, I)$ est la densité normale standard $|2\pi I|^{-\frac{1}{2}} e^{-\frac{1}{2}(x-0)^t I^{-1}(x-0)}$. S'interrogeant sur le mode de cette fonction, c'est à dire sur sa plus grande valeur, nous obtenons par simplification :

$$\begin{aligned} \Delta(r) &= (2\pi)^{-\frac{p}{2}} e^{-\frac{r^2}{2}} \int_{\|x\|=r} dr \\ &= (2\pi)^{-\frac{p}{2}} \mathcal{S}(r) e^{-\frac{r^2}{2}} \end{aligned} \tag{3.19}$$

où $\mathcal{S}(r)$ est la surface de l'hypersphère de rayon r. On sait que cette valeur s'écrit $K(p) r^{p-1}$ où $K(p)$ est une valeur ne dépendant que de p. La densité de la coquille de rayon r est donc égale à :

$$\Delta(r) = C r^{p-1} e^{-\frac{r^2}{2}} \tag{3.20}$$

où C est une constante. Par dérivation suivant r, on obtient :

$\frac{\delta \Delta(r)}{\delta r} = C\left((p-1) r^{p-2} - r^p\right) e^{-\frac{r^2}{2}}$ égal à 0 si et seulement si $(p-1) = r^2$

L'unique maximum est atteint pour la valeur $r_{\max} = \sqrt{(p-1)}$, qui n'est égale à 0 que lorsque $p = 1$. D'une manière assez peu prévisible, malgré que la loi normale

3. http ://ontopo.wordpress.com/2009/03/10/reasoning-in-higher-dimensions-measure/

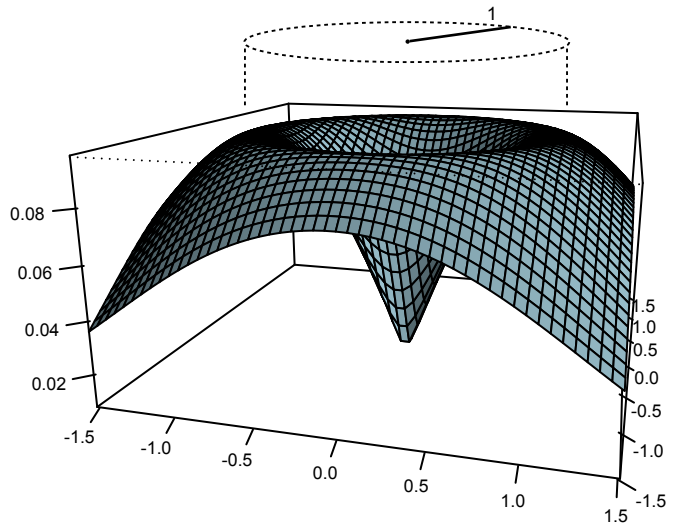

FIGURE 3.6 – *Densités des coquilles gaussiennes standards en dimension 2.*

"perde" de sa densité en s'éloignant de l'origine, la surface sphérique de densité maximale se trouve à $\sqrt{(p-1)}$ de celle-ci. Cette valeur correspond à l'estimateur de l'écart-type, rayon de l'hypersphère dans laquelle est concentré l'essentiel de la distribution.

La figure 3.6 illustre cette propriété en dimension $p = 2$. Les densités $\Delta(r)$ des coquilles de la loi $\mathcal{N}(0, \mathbf{I})$ atteignent leur maximum pour $r = \sqrt{p-1} = 1$.

En ce sens, un ensemble de données migrées sur cette surface présentera une plus forte vraisemblance vis à vis d'une loi normale standard, puisque les données voisineront alors avec une forte densité d'éléments normalement distribués. Bien entendu, la démarche de migration reste artificielle, mais l'unitarisation (normalisation de longueur) des i-vectors constitue une possible transformation pour améliorer leur caractère gaussien.

3.3 Conclusion

Les diverses caractéristiques statistiques et spatiales que nous avons étudiées ont montré certains défauts des i-vectors livrés par la procédure FA-total-var, en terme de conformité aux hypothèses précédant ou suivant leur extraction : ainsi, la normalité et la standardité théoriques des facteurs i-vectors ne sont pas acquises. L'outil de visualisation des variabilités (graphe spectral) s'avère utile pour apprécier rapidement et clairement le niveau de standardité d'un jeu de données, ainsi que les comportements dans et entre les classes-locuteurs. L'étude sur les normes des vecteurs, appuyée par la loi théorique du χ^2, confirme un manque de gaussianité des données.

En lieu et place d'une modélisation basée sur des postulats non-gaussiens, tels que ceux de la Heavy-tailed PLDA utilisant la loi de Student, des transformations sont envisageables pour corriger ces défauts, notamment la standardisation et la normalisation de longueur. Cette dernière technique, non-paramétrique, de blanchiment des données pour gaussianisation paraît expliquer l'efficacité du score par cosinus mis en place initialement dans le champ des i-vectors. Cette hypothèse, et plus généralement la prise en compte des conclusions de cette analyse, nous ont amenés à mettre en place une famille de transformations des i-vectors avant modélisation et scoring, qui sont présentées et détaillées au chapitre suivant.

Chapitre 4

Techniques de normalisation

Les résultats de l'analyse des i-vectors que nous avons conduite ont montré des défauts de la représentation i-vector, telle que fournie par l'extracteur FA-total-var. Ces défauts s'entendent en terme de conformité aux hypothèses, qu'il s'agisse de celles de l'extracteur en amont ou des modèles en aval. Nous envisageons la possibilité d'appliquer des transformations aux vecteurs, capables de corriger ou réduire ces défauts. L'objectif fondamental est de considérer les vecteurs après transformations comme des observations d'un modèle génératif probabiliste.

Nous présentons dans ce chapitre une famille de techniques de normalisation que nous avons mises en place comme transformations préliminaires avant l'élaboration d'un modèle génératif dans l'espace des i-vectors. Nous montrons comment il est possible de traiter les données préalablement pour fournir aux modèles LDA et PLDA des données plus proches de leurs hypothèses et mieux optimiser les critères qu'ils se sont fixés. Les analyses du chapitre précédent sont reprises sur les i-vectors après application de ces diverses transformations, justifiant leur pertinence.

Les transformations mises en place font chaque fois tendre les données vers un modèle théorique. Nous présentons ce modèle, puis recensons et démontrons un certain nombre de propriétés de ses vecteurs, qui éclairent et justifient leur rôle dans la qualité de la modélisation.

4.1 La transformation EFR (LΣ)

Dans (Bousquet et al., 2011b) , nous avons introduit une transformation non-linéaire des i-vectors issus de l'extraction FA-Total Var., destinée à préparer les données à un modèle génératif. Nous présentons l'algorithme que nous avons mis en place pour effectuer cette transformation, puis détaillons ses différents effets sur les données et un certain nombre de ses propriétés. La partie 4.5 montre son efficacité sur diverses expériences de discrimination du locuteur.

Présentation de la transformation Eigen Factor Radial (EFR

Etant donné un jeu de données d'apprentissage $\mathcal{T}$ et notant p la dimension des i-vectors, l'algorithme EFR de transformation des i-vectors procède en deux phases (apprentissage et test) :

Phase d'apprentissage

Pour $i = 1$ à $nb\_iterations$
Calculer la moyenne μ_i et la matrice de covariance totale $\mathbf{\Sigma}_i$ de $\mathcal{T}$
Pour chaque w de $\mathcal{T}$: $\quad w \leftarrow \dfrac{\mathbf{\Sigma}_i^{-\frac{1}{2}}(w-\mu_i)}{\left\|\mathbf{\Sigma}_i^{-\frac{1}{2}}(w-\mu_i)\right\|}$
Pour chaque w de $\mathcal{T}$: $\quad w \leftarrow \sqrt{p}w$

Durant la phase d'apprentissage, les i-vectors de $\mathcal{T}$ sont successivement standardisés puis normalisés par division par leur norme euclidienne. La matrice $\mathbf{\Sigma}_i^{-\frac{1}{2}}$ peut être calculée à partir d'une décomposition en valeurs singulières ou de Cholesky. A chaque itération, la même transformation est appliquée mais avec les paramètres successifs μ_i et $\mathbf{\Sigma}_i$ mis à jour à l'itération précédente.

La dernière opération (multiplication des i-vectors par le scalaire $\sqrt{p}$) est sans effet sur le processus décisionnel de discrimination du locuteur. Elle est destinée à clarifier les justifications théoriques qui suivront. Les paramètres statistiques successifs de moyennes $\{\mu_i\}$ et covariance $\{\mathbf{\Sigma}_i\}$ sont enregistrés pour la phase de test.

Phase de test

Etant donné un i-vector de test w_{test},
Pour $i = 1$ à $nb\_iterations$: $\quad w_{\text{test}} \leftarrow \dfrac{\mathbf{\Sigma}_i^{-\frac{1}{2}}(w_{\text{test}}-\mu_i)}{\left\|\mathbf{\Sigma}_i^{-\frac{1}{2}}(w_{\text{test}}-\mu_i)\right\|}$
$w_{\text{test}} \leftarrow \sqrt{p}w_{\text{test}}$

A partir des paramètres μ_i et $\mathbf{\Sigma}_i$ calculés durant la phase d'apprentissage, un i-vector de test w_{test} est modifié par application itérative de la transformation.

L'effet de cette transformation sur les données d'apprentissage peut être apprécié en dimension 2 sur la figure 4.1. Sur le fichier d'apprentissage initial préalablement centré (a), une rotation est appliquée dans la base des vecteur propres de la variabilité totale (b). Puis les données sont standardisées par division des coordonnées par leurs écart-type (c) et enfin normalisées par division par la norme (d). Au final, les données s'étendent sur la surface de l'hypersphère de rayon $\sqrt{p}$.

Nous nommons cette transformation la **normalisation EFR** (*Eigen Factor Radial*). Ce titre sera justifié dans la suite, qui inventorie les propriétés des vecteurs après cette transformation. Elle est aussi notée dans certaines publications L$\mathbf{\Sigma}$ [1].

1. pour "Length-normalization" et standardisation suivant la matrice de covariance totale $\mathbf{\Sigma}$.

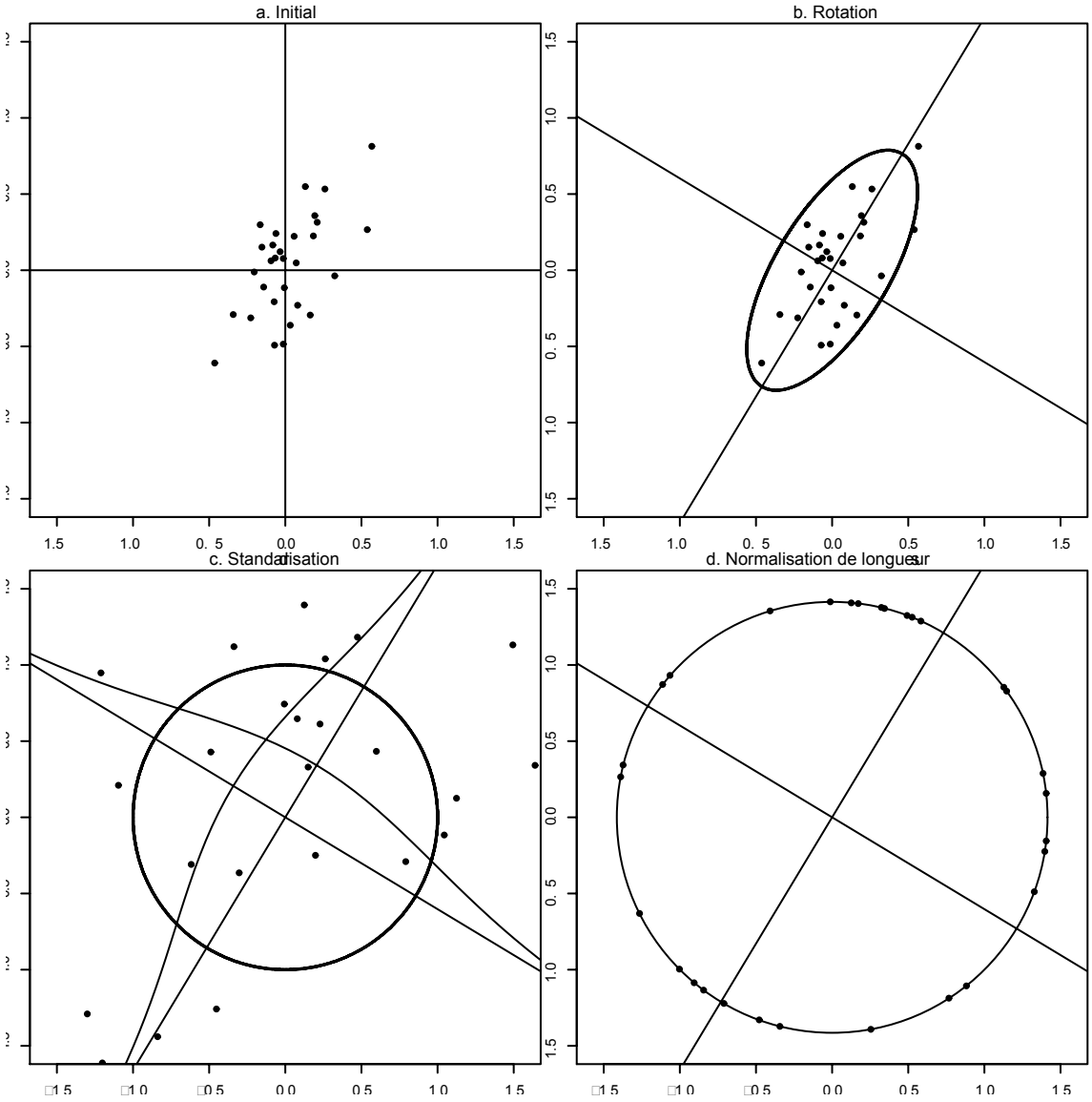

FIGURE *4.1* – *Visualisation en 2D d'une itération d'EFR.*

Notons que, puisque moyenne et matrice de covariance des données tendent vers 0 et **I**, la transformation appliquée à chaque itération de l'algorithme tend vers l'application identique. Ainsi, l'algorithme devient stationnaire au fil des itérations.

4.2 Premières propriétés de la normalisation EFR

Nous étudions le rôle joué par la transformation dans la mise à conformité des données aux hypothèses. Au fil des itérations de cet algorithme, les i-vectors d'apprentissage tendent vers un modèle théorique. Nous montrons ici la convergence des observations vers ce modèle qui possède un certain nombre de propriétés, répondant aux problématiques levées par l'analyse du chapitre précédent. Nous recensons ensuite un certain nombre de propriétés supplémentaires de ce modèle théorique, qui expliquent plus complètement le pouvoir discriminant des vecteurs après transformation EFR.

4.2.1 Convergence vers le modèle standard

Après chaque itération de l'algorithme précédent, les vecteurs sont normalisés en longueur, mais plus nécessairement standardisés. La propriété intuitive suivante a été constatée empiriquement, sur l'ensemble des jeux de données dont nous disposons :

Propriété : lorsque $i \to \infty$,

- la moyenne globale μ_i tend vers 0,
- la matrice de covariance totale $\mathbf{\Sigma}_i$ tend vers $\mathbf{I}$.

les convergences étant strictement monotones (la distance diminue strictement à chaque itération).

Démontrer cette propriété ne semble pas possible dans la mesure de nos connaissances et de la documentation accessible. Nous présentons ici sa validation expérimentale sur plusieurs jeux de données de i-vectors : LIA-hommes, BUT-hommes, BUT-femmes (détaillés en annexe A) produits par notre laboratoire ou fournis par le Brno University of Technology BUT. L'utilisation de plusieurs jeux de données issus de différentes configurations nous a aidé à vérifier la pertinence des résultats.

i) convergence de la moyenne μ_i

Les tables 4.1, 4.2, 4.3 affichent les normes de la moyenne globale des vecteurs pour chacun des jeux d'apprentissage, avant application de l'algorithme (initial "itération 0") puis après chaque itération de l'algorithme jusqu'à la $4^{\text{ème}}$ itération. Chaque fois, la dimension p de l'espace des i-vectors est rappelée. La norme de l'ensemble des i-vectors étant égale en sortie de l'algorithme à $\sqrt{p}$ (elle est de 1, puis les vecteurs sont multipliés par $\sqrt{p}$ pour simplifier les démonstrations qui suivent), comparer la norme de la moyenne μ_i à cette valeur $\sqrt{p}$ permet d'estimer de manière absolue la position de μ_i par rapport à la surface d'hypersphère d'origine 0 contenant les i-vectors.

LIA $p = 400$					
Après chaque itération, normes des i-vectors $= \sqrt{p} = \sqrt{400} = 20$					
Itération i	0	1	2	3	4
$\|\mu_i\|$	0.029	0.12	0.015	0.0023	0.00042

TABLE 4.1 – *Convergence de la moyenne globale vers 0 (LIA).*

BUT hommes $p = 600$					
Après chaque itération, normes des i-vectors $= \sqrt{p} = \sqrt{600} = 24.49$					
Itération i	0	1	2	3	4
$\|\mu_i\|$	10.95	0.087	0.012	0.0023	0.00054

TABLE 4.2 – *Convergence de la moyenne globale vers 0 (BUT-hommes).*

Pour chacun des jeux de données, la moyenne converge bien vers 0. De plus, cette convergence est chaque fois strictement monotone.

Par exemple, pour le troisième jeu de données (table 4.3), la norme de la moyenne est négligeable après deux itérations (0.083) en regard du rayon de l'hypersphère contenant les i-vectors (égal à 24.49).

Il découle deux remarques de ces tables :

- le comportement des données homme ou femme de BUT, issues des mêmes configurations GMM-UBM et extracteur d'i-vectors, est identique vis à vis de la convergence de la moyenne vers 0. Aucun biais n'est constaté suivant le genre.
- les normes initiales ("itération 0") diffèrent considérablement entre les données issues de deux extracteurs distincts (LIA et BUT). Mais nous verrons plus bas qu'elles doivent être comparées aux variances totales des données initiales.

ii) Convergence de la matrice de covariance $\mathbf{\Sigma}_i$ vers la matrice-identité $\mathbf{I}$

Nous mesurons la distance entre $\mathbf{\Sigma}_i$ et $\mathbf{I}$, au sens de la norme de Frobenius[2]. La norme de Frobenius d'une matrice $\mathbf{A}$ est définie par $\|\mathbf{A}\| = Tr\left(\sqrt{\mathbf{A}^t\mathbf{A}}\right)$ où $Tr\,()$ est l'opérateur Trace.

La distance entre $\mathbf{\Sigma}_i$ et $\mathbf{I}$ est alors mesurée par son ratio d'écart à $\mathbf{I}$ au sens des moindres carrés :

2. L'espace euclidien des matrices $p \times p$ étant de dimension finie, toutes les normes y sont équivalentes, et nous choisissons la plus usuelle.

BUT femmes $p = 600$					
Après chaque itération, normes des i-vectors $= \sqrt{p} = \sqrt{600} = 24.49$					
Itération i	0	1	2	3	4
$\|\mu_i\|$	11.58	0.083	0.010	0.0022	0.00070

TABLE 4.3 – *Convergence de la moyenne globale vers 0 (BUT-femmes).*

LIA $p = 400$					
Itération i	0	1	2	3	4
$\varepsilon(\mathbf{\Sigma}_i, \mathbf{I})$	0.99	0.10	0.013	0.0022	0.00041

TABLE 4.4 – *Convergence de la covariance vers l'identité (LIA).*

BUT hommes $p = 600$					
Itération i	0	1	2	3	4
$\varepsilon(\mathbf{\Sigma}_i, \mathbf{I})$	0.44	0.081	0.013	0.0027	0.00065

TABLE 4.5 – *Convergence de la covariance vers l'identité (BUT-hommes).*

$$\varepsilon(\mathbf{\Sigma}_i, \mathbf{I}) = \frac{\|\mathbf{\Sigma}_i - \mathbf{I}\|}{\|\mathbf{I}\|} = \sqrt{\frac{1}{p} Tr\left((\mathbf{\Sigma}_i - \mathbf{I})^2\right)} = \sqrt{\frac{1}{p} \sum_k \sum_l (\mathbf{\Sigma}_i - \mathbf{I})^2_{k,l}} \tag{4.1}$$

Les valeurs $\varepsilon(\mathbf{\Sigma}_i, \mathbf{I})$ sont indiquées sur les tables 4.4, 4.5, 4.6 pour les trois mêmes jeux que précédemment. Pour chacun des jeux de données, la matrice de covariance des données d'apprentissage tend bien vers la matrice-identité. De plus, cette convergence est chaque fois strictement monotone. Là encore, les comportements des données par genre sont comparables.

4.2.2 Gaussianité

Nous nous intéressons ici à la gaussianité des données, c'est à dire à leur degré de similarité avec un échantillon gaussien. L'analyse du chapitre précédent a montré que l'évolution de la gaussianité entre des jeux données soumis à une transformation pouvait être évaluée par la loi du χ^2. Nous reprenons donc les mesures effectuées précédemment, au fil des itérations de la phase d'apprentissage.

La figure 4.2 présente les histogrammes successifs, avant puis après 1 à 5 itérations de la normalisation EFR, des normes carrées des jeux de données standardisés d'apprentissage et d'évaluation BUT-hommes, ainsi que la densité du χ^2 à p degrés de liberté (ici $p = 600$).

$\Diamond$ Figure 4.2, graphique n° 1 :

Ce premier graphique, reproduction du graphique 3.4 du chapitre précédent, correspond aux données initiales (avant EFR). Rappelons que :

BUT femmes $p = 600$					
Itération i	0	1	2	3	4
$\varepsilon(\mathbf{\Sigma}_i, \mathbf{I})$	0.43	0.070	0.010	0.0029	0.0010

TABLE 4.6 – *Convergence de la covariance vers l'identité (BUT-femmes).*

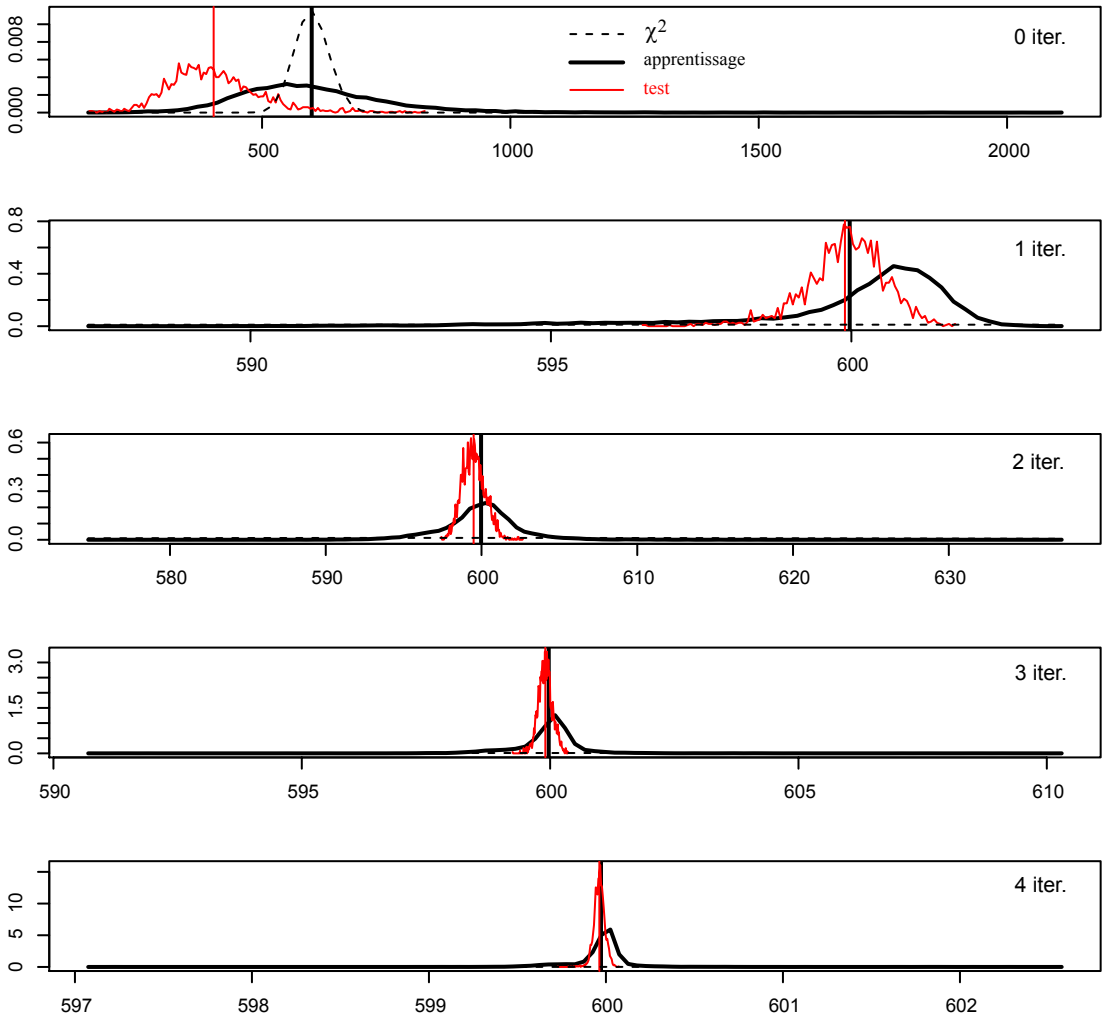

FIGURE 4.2 – *Histogrammes successifs des normes carrées des jeux de données standardisés d'apprentissage et d'évaluation BUT-hommes, et densité du χ^2 à p degrés de liberté ($p = 600$).*

- la courbe en pointillé est celle de la distribution théorique du χ^2 à p degrés de liberté. On rappelle que sa moyenne est p (ici 600) et son écart-type $\sqrt{2p}$ (ici $\approx$ 34.64).
- la courbe noire en gras est la distribution empirique des normes carrées des données d'apprentissage standardisées (la distribution S de l'équation 3.11).
- la courbe rouge est la distribution empirique des normes carrées des données d'évaluation standardisées.

Comme nous l'avons constaté au chapitre précédent, les données présentent plusieurs anomalies :

- asymétrie de la la distribution empirique (preuve de faible gaussianité)
- densité de l'apprentissage plus "plate" qu'un $\chi^2(p)$ c'est à dire que la proportion de valeurs extrêmes est plus importante qu'en situation gaussienne.
- densité empirique de l'évaluation plus proche en courbure d'un $\chi^2(p)$, mais présentant un décalage net de valeurs : la moyenne, indiquée par la droite verticale en rouge, est de l'ordre de 400, très inférieure aux 600 théoriques. Cela tient essentiellement à un biais entre les moyenne globales μ et μ^{test} des jeux de données.

$\Diamond$ Figure 4.2, graphique n° 2 :

Une itération de l'algorithme de normalisation EFR a été pratiquée et les distributions empiriques re-calculées. La procédure effectuée a été :

$$\text{Calculer les } w' = \frac{\mathbf{\Sigma}^{-\frac{1}{2}}(w-\mu)}{\left\|\mathbf{\Sigma}^{-\frac{1}{2}}(w-\mu)\right\|}$$
$$\text{Actualiser les paramètres de moyenne et covariance } (\mu, \mathbf{\Sigma}) \rightarrow (\mu', \mathbf{\Sigma}')$$
$$\text{Calculer } S' = \left\{ \left\| (\mathbf{\Sigma}')^{-\frac{1}{2}} \left(w_i' - \mu' \right) \right\|^2 \right\}_{i=1,\dots n}$$

Les distributions empiriques se sont nettement condensées autour de la moyenne : l'échelle des valeurs s'étend entre 595 et 603. A cette échelle, la distribution du $\chi^2(p)$ n'est plus visible sur le graphique. La distribution de l'apprentissage présente encore une asymétrie, celle de l'évaluation est quasiment symétrique et, surtout, correctement centrée autour de la valeur théorique p.

$\Diamond$ Figure 4.2, graphiques n° 3, 4, 5 : au fur et à mesure des itérations 2 à 4 de l'algorithme EFR, les distributions des normes se condensent progressivement autour de p (surveiller les échelles des abscisses : l'écart-type diminue de ± 10 à ± 1). L'asymétrie des distributions s'atténue. Ce résultat était prévisible : la convergence de μ_i et $\mathbf{\Sigma}_i$ vers 0 et $\mathbf{I}$, montrée au paragraphe précédent, entraîne que, pour toute observation w :

$$\left\| \mathbf{\Sigma}_{iter}^{-\frac{1}{2}} (w - \mu_{iter}) \right\|^2 \underset{i\rightarrow\infty}{\rightarrow} \left\| \Sigma^{-\frac{1}{2}} (w-0) \right\|^2 = \|w\|^2 = p = E\left[\chi^2(p)\right] \tag{4.2}$$

donc une forte proximité des carrés des normes des vecteurs standardisés avec le mode théorique.

Les normes carrées des i-vectors standardisés après normalisation EFR ne coïncident pas totalement avec une loi du $\chi^2(p)$. Mais elles s'en rapprochent nettement, en terme de moyenne p de cette loi, de symétrie et surtout d'adéquation entre les jeux d'apprentissage et d'évaluation : les i-vectors à évaluer ont été entraînés par l'algorithme vers la distribution de l'apprentissage et ces derniers vers une distribution gaussienne standard. Cette forme de conditionnement atténue fortement les dissemblances de moyenne et covariance entre apprentissage et évaluation, homogénéïsant ceux-ci suivant une allure gaussienne.

Dans (Garcia-Romero and Espy-Wilson, 2011), il est montré qu'après blanchiment (comme la standardisation l'effectue) et normalisation de longueur des données, le phénomène suivant se produit lorsqu'est pratiquée une "heavy-tailed" PLDA : le nombre de degrés de liberté du sous-espace de voix propres (*eigenvoices*) augmente nettement après cette normalisation alors que celui du sous-espace résiduel (*eigenchannels*) tend à diminuer. Ceci signifie que les voix propres (*eigenvoices*) ont une distribution à queue de probabilité moins forte, donc plus proche d'être gaussienne, le caractère non-gaussien se reportant plutôt sur les informations résiduelles.

L'étude de gaussianité montre que l'idée d'unitarisation des vecteurs, initialement introduite de manière implicite dans le score-cosinus, peut permettre de gaussianiser artificiellement les données, mais ce à condition de l'appliquer à des vecteurs préalablement "blanchis". La standardisation constitue un moyen efficace et statistiquement valide de réaliser cette opération.

En annexe C sont présentées les mêmes figures que la figure 4.2 pour les jeux de données BUT-femmes et LIA-hommes. Le premier permet notamment de vérifier le comportement similaire des distributions pour les fichiers féminins et donc l'indépendance au genre des constats précédents.

4.3 Propriétés générales de la normalisation EFR

Nous présentons dans la suite un certain nombre de propriétés vérifiées par tout jeu de données vectoriel simultanément de même norme $\sqrt{p}$ et standard (moyenne 0 et matrice de covariance égale à la matrice identité $\mathbf{I}$). Après l'algorithme de normalisation EFR précédent, les i-vectors vérifient presqu'exactement ces deux conditions, comme nous l'avons empiriquement montré. Toutefois, la quasi-validité de chacune de ces propriétés sur les i-vectors empiriques doit être ensuite vérifiée. Chaque propriété présentée est donc démontrée sur un modèle de données EFR théorique, puis sa validité est mesurée sur nos jeux de données expérimentaux.

Soit $\mathcal{X} = \{w_i\}_{i=1,\dots n}$ un jeu de n vecteurs de dimension p vérifiant les propriétés suivantes :

- la moyenne μ de $\mathcal{X}$ est égale à 0,
- la matrice de covariance $\mathbf{\Sigma}$ de $\mathcal{X}$ est la matrice $p \times p$ identité $\mathbf{I}$,

– $\mathcal{X}$ s'étend sur l'hypersphère de rayon $\sqrt{p}$ (tous les vecteurs de $\|w\|$ sont de norme $\sqrt{p}$).

4.3.1 Uniformité des probabilités

Propriété : les probabilités des observations, sous l'hypothèse de gaussianité la plus vraisemblable, sont égales.

En effet, les paramètres de la loi gaussienne la plus vraisemblable pour $\mathcal{X}$ sont sa moyenne $\mu = 0$ et sa matrice de covariance $\mathbf{I}$. Pour tout w de $\mathcal{X}$

$$\begin{aligned} P(w) &= (2\pi)^{-\frac{p}{2}} |\mathbf{I}|^{-\frac{1}{2}} \exp\left(-(w-\mu)^t \mathbf{I}^{-1} (w-\mu)\right) \\ &= (2\pi)^{-\frac{p}{2}} \exp\left(-\|w\|^2\right) = (2\pi)^{-\frac{p}{2}} \exp(-p) \end{aligned} \tag{4.3}$$

Ceci induit une propriété intéressante pour la suite :

Etant donnés deux i-vectors indépendants w_1 et w_2 de $\mathcal{X}$, leur probabilité conjointe sous l'hypothèse θ_{non} :

θ_{non} : "w_1 et w_2 n'appartiennent pas au même locuteur"

est constante. En effet :

$$P(w_1, w_2|\theta_{\text{non}}) = P(w_1) P(w_2) = (2\pi)^{-p} \exp(-2p) \tag{4.4}$$

Nous exploiterons cette propriété lors de la détermination d'un score de proximité entre les observations.

4.3.2 Dispersion maximale

Propriété : la dispersion de $\mathcal{X}$ est maximale sur la surface de l'hypersphère

L'indicateur de dispersion d'un jeu de données $\mathcal{X}$ est son écart-type $\sigma(\mathcal{X})$, racine carrée de sa variance, défini par :

$$\sigma(\mathcal{X}) = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (w_i - \mu)^t (w_i - \mu)} \tag{4.5}$$

qui est aussi la racine carrée de la trace de sa matrice de covariance :

$$\sigma(\mathcal{X}) = \sqrt{Tr(\mathbf{\Sigma}_{\mathcal{X}})} \tag{4.6}$$

LIA-hommes $p = 400$					
Valeur maximale de dispersion sur l'hypersphère $= \sqrt{p} = \sqrt{400} = 20$					
Itération i	0	1	2	3	4
$\sigma(\mathcal{X})$	0.39	19.99	19.99	20	20

TABLE 4.7 – *Convergence de l'écart-type (racine de la variance) des données d'apprentissage LIA-hommes vers son maximum $\sqrt{p}$ au fur et à mesure des itérations d'EFR*

BUT-hommes $p = 600$					
Valeur maximale de dispersion sur l'hypersphère $= \sqrt{p} = \sqrt{600} = 24.49$					
Itération i	0	1	2	3	4
$\sigma(\mathcal{X})$	22.53	24.49	24.49	24.49	24.49

TABLE 4.8 – *Convergence de l'écart-type (racine de la variance) des données d'apprentissage BUT-hommes vers son maximum $\sqrt{p}$ au fur et à mesure des itérations d'EFR*

L'écart-type indique le rayon de l'hypersphère dans laquelle est concentré l'essentiel des observations.

dém. : si $\mathcal{X}$ s'étend sur la surface de l'hypersphère de rayon $\sqrt{p}$, alors $\sigma^2(\mathcal{X})$ vérifie :

$$\begin{aligned}\sigma^2(\mathcal{X}) &= \frac{1}{n}\sum_{i=1}^{n} w_i^t w_i - \mu^t\mu \\ &= \frac{1}{n}\sum_{i=1}^{n} \|w_i\|^2 - \|\mu\|^2 = 1 - \|\mu\|^2 \end{aligned} \tag{4.7}$$

Cette valeur est maximale si et seulement si $\mu = 0$, ce qui est le cas ici par hypothèse.

◊ Vérification expérimentale sur les jeux de données d'apprentissage :

Les tables 4.7, 4.8 et 4.9 montrent les valeurs de l'écart-type de chacun des jeux de données utilisés précédemment après 0 à 4 itérations de l'algorithme. Chaque fois la dispersion maximale possible sur l'hypersphère est indiquée ($\sqrt{p} = \sqrt{400} = 20$ par exemple pour LIA-hommes).

Les dispersions de chacun des jeux de données tendent rapidement vers la valeur maximale. La convergence est quasiment achevée après deux itérations. Notons que

BUT-femmes $p = 600$					
Valeur maximale de dispersion sur l'hypersphère $= \sqrt{p} = \sqrt{600} = 24.49$					
Itération i	0	1	2	3	4
$\sigma(\mathcal{X})$	22.43	24.49	24.49	24.49	24.49

TABLE 4.9 – *Convergence de l'écart-type (racine de la variance) des données d'apprentissage BUT-femmes vers son maximum $\sqrt{p}$ au fur et à mesure des itérations d'EFR*

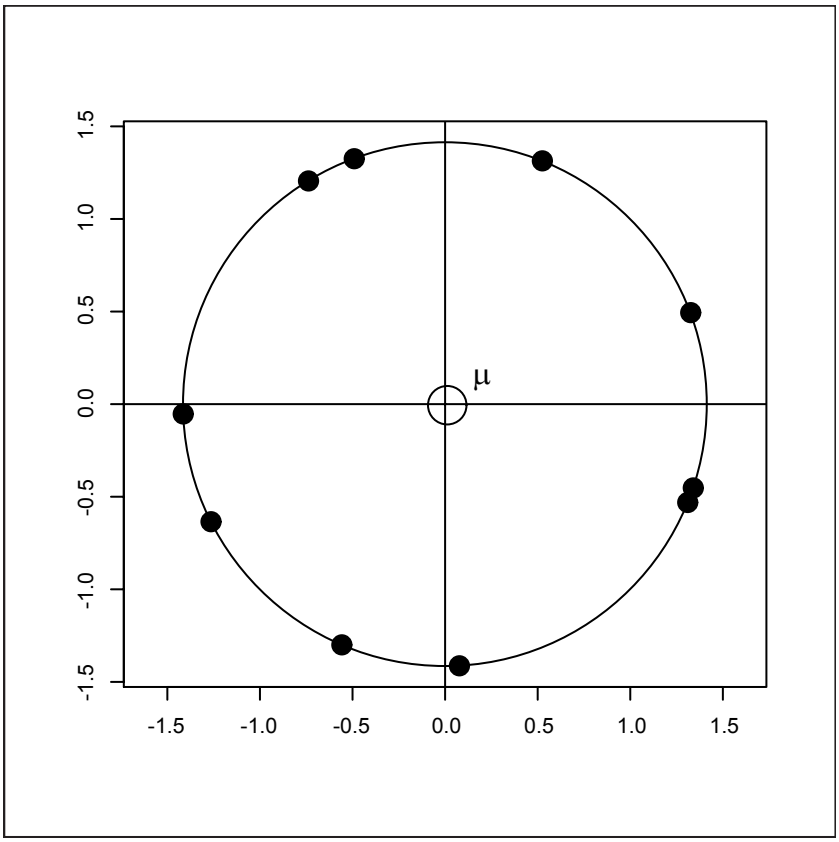

FIGURE 4.3 – *Nuage de points en 2D de moyenne nulle, matrice de covariance identité, normes des vecteurs égales à $\sqrt{p} = \sqrt{2}$ et donc variance maximale. Comme on peut le remarquer, les densités de points par régionnement uniforme sur la surface ne sont pas égales.*

les comportements des données hommes ou femmes d'un même extracteur sont identiques, vis à vis de la dispersion.

Par conséquent, les données d'apprentissage qui ont été utilisées par l'algorithme EFR pour calculer à chaque itération les moyenne et matrice de covariance forment une "couverture maximale" de l'espace de représentation, qui est ici une variété de surface finie. Pour éviter une image erronée, remarquons toutefois que dispersion maximale ne signifie pas que les données tapissent uniformément la surface. En effet, la gaussianité des données est pré-supposée. La figure 4.3 donne un exemple en dimension 2 d'un mini-nuage $\mathcal{X}$ de dix points vérifiant les caractéristiques précédentes : moyenne nulle, matrice de covariance identité, normes des vecteurs toutes égales à $\sqrt{p} = \sqrt{2}$ et donc variance maximale. Comme le montre cette figure, les densités de points par régionnement uniforme sur la surface ne sont pas égales.

4.3.3 Correspondances entre les décompositions en valeurs singulières locuteur et session

Propriété : les vecteurs propres des matrices de covariance intra- et inter-locuteur **W** et **B** sont exactement complémentaires : le 1$^{\text{er}}$ vecteur propre de B, associé à la plus grande valeur propre, est aussi le dernier de **W**, associé à la plus petite valeur propre et plus généralement :

$$\forall j = 1, ...p, \qquad v_j^{[\mathbf{B}]} = v_{p+1-j}^{[\mathbf{W}]} \tag{4.8}$$

où $\left\{v^{[\mathbf{B}]}\right\}_j$ (resp. $v^{[\mathbf{W}]}$) dénote les vecteurs propres de **B** (resp. **W**) dans l'ordre décroissant de leurs valeurs propres.

Propriété : les valeurs propres des matrices de covariance intra- et inter-locuteur **W** et **B** sont exactement complémentaires. La somme de la 1$^{\text{ère}}$ valeur propre de **B** (i.e. la plus grande) et de la dernière de **W** (i.e. la plus petite) est égale à 1 et plus généralement :

$$\forall j = 1, ...p, \qquad \lambda_j^{[\mathbf{B}]} + \lambda_{p+1-j}^{[\mathbf{W}]} = 1 \tag{4.9}$$

où $\left\{\lambda^{[\mathbf{B}]}\right\}_j$ (resp. $\lambda^{[\mathbf{W}]}$) dénotent les valeurs propres de **B** (resp. **W**) dans l'ordre décroissant.

Dém. : notant $\mathbf{\Sigma}$ la matrice de covariance totale de $\{w_i\}_{i=1,...n}$, le théorème de décomposition des variances par classes permet d'écrire

$$\mathbf{\Sigma} = \mathbf{B} + \mathbf{W} \tag{4.10}$$

Comme $\mathbf{\Sigma} = \mathbf{I}$, on en déduit que tout vecteur propre v de **B** pour une valeur propre λ vérifie

$$\begin{aligned} \mathbf{B}v &= \lambda v && (4.11)\\ &= (\mathbf{\Sigma} - \mathbf{W})\, v = (\mathbf{I} - \mathbf{W})\, v && (4.12)\\ &= v - \mathbf{W}v && (4.13) \end{aligned}$$

et donc que

$$\mathbf{W}v = (1 - \lambda)\, v \tag{4.14}$$

v est donc vecteur propre de **W** pour la valeur propre $(1 - \lambda)$.

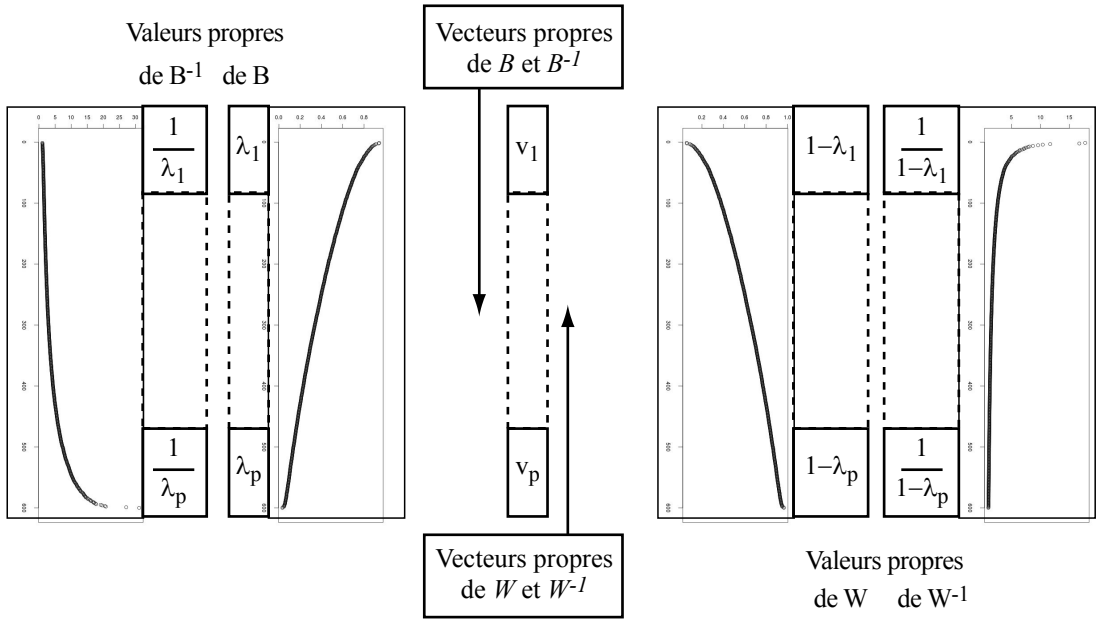

FIGURE 4.4 – *Correspondances entre vecteurs et valeurs propres de B et W après EFR.*

Les vecteurs propres de **B** et **W** sont donc les mêmes, mais dans l'ordre inverse des valeurs propres. Et les spectres de valeurs propres de **B** et **W** vérifient donc la propriété énoncée.

Ces correspondances, que nous exploiteront par la suite, sont décrites sur la figure 4.4.

La base de vecteurs propres $\mathcal{V} = \{v_j\}_{j=1,...,p}$ de **B** est aussi celle de **W**, en ordre inverse du classement des valeurs propres. Notons que les bases de vecteurs propres des matrices inverses de **B** et **W** sont aussi égales à $\mathcal{V}$. La figure 4.4 rappelle les énergies correspondantes de $\mathbf{B}^{-1}$ et $\mathbf{W}^{-1}$ (i.e. les spectres de valeurs propres) qui seront souvent exploitées lors de la phase de scoring.

◊ Vérification expérimentale sur les jeux de données d'apprentissage :

Vecteurs propres : pour tout j de $1, ..., p$ on souhaite étudier la proximité entre le $j^{\text{ème}}$ vecteur propre $v_j^{[\mathbf{B}]}$ de **B** et le $(p+1-j)^{\text{ème}}$vecteur propre $v_{p+1-j}^{[\mathbf{W}]}$ de **W**. Les deux vecteurs étant de norme 1, leur cosinus carré est égal au produit scalaire :

$$\cos^2\left(v_j^{[\mathbf{B}]}, v_{p+1-j}^{[\mathbf{W}]}\right) = \left(v_j^{[\mathbf{B}]}.v_{p+1-j}^{[\mathbf{W}]}\right)^2 \tag{4.15}$$

L'utilisation du carré évite le piège de la décomposition en valeurs singulières, qui peut indifféremment renvoyer un vecteur propre ou son opposé. Cette valeur est dans $[0, 1]$ et une valeur de 1 indique une parfaite égalité entre les deux vecteurs.

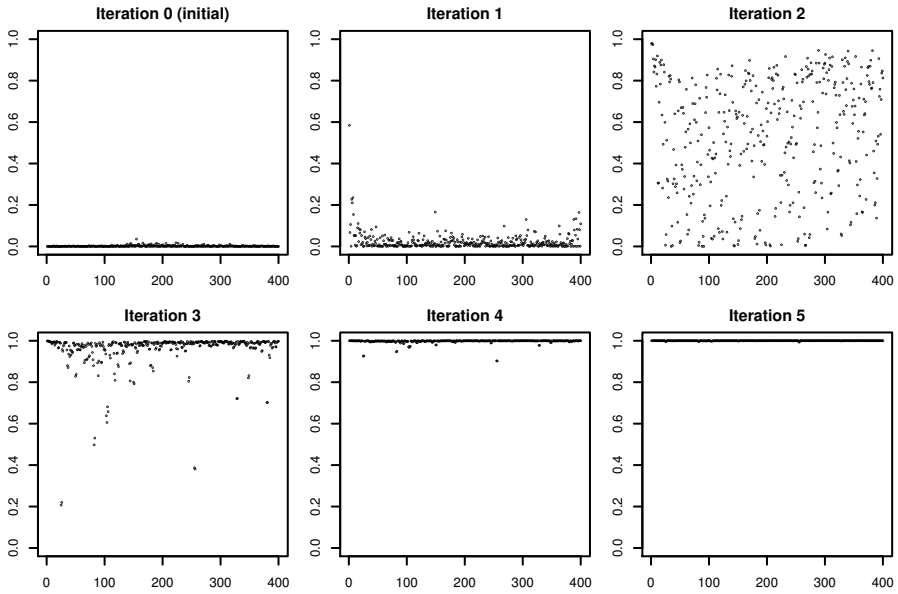

FIGURE 4.5 – *Convergence empirique des vecteurs propres de* **B** *et* **W** *durant EFR.*

La figure 4.5 affiche la série des $p = 400$ valeurs de cosinus carrés $\left(v_j^{[\mathbf{B}]}.v_{p+1-j}^{[\mathbf{W}]}\right)^2$ du jeu de données d'apprentissage LIA-hommes, initialement puis après chaque itération de l'algorithme EFR jusqu'à la 5$^{\text{ème}}$ itération. Les couples de vecteurs propres B et **W** sont initialement décorrélés, puis tendent vers la valeur optimale commune de proximité, égale à 1, jusqu'à une convergence quasiment parfaite au bout de 5 itérations.

Cette valeur de 5 itérations peut constituer une borne pour l'algorithme itératif si celui-ci est utilisé avant une technique de compensation de l'effet-session basée sur ces matrices **B** et **W**.

Valeurs propres : de façon similaire, la correspondance théorique entre les valeurs propres de **B** et **W** peut être estimée sur les données d'apprentissage en affichant la série des valeurs :

$$\left\{\left(\lambda_j^{[\mathbf{B}]}, \lambda_{p+1-j}^{[\mathbf{W}]}\right), j = 1, ..., p\right\} \tag{4.16}$$

Puisque la somme de deux de ces valeurs doit être égale à 1, ce nuage de points doit se rapprocher de la droite $y = 1 - x$.

La figure 4.6 affiche ces nuages de points pour le jeu LIA-hommes, initialement puis après 1 et 2 itérations de l'algorithme. La droite $y = 1 - x$ est tracée en pointillés. La propriété, non vérifiée initialement (itération 0), est rapidement validée expérimentalement. La correspondance théorique entre les valeurs propres de **B** et **W** est quasiment achevée à la 2$^{\text{ème}}$ itération.

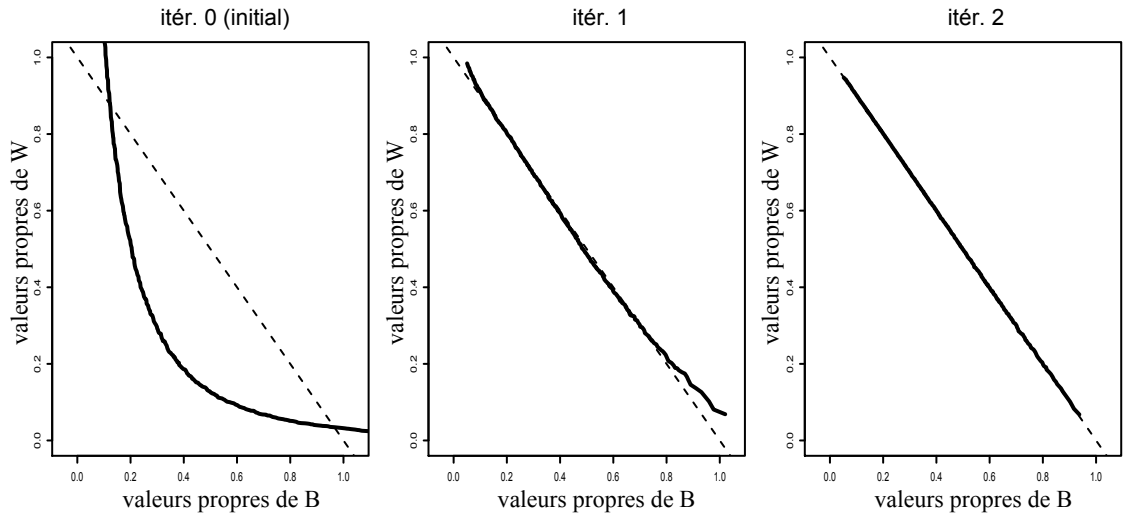

FIGURE 4.6 – *Convergence empirique des valeurs propres de* **B** *et* **W** *durant EFR.*

4.3.4 Equivalence des techniques LDA et NAP

Etant fixé un rang r ($r < p$), la méthode NAP (Campbell et al., 2006) de compensation de l'effet session soustrait à chaque vecteur sa projection sur le sous-espace principal des r premiers vecteurs propres de $\mathbf{W}$.

Propriété : les techniques de projection LDA et NAP sont équivalentes.

Dém. : comme nous l'avons indiqué au 2.6.1.3, la solution du problème d'optimisation de la LDA (maximiser le quotient de Rayleigh) est constituée des vecteurs propres de la matrice $\mathbf{B}\mathbf{W}^{-1}$ dans l'ordre décroissant de ses valeurs propres. Dénotant v le vecteur propre de $\mathbf{B}$ associé à sa plus grande valeur propre λ, v est vecteur propre de $\mathbf{W}$ pour la valeur propre $1-\lambda$ et donc également de $\mathbf{W}^{-1}$ pour la valeur propre $\frac{1}{1-\lambda}$. On a :

$$\mathbf{W}^{-1}\mathbf{B}v = \frac{\lambda}{1-\lambda}v \tag{4.17}$$

ce qui montre que v est également vecteur propre de $\mathbf{W}^{-1}\mathbf{B}$. De plus, la matrice $\mathbf{B}$ étant définie positive, toutes ses valeurs propres sont positives et elles sont également inférieures à 1. En effet, ces valeurs propres sont les variances inter-locuteur par dimension, qui sont nécessairement inférieures aux variances totales par dimension. Ces dernières sont toutes égales à 1, la matrice de variabilité totale étant l'identité I. Or, $\frac{\lambda}{1-\lambda}$ est croissant avec λ, pour tout λ de $[0,1]$, donc maximal quand λ est maximal. Il

s'en déduit que les vecteurs propres de **B** sont aussi ceux de $\mathbf{W}^{-1}\mathbf{B}$, dans le même ordre des valeurs propres décroissantes.

B et **W** partagent la même base de vecteurs propres en ordre inverse et la transformation NAP revient donc à projeter sur le sous-espace principal de rang r de **B**.

Les points précédents montrent que les méthodes NAP et LDA sont strictement identiques sous nos hypothèses, projetant les vecteurs sur le même sous-espace principal issu de la matrice **B**. La comparaison des techniques de compensation de l'effet session NAP et LDA a occupé les premières recherches sur les i-vectors (Dehak et al., 2009). La normalisation des vecteurs par un algorithme de type EFR met fin à leur concurrence, en portant les données vers une configuration optimale. Celle-ci satisfait le double critère de supprimer la part la plus importante possible de variabilité nuisible, tout en conservant au mieux la variabilité-cible du locuteur.

4.3.5 Optimalité de la LDA

La propriété précédente a une conséquence fondamentale :

La solution de la LDA, comme nous l'avons vu ci-dessus, est constituée des r premiers vecteurs propres de B. Or ceux-ci sont également les r derniers vecteurs propres de W. Ceci implique que :

Propriété : dans le quotient de Rayleigh,

- le numérateur est maximal. En effet $v^t\mathbf{B}v = \lambda$ est maximal si v est le vecteur propre de **B** associé à sa plus grande valeur propre λ,

- le dénominateur est minimal. En effet, étant donnée la correspondance entre les spectres, $v^t\mathbf{W}v = \lambda$ est minimal si v est le vecteur propre de **B** associé à sa plus petite valeur propre λ,

et donc le sous-espace de projection maximise le quotient de Rayleigh.

Ce fait est important : le critère d'optimisation du problème de séparation des variances (intra- et inter-locuteur) a été conçu comme un quotient parce qu'il n'existait pas a priori de solution au problème double d'optimisation : maximiser la variance inter-locuteur tout en minimisant la variance intra-locuteur.

Le nuage de points standardisé et unitarisé répond, lui, parfaitement à cette double contrainte. Les figures 4.7, 4.8 et 4.9 montrent le graphe spectral **B** du même jeu de données BUT-hommes que celui utilisé sur la figure 3.3, après une (figure 4.7), deux (figure 4.8) puis trois itérations (figure 4.9) de l'algorithme de normalisation EFR. Chaque fois, les variances spectrales sont représentées dans la base de vecteurs propres de **B**. La convergence de la matrice de covariance totale $\mathbf{\Sigma}_i$ vers **I** est quasiment achevée à la troisième itération. Après cette normalisation, les variances locuteur et résiduelle présentent un fort degré de complémentarité, les premiers vecteurs propres de B (à gauche) contenant la part très majoritaire de variabilité locuteur et une part devenue mineure

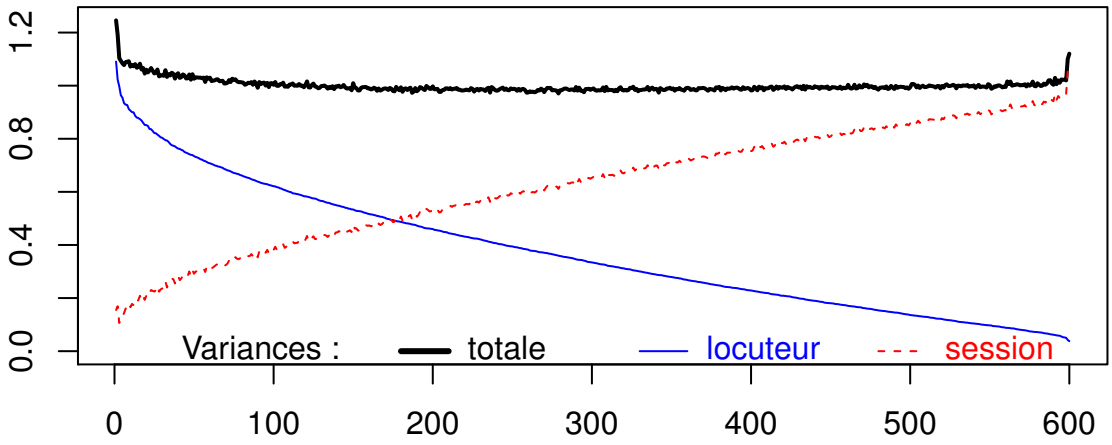

FIGURE 4.7 – *Graphe spectral-***B** *du jeu de données BUT-hommes après 1 itération d'EFR.*

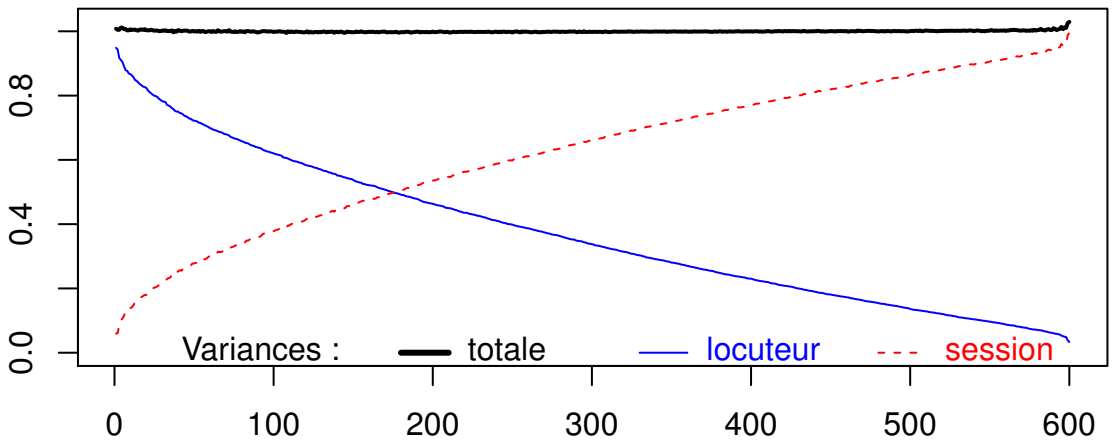

FIGURE 4.8 – *Graphe spectral-***B** *du jeu de données BUT-hommes après 2 itérations d'EFR.*

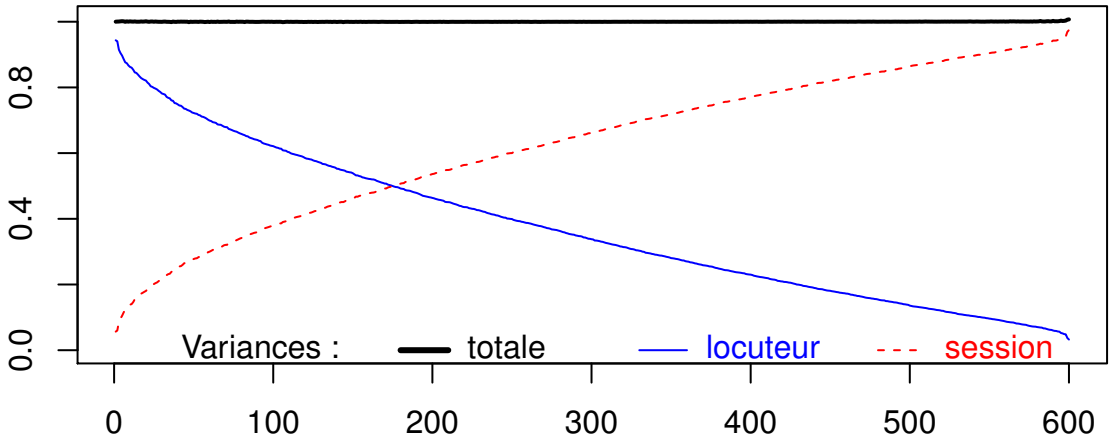

FIGURE 4.9 – *Graphe spectral-***B** *du jeu de données BUT-hommes après 3 itérations d'EFR.*

de variabilité session. Les derniers vecteurs propres contiennent, eux, une part très majoritaire de variabilité session et une part devenue mineure de variabilité locuteur.

Si des méthodes probabilistes comme la PLDA sont considérées actuellement comme les plus performantes pour le traitement des i-vectors, ce résultat montre que des méthodes géométriques gagnent à être précédées de la normalisation EFR, qui optimise leurs critères fondamentaux. Mais nous verrons plus loin qu'un modèle génératif probabiliste comme la PLDA, bien initialisé grâce aux estimations géométriques des variances intra- et inter-locuteur fournies par **B** et **W**, gagne également à exploiter l'optimalité -sur critères déterministes- procurée par la normalisation EFR.

4.3.6 JFA vectorielle (notion d'eigenfactors)

Propriété : le résultat précédent peut s'écrire sous forme d'une décomposition des i-vectors suivant un principe de Joint Factor Analysis. Soit w un i-vector d'un jeu de données $\mathcal{X}$ vérifiant les propriétés précédentes. Soit $\mathbf{P}^{(\mathbf{B})}$ la matrice $p \times p$ des vecteurs propres de **B** et, pour tout $r < p$, la sous-matrice $\mathbf{P}^{(\mathbf{B})}_{(r+1):p}$ des colonnes (i.e. vecteurs propres) $(r+1)^{\text{ème}}$ à $p^{\text{ème}}$ de $\mathbf{P}^{(\mathbf{B})}$. Soit $\mathbf{P}^{(\mathbf{W})}$ la matrice $p \times p$ des vecteurs propres de **W** et, pour tout $r < p$, la sous-matrice $\mathbf{P}^{(\mathbf{W})}_{(p-r):1}$ des colonnes $(p-r)$ à 1 de $\mathbf{P}^{(\mathbf{W})}$. La correspondance entre les spectres de **B** et **W** implique que, pour tout $r < p$:

$$\mathbf{P}^{(\mathbf{B})}_{(r+1):p} = \mathbf{P}^{(\mathbf{W})}_{(p-r):1} \tag{4.18}$$

Le vecteur w est la somme de ses projections sur les sous-espaces engendrés par $\mathbf{P}^{(B)}_{1:r}$ et $\mathbf{P}^{(\mathbf{B})}_{(r+1):p}$, ce qui s'écrit :

$$
\begin{aligned}
w &= \mathbf{P}^{(\mathbf{B})}_{1:p} P^{(\mathbf{B})^t}_{1:p} w \\
&= \mathbf{P}^{(\mathbf{B})}_{1:r} P^{(\mathbf{B})^t}_{1:r} w + \mathbf{P}^{(\mathbf{B})}_{(r+1):p} \mathbf{P}^{(\mathbf{B})^t}_{(r+1):p} w \\
&= \mathbf{P}^{(\mathbf{B})}_{1:r} \left(\mathbf{P}^{(\mathbf{B})^t}_{1:r} w \right) + \mathbf{P}^{(\mathbf{W})}_{(p-r):1} \left(\mathbf{P}^{(\mathbf{W})^t}_{(p-r):1} w \right) \\
&= \mathbf{V}z + \mathbf{U}x
\end{aligned} \tag{4.19}
$$

où

- $\mathbf{V} = \mathbf{P}^{(\mathbf{B})}_{1:r}$ est une matrice $p \times r$ orthogonale,
- $\mathbf{U} = \mathbf{P}^{(\mathbf{W})}_{(p-r):1}$ est une matrice $p \times (p-r)$ également orthogonale,
- $z = P^{(\mathbf{B})^t}_{1:r} w$ est un vecteur de dimension r de moyenne $E[z] = E\left[P^{(\mathbf{B})^t}_{1:r} w\right] = P^{(\mathbf{B})^t}_{1:r} E[w] = 0$ et de covariance :

$$E\left[zz^t\right] = E\left[\mathbf{P}^{(\mathbf{B})^t}_{1:r} w\right] = \frac{1}{n} \mathbf{P}^{(\mathbf{B})^t}_{1:r} w w^t \mathbf{P}^{(\mathbf{B})}_{1:r} = \mathbf{P}^{(\mathbf{B})^t}_{1:r} \mathbf{I}\, \mathbf{P}^{(\mathbf{B})}_{1:r} = \mathbf{I},$$

- $x = \mathbf{P}^{(\mathbf{W})^t}_{(p-r):1} w$ est un vecteur de dimension$(p-r)$, de mêmes moyenne covariance.

La matrice de covariance $cov(w)$ des i-vectors est égale à $\mathbf{I}$ mais aussi à :

$$
\begin{aligned}
cov(w) &= E\left[(\mathbf{V}z + \mathbf{U}x)(\mathbf{V}z + \mathbf{U}x)^t\right] \\
&= \mathbf{V}\mathbf{V}^t + \mathbf{V}\mathbf{U}^t + \mathbf{U}\mathbf{V}^t + \mathbf{U}\mathbf{U}^t
\end{aligned} \tag{4.20}
$$

Puisque $\mathbf{P}^{(\mathbf{B})}$ est orthogonale, $\mathbf{V}\mathbf{U}^t = \mathbf{P}^{(\mathbf{B})}_{1:r} \mathbf{P}^{(\mathbf{B})^t}_{r:(p+1-r)} = 0$ et de même $\mathbf{U}\mathbf{V}^t = 0$. De plus, comme le montre la figure 4.9, la majorité de la variabilité de $\mathbf{V}z$ est attribuable au locuteur et la majorité de celle de $\mathbf{U}x$ à la session (résidu). Le sous-espace engendré par $\mathbf{V}$ est de dimension r et peut constituer un sous-espace initial de "voix propres" (*eigenvoices*). Celui engendré par $\mathbf{U}$, de dimension $p-r$, peut constituer de même un sous-espace initial de "canaux propres" (*eigenchannels*). Ces deux sous-espaces sont des complémentaires orthogonaux.

Le fait qu'une même base de l'espace des i-vectors contienne de possibles *eigenvoices* dans ses r premières dimensions et *eigenchannels* dans ces $(p-r)$ dernières est suffisamment important pour introduire ici la notion de "facteurs propres" (*eigenfactors*) qui donnent leur nom à notre algorithme : l'algorithme EFR fournit une spatialisation des i-vectors commune aux deux variabilités locuteur et session et une solution vectorielle initiale à un problème de décomposition de ces variances en facteurs joints (initialisation par une solution géométrique d'un problème de JFA).

Nous verrons dans la suite de ce chapitre que de telles propriété des algorithmes de normalisation sur la sphère peuvent être exploitées pour initialiser la PLDA gaussienne.

4.3.7 Généralisation des métaparamètres

L'objectif de la normalisation est de transformer les données pour les préparer à une technique de compensation de l'effet-session. Plus précisément d'extraire de l'apprentissage, après transformation, des métaparamètres plus précis en termes de discrimination du locuteur. Mais la qualité du résultat de la discrimination dépend essentiellement de la capacité à appliquer ces paramètres aux données de test. Ceci suppose que les données d'apprentissage ont été correctement choisies, reflétant au mieux les caractéristiques des données de test et que leurs résultats sont généralisables. Dans le sens où nous transformons ces données, la question se pose de savoir si ces dernières, après transformation, suivent avec une bonne précision les paramètres et lois issus de l'apprentissage.

Ainsi, les techniques de compensation s'appuient sur la convergence des données unitarisées vers une loi standard. Nous pouvons mesurer l'adéquation de données de tests à ce modèle en calculant les moyenne globale et matrice de covariance d'un grand jeu de données d'évaluation *étalon*. Les organismes comme NIST en fournissent. Il est entendu que rien n'indique que ces données soient constituées aléatoirement, ni qu'elles coïncident d'une campagne de NIST à l'autre en termes de caractéristiques acoustiques. Mais la mesure de proximité entre paramètres empiriques et théoriques peut convaincre -ou dissuader- de mener une expérimentation basée sur la normalisation EFR.

Nous avons analysé la capacité de la transformation EFR à mettre en conformité les données d'évaluation à celles de développement. Cette capacité se mesure par la coïncidence des paramètres de tendance centrale de l'évaluation à ceux de l'apprentissage.

Les tables 4.10, 4.11, 4.12 indiquent :

- les normes des moyennes globales des fichiers d'évaluation (ensemble des segments cibles et tests) au fur et à mesure des itérations de l'algorithme EFR suivant le jeu d'apprentissage indiqué.

- les variances des fichiers d'évaluation après chaque itération, suivant la même configuration.

En fait, les données d'évaluation s'étendant sur la surface de l'hypersphère de rayon $\sqrt{p}$, ces deux paramètres sont reliés par la relation $\sigma^2(\mathcal{X}) = p - \|\mu_i\|^2$.

Pour chacun des trois jeux de données, la variance des données d'évaluation tend vers sa valeur maximale. Le biais tient à une légère déviation de la moyenne, de l'ordre de 15%, entre l'origine 0 et le rayon $\sqrt{p}$. Cette déviation pouvant découler d'une constitution non aléatoire du fichier d'évaluation et son intensité étant faible, nous pouvons considérer que l'algorithme EFR a bien contribué à conditionner les données d'évaluation par rapport à celles de développement, au moins en termes de moyenne et va-

jeu LIA-hommes $p = 400$					
Après chaque itération, normes des i-vectors = $\sqrt{p} = \sqrt{400} = 20$					
Itération i	0	1	2	3	4
$\|\mu_{\text{eval},i}\|$	0.10	2.98	2.84	2.83	2.83
$\sigma(\mathcal{X}_{\text{test}})$	0.36	19.77	19.79	19.79	19.79

TABLE *4.10* – *Convergences de la moyenne globale et de la variance des données de test vers 0 et $\sqrt{p}$ (LIA).*

jeu BUT-hommes $p = 600$					
Après chaque itération, normes des i-vectors = $\sqrt{p} = \sqrt{600} = 24.49$					
Itération i	0	1	2	3	4
$\|\mu_{\text{eval},i}\|$	11.20	3.75	3.66	3.65	3.65
$\sigma(\mathcal{X}_{\text{test}})$	18.25	24.20	24.21	24.22	24.22

TABLE *4.11* – *Convergences de la moyenne globale et de la variance des données de test vers 0 et $\sqrt{p}$ (BUT-hommes).*

jeu BUT-femmes $p = 600$					
Après chaque itération, normes des i-vectors = $\sqrt{p} = \sqrt{600} = 24.49$					
Itération i	0	1	2	3	4
$\|\mu_{\text{eval},i}\|$	11.94	3.83	3.77	3.76	3.76
$\sigma(\mathcal{X}_{\text{test}})$	18.72	24.19	24.20	24.20	24.20

TABLE *4.12* – *Convergences de la moyenne globale et de la variance des données de test vers 0 et $\sqrt{p}$ (BUT-femmes).*

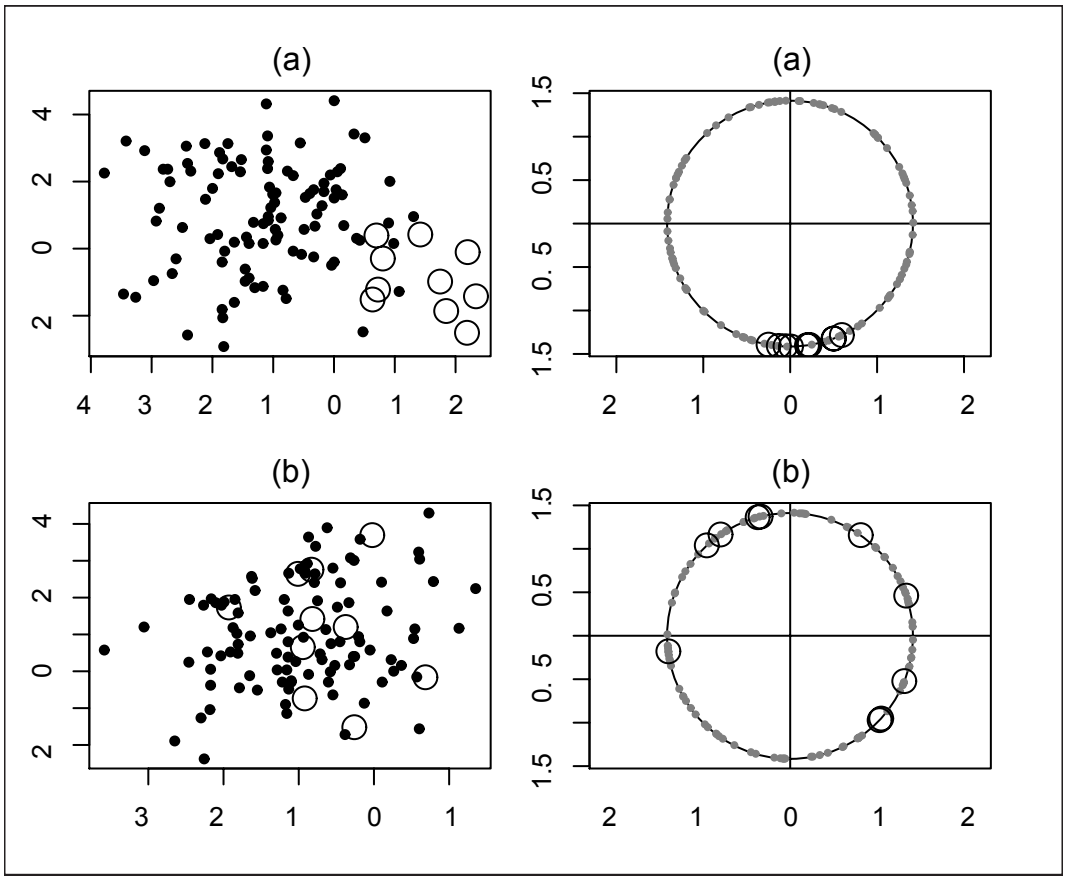

FIGURE 4.10 – *Deux exemples en 2D de limite de la mise à conformité des données d'évaluation opérée par l'algorithme EFR. Dans les deux cas (a) et (b), le biais initial entre données de développement (points noirs) et d'évaluation (points blancs) n'est pas redressé par la transformation.*

riance. L'adéquation de l'évaluation au développement en termes de gaussianité sera étudiée dans la suite.

La figure 4.10 montre deux exemples, en dimension 2, de mise à conformité des données d'évaluation opérée par l'algorithme EFR. Dans le cas (a), les données d'évaluation (points blancs) présentent un biais initial important par rapport au développement, en terme de moyenne et covariance. Ces données d'évaluation se retrouvent (graphique à droite) dans une région précise de la surface sphérique. Dans le cas (b), les données d'évaluation suivent la même loi que le développement, mais la quantité réduite d'observations comparée aux données de développement induit quand même un biais final, sur la moyenne et la covariance (la distribution des données d'évaluation après sphérisation n'est pas centrée à l'origine). Ceci rappelle que l'évaluation ne peut être totalement "redressée" pour se conformer aux paramètres issus du développement.

4.4 Opportunité de l'algorithme EFR après LDA

Après projection des i-vectors par LDA, ceux-ci ne s'étendent plus sur la surface sphérique initiale. La projection ne pouvant augmenter la norme, les données de norme initiale $\sqrt{p}$ sont envoyées sur l'intérieur d'une boule de dimension r, où r est le rang de projection. Nous considérons qu'une nouvelle application de l'algorithme EFR à ces données est alors nécessaire, pour satisfaire les objectifs de gaussianité.

4.5 Performances comparées des scorings après normalisation EFR

Nous présentons ici une comparaison de systèmes basés sur les i-vectors utilisant différents modèles et scorings, avec ou sans normalisation EFR. Trois modèles sont testés :

- le modèle LDA-WCCN-cosine scoring, qui enchaîne une réduction de dimensionnalité par LDA et le scoring par WCCN-cosine. Ce dernier inclut implicitement une transformation des données (normalisation WCCN puis division par la norme)
- Le modèle two-covariance, cas particulier déterministe de la PLDA. Celui-ci est testé sans puis avec normalisation EFR, pour estimer le gain de cette transformation en terme de performance. La LDA peut être pratiquée à partir des matrices $(\mathbf{B}, \mathbf{W})$ ou bien $(S_{\mathbf{B}}, S_{\mathbf{W}})$, définies en 2.41, 2.42 et 2.44, 2.45.
- Le modèle de Mahalanobis, cas particulier du modèle two-covariance, où aucune hypothèse n'est émise sur la variabilité locuteur ($\mathbf{B}^{-1} = 0$).

La comparaison de ces résultats va permettre d'estimer l'impact de la normalisation EFR, mais aussi l'efficacité d'une modélisation de la variabilité locuteur et de déterminer les paramètres de la LDA les plus pertinents ($\mathbf{B}, \mathbf{W}$ ou $S_{\mathbf{B}}, S_{\mathbf{W}}$).

Afin d'affiner l'analyse de la transformation, des systèmes décomposant cette procédure ont été mis en place : une simple normalisation de longueur sur des vecteurs centrés ("Lnorm") ou 1 puis 2 itérations d'EFR.

Ces expériences utilisent les jeux de i-vectors BUT-hommes et BUT-femmes et les conditions d'évaluation hommes et femmes NIST-SRE-2010-det5Extended (détaillés en annexe A). La table 4.13 affiche les résultats, en terme d'EER et DCF min, de systèmes basés sur différentes options et alternatives : la normalisation peut ne pas être effectuée (**non**) hormis le centrage suivant la moyenne du fichier d'apprentissage, ou bien l'être implicitement par le modèle et scoring **LDA-WCCN-cosine**, par seule normalisation de longueur des vecteurs centrés (**Lnorm**) ou par EFR, avec 1 ou 2 itérations (**EFR 1 iter.**, **EFR 2 iter.**).

Une LDA peut être pratiquée suivant les matrices $\mathbf{B}, \mathbf{W}$ ($\mathbf{LDA}_{\mathbf{B},\mathbf{W}}$) ou bien $\mathbf{S_B}, S_{\mathbf{W}}$ ($\mathbf{LDA}_{S_{\mathbf{B}},S_{\mathbf{W}}}$). Le modèle utilisé peut être celui de **Mahalanobis** ou bien two-covariance (**2Cov**).

			hommes		femmes	
	normalisation	scoring	EER	DCF	EER	DCF
1	non	2Cov	4.47	0.44	4.75	0.48
2	non	LDA$_{\mathbf{B,W}}$-2Cov	2.45	0.41	3.42	0.41
3	LDA$_{\mathbf{B,W}}$-WCCN-cosine		1.90	0.41	2.83	0.39
4	EFR 2 iter.	LDA$_{\mathbf{B,W}}$-Mahalanobis	1.47	0.33	2.24	0.32
5	Lnorm	2Cov	1.38	0.34	2.34	0.44
6	Lnorm	LDA$_{S_{\mathbf{B}},S_{\mathbf{W}}}$-2Cov	1.27	0.31	2.27	0.38
7	EFR 1 iter.	LDA$_{S_{\mathbf{B}},S_{\mathbf{W}}}$-2Cov	1.36	0.33	2.29	0.39
8	EFR 1 iter.	LDA$_{\mathbf{B,W}}$-2Cov	1.36	0.30	1.89	0.35
9	EFR 2 iter.	LDA$_{S_{\mathbf{B}},S_{\mathbf{W}}}$-2Cov	1.30	0.32	2.30	0.39
10	EFR 2 iter.	LDA$_{\mathbf{B,W}}$-2Cov	1.27	0.31	1.89	0.35

TABLE *4.13* – *Performances en termes d'EER et DCF min de différents systèmes pour la condition det 5 extended de NIST-SRE 2010 par genre.*

La comparaison des deux premiers systèmes (sans normalisation et basés sur le modèle two-covariance, mais avec ou sans LDA) montre l'opportunité d'une réduction de dimensionnalité par LDA. Le rang r optimal de la LDA est de 80 (pour une dimension initiale de 600). Ce ce rang s'est avéré optimal pour toutes les expériences utilisant la LDA. Cette technique procure une amélioration relative d'EER de l'ordre de 36% et de DCF minimale de l'ordre de 10%.

Les scorings de Mahalanobis et two-covariance, basés sur le log-ratio d'hypothèses complémentaires, prévalent sur le scoring cosinus de WCCN-cosine (pour les expériences hommes, par exemple, l'EER passe de 1.90% à 1.47% et 1.27%). Le gain procuré par le modèle two-covariance sur celui de Mahalanobis montre l'opportunité d'une modélisation de la variabilité inter-locuteur.

L'utilisation des matrices $\mathbf{B}$ et $\mathbf{W}$ durant la LDA, plutôt que $\mathbf{S_B}$ et $S_{\mathbf{W}}$, améliore nettement les performances pour les expériences femmes (1.89% au lieu de 2.27% en terme d'EER, 0.35 au lieu de 0.38 en terme de DCF). Si le système 6 procéde sans standardisation et sur les matrices $\mathbf{S_B}$ et $S_{\mathbf{W}}$, il rejoint pourtant les performances du système 10 pour l'évaluation hommes, alors que ses performances pour l'évaluation femmes sont faibles. Le système 10 s'avére donc plus robuste.

La comparaison des quatre dernières lignes montre une légère amélioration à la deuxième itération d'EFR. Précisons qu'aucun des résultats n'a varié au delà de deux itérations.

L'objectif principal de ces expériences était d'étudier l'impact des techniques de normalisation sur la qualité des systèmes par i-vectors. La comparaison du second système avec les suivants montre clairement leur rôle dans la qualité de la détection. Le second système applique le meilleur des modèles mis en compétition dans cette évaluation (LDA sur $\mathbf{B}$,$\mathbf{W}$ et modèle two-covariance) mais sans normalisation, hormis un centrage des données. Il génère pourtant des taux d'erreur et coûts de décision plus élevés que l'ensemble des systèmes appliquant une normalisation. Le gain relatif d'EER entre le

système le plus performant (dernière ligne) et ce système est de l'ordre de 46% en terme d'EER et de 19% en terme de DCF. Ces valeurs quantifient et confirment l'impact significatif de la technique EFR dans la performance d'un système de vérification du locuteur basé sur les i-vectors. Les procédures itératives de standardisation et unitarisation, à partir des paramètres de distribution d'un fichier d'apprentissage, constituent un moyen efficace de préparer les i-vectors à la phase de modélisation.

La table 4.13 montre les résultats de systèmes basés sur un maximum de deux itérations de l'algorithme EFR. Au delà , la convergence vers le modèle standard unitaire n'influe plus sur la performance, la transformation tendant à être stationnaire. Le nombre d'itérations de l'algorithme EFR varie suivant les configurations et suivant les contextes de reconnaissance (il atteint 5 dans (Rouvier and Meignier, 2012a) (Rouvier and Meignier, 2012b)).

Une famille de transformations est envisageable dans le champ des i-vectors, combinant les phases de standardisation et unitarisation. En effet, la première de ces phases, basée dans EFR sur la variabilité totale, peut être aussi effectuée suivant une variabilité ciblée. L'objectif est alors de combattre des défauts de modélisation tout en accentuant le caractère discriminant de la représentation. Ces considérations ont fait l'objet d'une étude que nous présentons dans la section suivante et qui s'est avérée concluante dans le cadre des systèmes utilisant la modélisation par PLDA gaussienne.

4.6 Adaptation de la normalisation au modèle PLDA

L'algorithme de normalisation EFR peut être appliqué aux données avant un processus décisionnel par PLDA Gaussienne. Comme indiqué précédemment, cette normalisation contribue à améliorer la gaussianité des données, qui constitue un postulat de la PLDA Gaussienne. C'est d'ailleurs par celle-ci que la méthode a rejoint les performances de la Heavy-tailed PLDA de Kenny (Garcia-Romero and Espy-Wilson, 2011). Cette dernière, qui s'appuie sur la loi de Student, tente de prendre en compte des anomalies des données empiriques par rapport à des a priori gaussiens. En particulier, l'aspect de la distribution de vraisemblance, à forte queue de probabilité, se rapproche plus d'une loi de Student que gaussienne. L'appartenance de la loi de Student à la famille des lois exponentielles a permis la mise en place d'un algorithme EM itératif comparable à l'état de l'art gaussien, mais avec une augmentation de complexité des calculs : une borne inférieure variationnelle est utilisée comme approximation pour chacune des vraisemblances marginales impliquées dans le log-ratio de vraisemblance (Kenny, 2010). Le but d'une normalisation dans le cadre d'une technique de décomposition de vecteurs par EM est seulement intermédiaire : il s'agit de rendre plus optimal l'algorithme itératif par maximisation de borne inférieure de la vraisemblance.

Les postulats gaussiens sont connus pour aboutir à des modèles peu complexes. La technique de normalisation autorise donc de conserver un cadre d'a priori gaussien à la PLDA, évitant l'accroissement de complexité des modélisations basées sur d'autres lois. Mais le caractère probabiliste de la PLDA, en particulier l'estimation par maximum de

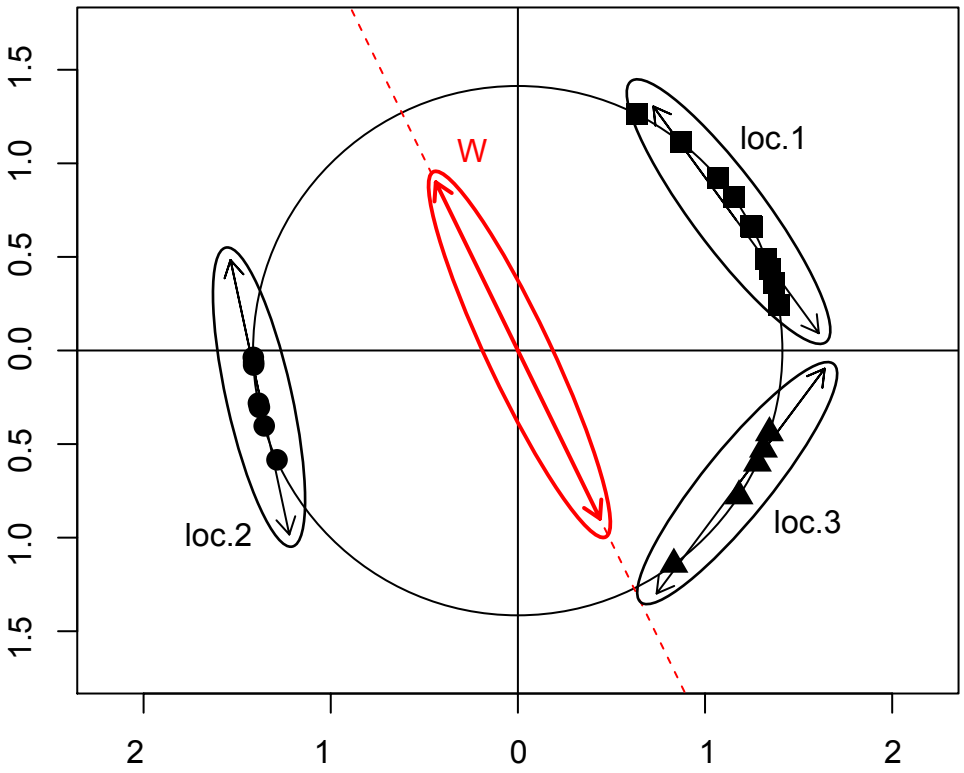

FIGURE 4.11 – *Un exemple en 2D de variabilité session sur une surface sphérique. Les i-vectors de trois locuteurs sont affichés. Les fléches indiquent le 1er axe principal session de chacun. L'axe rouge indique le 1er axe principal intra-locuteur (***W***).*

vraisemblance de métaparamètres matriciels de variabilité, nous a amené à envisager une transformation des données plus adaptée qu'EFR à cette modélisation.

4.6.1 Investigation préliminaire

Nous avons remarqué le fait suivant : une fois que des i-vectors s'étendent sur une surface finie et sphérique, comme c'est le cas après normalisation de longueur, il est difficile d'estimer une matrice de covariance intra-locuteur. En effet, cette estimation linéaire s'adapte mal à la nature non-linéaire de la surface.

La figure 4.11 illustre cet état de fait en deux dimensions : un groupe d'i-vectors de trois locuteurs s'étend sur un cercle, avec un bon niveau de séparabilité entre les classes-locuteur. Leurs ellipses de covariance ont été tracées sur la figure. La matrice **W** de covariance intra-locuteur a été ensuite calculée, qui, rappelons-le, est obtenue par la moyenne des matrices covariance de chacune des classes. L'ellipse de covariance correspondante est indiquée sur la figure. La comparaison de cette ellipse aux ellipses

par classe montre clairement la médiocrité du résultat, qui ne reflète pas la réalité des distributions intra-classes en terme d'orientation. Malgré le bon niveau de séparabilité des classes, la vérification du locuteur sera pénalisée par une estimation linéaire de la variabilité intra-locuteur.

Remarquons qu'un tel défaut d'estimation ne se produirait pas si les classes locuteurs se trouvaient toutes dans la même zone de dispersion réduite, sur la surface sphérique. Mais l'algorithme de normalisation garantit une dispersion maximale des vecteur, interdisant qu'une telle situation se produise.

La PLDA gaussienne modélise les variabilités locuteur et résiduelle sous forme linéaire (matrices $\mathbf{\Gamma}$, $\mathbf{\Phi}$ et $\mathbf{\Sigma}$ du paragraphe 2.6.5.1). L'imprécision de la matrice $\mathbf{W}$ sur la figure 4.11 montre la difficulté à cerner la part résiduelle dans des métaparamètres matriciels $\mathbf{\Phi}$ et $\mathbf{\Sigma}$. La voie qui nous a paru la plus appropriée consiste à envoyer les données sur une surface sphérique par une normalisation de longueur (afin d'améliorer leur gaussianité), mais sans privilégier aucune direction principale de variabilité intra-locuteur, ni aucune dépendance entre ces directions. Par conséquent, la matrice de covariance intra-locuteur après transformation doit être sphérique. Une matrice sphérique est une matrice du type $\sigma^2\mathbf{I}$, où σ est un scalaire et $\mathbf{I}$ la matrice identité.

Le modèle obtenu est celui de la probabilistic PCA (PPCA) ou modèle isotropique (Tipping and Bishop, 1999a). Il constitue un cas particulier de la Factor analysis et peut donc être obtenu par algorithme de type EM-ML. Mais l'efficacité des techniques de normalisation précédentes nous a amené à en envisager une variante, adaptée à la contrainte d'isotropie. Trouver une solution qui concilie normalisation des longueurs et matrice résiduelle sphérique est grandement facilité par l'algorithme de normalisation EFR précédent. Se basant sur le fait que cette normalisation fait tendre la matrice de covariance totale vers l'identité, qui est sphérique, nous proposons une transformation similaire, préparatoire à une PLDA Gaussienne, qui remplace matrice de covariance **totale** par matrice de covariance **intra-locuteur**. Nous décrivons ci-dessous l'algorithme proposé puis détaillons ces propriétés.

4.6.2 Normalisation spécifique à la PLDA Gaussienne : Spherical Nuisance (LW)

Etant donné un jeu de données d'apprentissage $\mathcal{T}$ et notant p la dimension des i-vectors, l'algorithme de transformation des i-vectors procède en deux phases : apprentissage et test :

Phase d'apprentissage

Pour $i = 1$ à $nb\_iterations$

Calculer la moyenne μ_i et la matrice de covariance intra-locuteur $\mathbf{W}_i$ de $\mathcal{T}$

Pour chaque w de $\mathcal{T}$: $\quad w \leftarrow \dfrac{\mathbf{W}_i^{-\frac{1}{2}}(w-\mu_i)}{\left\|\mathbf{W}_i^{-\frac{1}{2}}(w-\mu_i)\right\|}$

Pour chaque w de $\mathcal{T}$: $\quad w \leftarrow \sqrt{p}w$

A chaque itération, les i-vectors du jeu de données d'apprentissage sont successivement standardisés suivant $\mathbf{W}$ puis normalisés en les divisant par leur norme euclidienne.

La dernière opération (multiplication des i-vectors par le scalaire $\sqrt{p}$) est sans effet sur le processus décisionnel de discrimination du locuteur. Elle est destinée à clarifier les justifications théoriques qui suivront. Les paramètres statistiques successifs de moyennes $\{\mu_i\}$ et covariance $\{\mathbf{W}_i\}$ sont enregistrés pour la phase de test.

Phase de test

Etant donné un i-vector d'évaluation w_{test},

Pour $i = 1$ à $nb\_iterations$: $\quad w_{\text{test}} \leftarrow \dfrac{\mathbf{W}_i^{-\frac{1}{2}}(w_{test} - \mu_i)}{\left\|\mathbf{W}_i^{-\frac{1}{2}}(w_{test} - \mu_i)\right\|}$

$w_{\text{test}} \leftarrow \sqrt{p} w_{\text{test}}$

A partir des paramètres μ_i et $\mathbf{W}_i$ calculés durant la phase d'apprentissage, un i-vector de test w_{test} est modifié par application itérative de la transformation.

Nous appelons "*Spherical Nuisance*" (et en abrégé *SphN*) cette technique de normalisation, l'ensemble des variabilités considérées comme nuisibles présentant alors une matrice de covariance a priori sphérique. Elle est aussi notée dans certaines publications L$\mathbf{W}$ [3].

Il reste cependant à montrer la convergence des données vers un tel modèle.

4.6.3 Convergence vers le modèle isotropique

De manière analogue à l'algorithme EFR, nous devons vérifier expérimentalement la propriété suivante :

Propriété : lorsque $i \to \infty$,

- la moyenne globale μ_i tend vers 0,

- la matrice de covariance intra-locuteur $\mathbf{W}_i$ tend vers une matrice isotropique $\sigma^2\mathbf{I}$ ($\sigma \in \mathbb{R}$),

les deux convergences étant strictement monotones (la distance diminue strictement à chaque itération).

$\Diamond$ i) convergence de la moyenne μ_i

Les tables 4.14, 4.15 et 4.16 montrent les normes des moyennes globales des trois jeux de données, initialement ("itération 0") puis après chaque itération jusqu'à la quatrième, de l'algorithme SphN. La convergence stricte des moyennes vers 0 est confirmée empiriquement sur ces jeux de données.

3. pour "Length-normalization" et standardisation suivant la matrice de covariance intra-locuteur $\mathbf{W}$.

jeu LIA-hommes $p = 400$					
Après chaque itération, normes des i-vectors = $\sqrt{p} = \sqrt{400} = 20$					
Itération i	0	1	2	3	4
$\|\mu_i\|$	0.029	0.20	0.020	0.0026	0.00037

TABLE *4.14* – *Convergences de la moyenne globale vers 0 (LIA).*

jeu BUT-hommes $p = 600$					
Après chaque itération, normes des i-vectors = $\sqrt{p} = \sqrt{600} = 24.49$					
Itération i	0	1	2	3	4
$\|\mu_i\|$	10.95	0.13	0.015	0.0026	0.00057

TABLE *4.15* – *Convergences de la moyenne globale vers 0 (BUT-hommes).*

$\Diamond$ ii) Convergence de la matrice de covariance $\mathbf{W}_i$ vers une matrice isotropique $\sigma^2\mathbf{I}$ ($\sigma \in I\!R$) :

Pour montrer cette convergence, nous devons estimer, initialement puis après chaque itération de l'algorithme de normalisation SphN, la distance entre la matrice de covariance intra-locuteur $\mathbf{W}_i$ et la droite vectorielle des matrices isotropiques, notée $vect\{\mathbf{I}\}$. Cette distance est l'erreur au sens des moindres carrés entre $\mathbf{W}_i$ et son projeté sur $vect\{\mathbf{I}\}$:

$$\left\|\mathbf{W}_i - p_{vect\{\mathbf{I}\}}(\mathbf{W}_i)\right\| \tag{4.21}$$

où $p_{vect\{\mathbf{I}\}}(\mathbf{W}_i)$ est le projeté orthogonal de $\mathbf{W}_i$ sur $vect\{\mathbf{I}\}$ et la norme est celle, par exemple, de la trace. Le coefficient de détermination R^2 est alors donné par :

$$R^2 = \frac{\left\|\mathbf{W}_i - p_{vect\{\mathbf{I}\}}(\mathbf{W}_i)\right\|^2}{\|\mathbf{W}_i\|^2} \tag{4.22}$$

qui, de par l'orthogonalité entre l'erreur et le projeté, est compris entre 0 et 1. La valeur nulle du R^2 correspond au cas où $\mathbf{W}_i$ est exactement isotropique.

Au sens de la norme et du produit scalaire de la trace de Frobenius, R^2 s'écrit :

jeu BUT-femmes $p = 600$					
Après chaque itération, normes des i-vectors = $\sqrt{p} = \sqrt{600} = 24.49$					
Itération i	0	1	2	3	4
$\|\mu_i\|$	11.58	0.15	0.016	0.0031	0.00089

TABLE *4.16* – *Convergences de la moyenne globale vers 0 (BUT-femmes).*

jeu LIA-hommes $p = 400$					
Itération i	0	1	2	3	4
R^2	0.31	0.0079	$9.348e-05$	$1.75e-06$	$3.77e-08$

TABLE 4.17 – *Convergences de* $\mathbf{W}$ *vers l'identité (LIA).*

jeu BUT-hommes $p = 600$					
Itération i	0	1	2	3	4
R^2	0.20	0.0067	0.00015	$8.78e-06$	$7.19e-07$

TABLE 4.18 – *Convergences de* $\mathbf{W}$ *vers l'identité (BUT-hommes).*

$$R^2 = \frac{\left\| \mathbf{W}_i - \frac{\mathbf{W}.\mathbf{I}}{\|\mathbf{I}\|}\frac{\mathbf{I}}{\|\mathbf{I}\|} \right\|^2}{\|\mathbf{W}_i\|^2} = \frac{Tr\left[\left(\mathbf{W}_i - \frac{Tr(\mathbf{W})}{p}\mathbf{I}\right)^2\right]}{Tr\left[(\mathbf{W}_i)^2\right]} \tag{4.23}$$

Rem. : le projeté de $\mathbf{W}_i$ est nécessairement de la forme $\sigma^2\mathbf{I}$ (i.e. un scalaire positif), la matrice $\mathbf{W}_i$ étant positive.

Les valeurs R^2 sont indiquées sur les tables 4.17, 4.18 et 4.19 pour les trois mêmes jeux que précédemment. La convergence stricte des matrices $\mathbf{W}_i$ vers des configurations matricielles isotropiques est clairement constatée sur l'ensemble de ces tables.

4.6.4 Généralisation des métaparamètres

Nous mesurons ici, de la même manière que pour la normalisation EFR (paragraphe 4.3.7), la coïncidence des paramètres de tendance centrale de l'évaluation à ceux de l'apprentissage après normalisation SphN.

Les tables 4.20, 4.21 et 4.22 indiquent les normes de la moyenne de l'évaluation, pour les mêmes jeux de développement que précédemment ainsi que pour les jeux d'évaluation correspondants. Comme après l'algorithme EFR, la moyenne globale des données d'évaluation tend vers une valeur non nulle. Cette déviation à l'origine (4.51 après 4 itérations, de l'ordre de 18% du rayon $\sqrt{p}$) reste minime. L'algorithme SphN a donc partiellement conditionné les données d'évaluation à celles de développement, en terme de moyenne.

jeu BUT-femmes $p = 600$					
Itération i	0	1	2	3	4
R^2	0.20	0.0056	0.00022	$2.41e-05$	$3.25e-06$

TABLE 4.19 – *Convergences de* $\mathbf{W}$ *vers l'identité (BUT-femmes).*

jeu LIA-hommes $p = 400$					
Après chaque itération, normes des i-vectors = $\sqrt{p} = \sqrt{400} = 20$					
Itération i	0	1	2	3	4
$\|\mu_{\text{eval},i}\|$	0.10	3.90	3.72	3.70	3.70

TABLE 4.20 – *Convergences de la moyenne des tests vers 0 (LIA).*

jeu BUT-hommes $p = 600$					
Après chaque itération, normes des i-vectors = $\sqrt{p} = \sqrt{600} = 24.49$					
Itération i	0	1	2	3	4
$\|\mu_{\text{eval},i}\|$	11.20	4.31	4.21	4.20	4.20

TABLE 4.21 – *Convergences de la moyenne des tests vers 0 (BUT-hommes).*

4.6.5 Stationnarité

Puisque moyenne et matrice de covariance intra-locuteur **W** des données tendent de manière strictement monotone vers 0 et $\sigma^2\mathbf{I}$, la transformation appliquée à chaque itération de l'algorithme tend vers l'application identique. Ainsi, l'algorithme devient stationnaire au fil des itérations.

4.6.6 Graphe spectral

Conjointement au graphe spectral des données d'origine déjà présenté sur la figure 3.3, la figure 4.12 affiche le graphe spectral du jeu de données BUT-hommes, dans la base des vecteurs propres de la matrice de covariance inter-locuteur **B**, après trois itérations de l'algorithme. Sur la figure, le spectre de **W** est presque exactement plat. Le point à retenir ici est que l'énergie du spectre de **B** n'a pas été diminuée par l'algorithme : c'est à dire que la pente de ses valeurs propres s'est maintenue, moins de 200 axes sur 600 initiaux contenant une part majoritaire de variabilité locuteur. La part de variance locuteur dans la variance totale s'est même accrue (de 40% à 47% pour le jeu BUT-femmes et de de 41% à 50% pour le jeu BUT-hommes).

La PLDA recherche un sous-espace locuteur (*eigenvoices*) de plus grande vraisemblance, avec la contrainte d'un résidu à distribution aléatoire. Il est évident sur cette figure que les premiers axes (1 à 100 pour le jeu de données testé) forment une base de départ pertinente pour ce sous-espace. La séparation d'un i-vector en une compo-

jeu BUT-femmes $p = 600$					
Après chaque itération, normes des i-vectors = $\sqrt{p} = \sqrt{600} = 24.49$					
Itération i	0	1	2	3	4
$\|\mu_{\text{eval},i}\|$	11.94	4.60	4.52	4.51	4.51

TABLE 4.22 – *Convergences de la moyenne des tests vers 0 (BUT-femmes).*

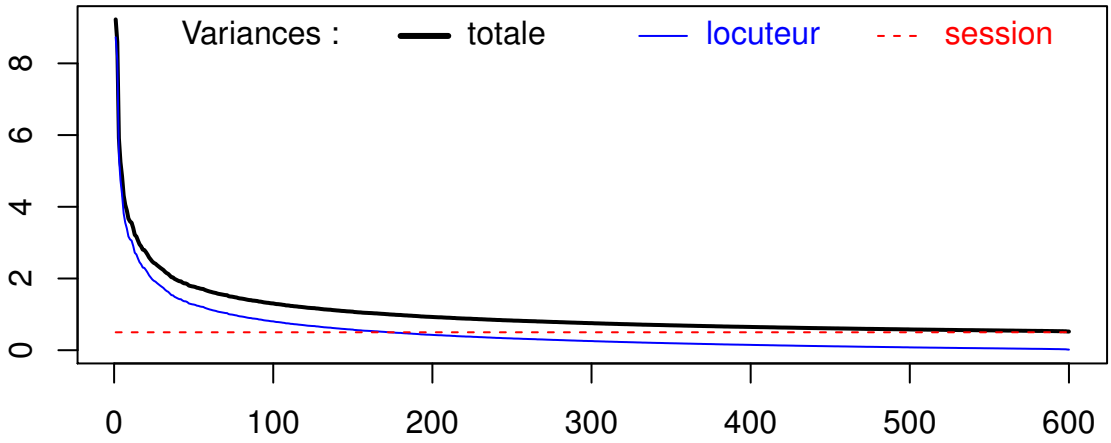

FIGURE 4.12 – *Graphe spectral-***B** *du jeu de données BUT-hommes après normalisation Spherical Nuisance.*

sante de ce sous-espace et en un résidu de plein rang, dont la variabilité présente les caractéristiques isotropiques d'un bruit aléatoire, constitue donc une initialisation déterministe d'excellente qualité à une modélisation PLDA basée sur une vraisemblance probabiliste. Ce fait est confirmé dans le paragraphe suivant.

4.6.7 Initialisation de la PLDA

Les métaparamètres de la PLDA sont estimés par un algorithme itératif "Expectation Maximization" (EM). En l'absence de toute autre information, les matrices locuteur et session $\mathbf{\Phi}$ et $\mathbf{\Gamma}$ de l'équation 2.74 sont initialisées aléatoirement après l'algorithme précédent. Cependant, notre nouvelle représentation peut être avantageusement utilisée pour initialiser les matrices de la PLDA. Pour procéder, les i-vectors d'apprentissage et toute observation d'une évaluation sont exprimés dans la base de la matrice de covariance inter-locuteur $\mathbf{B}$: chaque vecteur est multiplié par la matrice des vecteurs propres de $\mathbf{B}$ dans l'ordre décroissant des valeurs propres. Dans cette base, qui est celle de la figure 4.12, les r premières dimensions contiennent la variabilité locuteur "principale". Aussi, nous initialisons la matrice locuteur par :

$$\mathbf{\Phi} = \begin{bmatrix} \mathbf{I}_{r\times r} \\ \mathbf{0}_{(p-r)\times r} \end{bmatrix} \tag{4.24}$$

où $\mathbf{I}_{r\times r}$ est la matrice $r \times r$ identité et $\mathbf{0}_{(p-r)\times r}$ est la matrice $(p-r) \times r$ nulle.

La matrice $\mathbf{\Gamma}$ est initialisée par la décomposition de Cholesky de $\mathbf{W}$, en sorte que :

$$\mathbf{W} = \mathbf{\Gamma}\mathbf{\Gamma}^t \tag{4.25}$$

La PLDA sera alors exécutée avec des rangs locuteur r et session plein p. De cette manière, l'algorithme EM est initialisé à un point significatif du champ d'optimisation de la PLDA. En effet :

- la base canonique à partir de laquelle débute l'EM est une version triée des axes orthonormés principaux du locuteur. Les r premières dimensions forment un sous-espace contenant la plus forte part de variabilité locuteur qu'il soit possible d'obtenir d'une base orthogonale, à partir des données d'apprentissage. Nous supposons que le meilleur sous-espace d'*eigenvoices* sous contrainte gaussienne est proche de ce sous-espace,
- l'information résiduelle est variée suivant une loi isotropique.

La PLDA va raffiner ce modèle initial, en distribuant la variabilité locuteur dans son sous-espace et en normalisant les nuisances anisotropiques.

4.7 Propriétés après normalisation "Spherical Nuisance"

Comme l'algorithme EFR, l'algorithme de normalisation "Spherical Nuisance" entraîne les données d'apprentissage vers un modèle doté d'un certain nombre de propriétés. Ces propriétés sont similaires à celles de l'EFR. Etant donné un jeu $\mathcal{X} = \{w_i\}_{i=1,...n}$ de n vecteurs de dimension p, vérifiant les propriétés suivantes :

- la moyenne μ de $\mathcal{X}$ est égale à 0

- la matrice de covariance intra-classes $\mathbf{W}$ de $\mathcal{X}$ est la matrice $p \times p$ identité I.

- $\mathcal{X}$ s'étend sur l'hypersphère de rayon $\sqrt{p}$ (tous les vecteurs de $\|w\|$ sont de norme $\sqrt{p}$).

4.7.1 Dispersion maximale

Les vecteurs après transformation sont unitaires et leur moyenne est égale à 0. La variance, indicateur de dispersion de $\mathcal{X}$, est donc à nouveau maximale sur la surface de l'hypersphère (le raisonnement est le même qu'au paragraphe 4.3.2).

4.7.2 Correspondance entre les décompositions en valeurs singulières. Base orthonormée de facteurs propres

Comme dans le cas d'EFR, il existe une base orthonormée de "facteurs propres". Pour tout vecteur propre v de $\mathbf{B}$ associé à la valeur propre λ, on a :

$$\mathbf{\Sigma}v = \mathbf{B}v + \mathbf{W}v \tag{4.26}$$

et, comme $\mathbf{W} = \mathbf{I}$,

$$\mathbf{\Sigma} v = (1 + \lambda)\, v \tag{4.27}$$

$\mathbf{B}$ et $\mathbf{\Sigma}$ possèdent strictement (i.e. dans l'ordre décroissant des valeurs propres) la même base de vecteurs propres. Après transformation SphN et passage dans cette base, les premières dimensions constituent une expression déterministe des *eigenvoices* et les dernières des *eigenchannels*. Avant même la détermination par EM-ML des métaparamètres les plus vraisemblables, les données correspondent partiellement à la situation escomptée par la PLDA. Seul le caractère gaussien des différentes distributions n'est pas a priori assuré. Nous verrons dans la partie expérimentale qu'il est en grande partie acquis, à partir du moment où l'initialisation non-aléatoire du paragraphe précédent, basée sur les facteurs propres, a été réalisée.

4.7.3 LDA

Le critère de Rayleigh après SphN s'écrit

$$\frac{v^t \mathbf{B} v}{v^t \mathbf{W} v} = \frac{v^t \mathbf{B} v}{v^t v} = \frac{v^t \mathbf{B} v}{\|v\|^2} = \frac{v^t \mathbf{B} v}{1} = v^t \mathbf{B} v \tag{4.28}$$

qui est maximal pour v vecteur propre de $\mathbf{B}$ associé à sa plus grande valeur propre. On retrouve donc le fait que le meilleur sous-espace de projection au sens de l'analyse discriminante linéaire est le sous-espace principal de $\mathbf{B}$, c'est à dire celui des *eigenvoices*. Le critère de Rayleigh est optimisé après cet algorithme. L'application du modèle LDA-two covariance aux vecteurs, mais aussi d'un modèle PLDA succédant à une réduction de dimensionnalité par LDA, est donc très pertinent sur les données transformées par la normalisation SphN.

4.7.4 Analyse des métaparamètres après initialisation

Les métaparamètres $\mathbf{\Phi}$ et Γ de la PLDA Gaussienne sont initialisés de manière déterministe suivant la méthode proposée au paragraphe 4.6.7, puis affinés suivant l'algorithme par maximum de vraisemblance de la PLDA. Nous nous intéressons ici à la part attribuable à chacune des deux étapes dans la détermination de ces paramètres.

La matrice $\mathbf{\Phi}$ a été initialisée selon l'équation 4.24 par un bloc identité et un bloc nul et $\mathbf{\Gamma}$ par la matrice triangulaire issue de la décomposition de Cholesky de $\mathbf{W}$ dans la base de $\mathbf{B}$ (équations 4.24 et 4.25). La PLDA raffine alors ce modèle initial en distribuant la variabilité locuteur dans son sous-espace et en normalisant les nuisances anisotropiques, suivant des contraintes de normalité des facteurs. Plus l'initialisation sera pertinente, plus la part dévolue à l'algorithme par maximum de vraisemblance sera réduite. Ceci se traduit par le fait que le sous-espace locuteur initial doit être approximativement conservé après EM. Cet algorithme se contentera alors d'ajuster les différences de variance entre les vecteurs de $\mathbf{\Phi}$. La base d'*eigenvoices* produite par la

	hommes	femmes
Orthogonalité de $\mathbf{\Phi}$	0.985	0.9799
Orthogonalité de $\mathbf{\Gamma}$	0.993	0.9949

TABLE 4.23 – *Orthogonalité des métaparamètres matriciels Φ et Γ de la PLDA (BUT hommes et femmes) après initialisation non aléatoire.*

PLDA sera alors proche de l'orthogonalité, mais non normée. De même, l'indépendance statistique des résidus doit entraîner la quasi-orthogonalité des *eigenchannels* contenus dans les colonnes de Γ et une légère variabilité des normes liée aux anomalies anisotropiques. Ces éléments peuvent être mesurés par l'analyse des matrices $\mathbf{\Phi}$, $\mathbf{\Gamma}$ après PLDA Gaussienne. La table 4.23 présente un diagnostic de ces matrices après l'algorithme EM de la PLDA, effectué à partir de l'expérimentation précédente sur les jeux d'apprentissage et d'évaluation BUT femmes et hommes. Le rang optimal est de $r = 80$ sur un rang total de 600, la matrice $\mathbf{\Phi}$ est donc de dimension 600×80 et la matrice $\mathbf{\Gamma}$ est de dimension 600×600.

Le degré d'orthogonalité des *eigenvoices* (les vecteurs-colonnes de $\mathbf{\Phi}$) est calculé à partir de la matrice de Gram des produits scalaires entre ces vecteurs et de la distance entre cette matrice et sa version diagonale. La valeur dans $[0, 1]$ est maximale si l'orthogonalité entre les vecteurs est totale. Ces valeurs sont :

$$\frac{Tr\left[diag\left(\mathbf{\Phi}^t\mathbf{\Phi}\right)^2\right]}{Tr\left[\left(\mathbf{\Phi}^t\mathbf{\Phi}\right)^2\right]} \text{ et } \frac{Tr\left[diag\left(\mathbf{\Gamma}^t\mathbf{\Gamma}\right)^2\right]}{Tr\left[\left(\mathbf{\Gamma}^t\mathbf{\Gamma}\right)^2\right]} \tag{4.29}$$

où $diag\,(.)$ est la matrice diagonale obtenue par annulation des valeurs non-diagonales.

La table 4.23 affiche ces mesures d'orthogonalité des vecteurs-colonnes de $\mathbf{\Phi}$ et Γ, donc des bases des sous-espaces locuteur et session après l'algorithme EM-ML de la PLDA (nous rappelons que ces valeurs immédiatement après l'initialisation sont toutes égales à 1). L'orthogonalité des matrices est quasiment exacte pour les jeux de données des deux genres. L'impact de l'algorithme par maximum de vraisemblance est donc réduit après transformation SphN et l'initialisation proposée. Ces deux dernières opérations modélisent les données d'une manière quasiment optimale, au regard des critères de la décomposition en facteurs PLDA.

Un i-vector peut donc s'écrire après PLDA :

$$w = \mu + \begin{pmatrix} \phi_1 y_1 \\ \dots \\ \phi_r y_r \\ 0 \\ \dots \\ 0 \end{pmatrix} + \mathbf{\Gamma} z + \varepsilon \tag{4.30}$$

		hommes	
normalisation	scoring	EER	DCF
non	G-PLDA	2.64 [2.62 ;2.65]	0.41 [0.40 ;0.42]
EFR (1 itér.)	G-PLDA	1.34 [1.29 ;1.38]	0.34 [0.34 ;0.34]
EFR (2 itér.)	G-PLDA	1.29 [1.24 ;1.35]	0.34 [0.34 ;0.35]
SphN	G-PLDA	1.08 [1.04 ;1.13]	0.31 [0.29 ;0.32]
SphN	G-PLDA initialisée	1.04	0.29

TABLE 4.24 – *Performances de différents systèmes basés sur la PLDA gaussienne (évaluation det 5 extended 2010 hommes).*

		femmes	
normalisation	scoring	EER	DCF
non	G-PLDA	3.23 [3.18 ;3.26]	0.38 [0.38 ;0.38]
EFR (1 itér.)	G-PLDA	1.92 [1.91 ;1.94]	0.35 [0.34 ;0.35]
EFR (2 itér.)	G-PLDA	1.94 [1.94 ;1.94]	0.35 [0.35 ;0.35]
SphN	G-PLDA	1.77 [1.73 ;1.84]	0.34 [0.33 ;0.34]
SphN	G-PLDA initialisée	1.73	0.33

TABLE 4.25 – *Performances de différents systèmes basés sur la PLDA gaussienne (évaluation det 5 extended 2010 femmes)*

où les scalaires ϕ_i représentent les écart-types par dimension du sous-espace des *eigenvoices*, les facteurs y_i, z, ε suivent une loi normale standard et où la matrice Γ est de plein rang et orthogonale.

4.8 Performance de la PLDA après normalisation SphN

Les tables 4.24 et 4.25 présentent les résultats de cinq systèmes basés sur la PLDA Gaussienne. Le protocole expérimental est le même qu'au paragraphe 4.5 précédent.

Les cinq systèmes comparés sont les suivants :

- **G-PLDA** : exécution sur les i-vectors non normalisés (hormis un centrage) d'une PLDA Gaussienne avec initialisation aléatoire des paramètres $\mathbf{\Phi}$ et $\mathbf{\Gamma}$,
- **EFR (1 itération) + G-PLDA** : 1 itération de l'algorithme EFR suivie d'une PLDA Gaussienne initialisée aléatoirement,
- **EFR (2 itérations) + G-PLDA** : 2 itérations de l'algorithme EFR suivies d'une PLDA Gaussienne initialisée aléatoirement,
- **SphN + G-PLDA** : 3 itérations de l'algorithme SphN suivies d'une PLDA Gaussienne initialisée aléatoirement,
- **SphN + G-PLDA** : 3 itérations de l'algorithme SphN suivies d'une PLDA Gaussienne initialisée suivant la méthode proposée en 4.6.7.

Chaque fois, le rang optimal de la matrice locuteur $\mathbf{\Phi}$ est $r = 80$, celui de la matrice session $\mathbf{\Gamma}$ de 600 (pleine dimension). Dans tous les systèmes, hormis le dernier, l'éva-

	LDA-WCCN- -cosine scoring	PLDA Gaussienne	SPhN + PLDA gaussienne
normalisation (1)	-	-	**W**-norm + unitarisation
Détermination sous-espace de discrimination du locuteur	LDA	PLDA **Φ**,**Γ**	PLDA **Φ**,**Γ**
normalisation (2)	WCCN-norm + unitarisation	-	-
scoring	produit scalaire	log-ratio hypo. complémentaires	log-ratio hypo. complémentaires

TABLE 4.26 – *Détails de trois méthodologies en reconnaissance du locuteur basée sur les i-vectors.*

luation hommes a nécessité 100 itérations pour atteindre la meilleure performance, celle des femmes en a nécessité 300. De plus, les initialisations étant aléatoires, les résultats sont sujets à variabilité, en terme d'EER comme de DCF min. Les valeurs affichées sont des performances moyennes obtenues sur un ensemble de 10 répétitions de la même expérimentation. Les valeurs entre crochets indiquent les minima et maxima obtenus lors des répétitions de l'expérience.

L'application de l'algorithme de normalisation EFR 1 itération améliore nettement les performances d'un système basée sur la PLDA Gaussienne. Il est à noter, comme cela l'a été dans (Garcia-Romero and Espy-Wilson, 2011), que ces performances rejoignent alors celles d'une heavy-tailed PLDA. Une expérience comparant ces systèmes sur une configuration i-vectors très proche de celle que nous avons utilisée et par une méthode avoisinante (*whitening*, puis une standardisation et normalisation de longueur) a permis de pleinement le constater.

Mais un nombre d'itérations supérieur à 1 n'apporte pas forcément la stabilisation des performances à leur optimum, tel qu'observé pour le système LDA-two–covariance. Les $4^{ème}$ et $5^{ème}$ systèmes sont basés sur la normalisation Spherical-Nuisance préliminaire à la PLDA Gaussienne. Les performances pour les deux genres marquent une amélioration, significative étant données les grandes quantités de tests effectués dans ces évaluations. Le dernier système est basé sur l'initialisation des matrices décrite précédemment. En l'absence d'aléa initial, les résultats ne sont donc plus soumis à variabilité. Seulement 10 itérations ont été nécessaires pour atteindre la performance optimale sur l'évaluation hommes, pour 2 itérations sur l'évaluation femmes.

La normalisation Spherical-Nuisance s'avère donc plus efficace et robuste qu'EFR dans le cadre de la PLDA gaussienne. La comparaison des tables 4.13, 4.24 et 4.25 montre que la méthode enchaînant la normalisation Spherical-Nuisance et un modèle de PLDA gaussienne procure les meilleures performances, en terme d'EER comme de DCFmin, quel que soit le genre. Nous reviendrons en détail dans le chapitre suivant sur la comparaison des différentes méthodes dans les système basés sur les i-vectors.

La normalisation Spherical-Nuisance inclut une "**W**-normalisation", c'est à dire une normalisation suivant la matrice de covariance intra-classes, qui s'apparente à la pro-

cédure WCCN de (Dehak et al., 2011). La table 4.26 permet d'apprécier les différentes stratégies mises en oeuvre lors des trois approches suivantes :

- LDA-WCCN-cosine scoring tel qu'initialement proposée par (Dehak et al., 2011) dans le champ des i-vectors,
- PLDA gaussienne telle que proposée initialement par (Prince and Elder, 2007),
- PLDA gaussienne avec normalisation Spherical-nuisance telle que nous l'avons élaborée. La démarche de (Dehak et al., 2011) s'est avérée pertinente par la normalisation (2) des données suivant la variabilité intra-locuteur suivie d'une unitarisation, mais pêche par un scoring insuffisant (simple produit scalaire) et par une LDA non-probabiliste et pratiquée avant normalisation. La PLDA gaussienne probabilise les étapes de détermination du sous-espace principal de discrimination et de scoring, mais ignore une phase essentielle : la normalisation des données. Parmi les techniques de normalisation par sphérisation des données, celle basée sur $\mathbf{W}$ (Spherical-nuisance) limite l'imprécision d'un paramètre linéaire global de variabilité intra-locuteur en privilégiant l'isotropie.

Un dernier point reste à signaler pour conclure cette étude : les résultats précédents peuvent interroger sur l'algorithme EM de la PLDA appliqué sur les i-vectors. La faible dimensionnalité de la représentation par i-vectors autorise, ce qui n'était pas le cas jusqu'alors dans le domaine, d'itérer l'algorithme EM-ML un nombre considérable de fois. D'une part, la convergence vers un point optimal pour le jeu de données jeu BUT-femmes semblait acquise après 100 itérations. Poussant à 300 itérations, un nouveau maximum de performance a été atteint. Mais d'autre part, et c'est le fait qui mérite pour nous le plus d'attention, après l'initialisation proposée des matrices $\mathbf{\Phi}$ et $\mathbf{\Gamma}$ la convergence vers un maximum de performance s'effectue en 10 (resp. 2) itérations pour les jeux hommes (resp. femmes), contre 100 et 300. La performance atteinte est significativement meilleure dans ce cas. Remarquons que la divergence en nombre d'itérations suivant le genre reste à expliquer.

Dans le domaine de la reconnaissance du locuteur, il est souvent admis, à juste titre, que les quantités considérables de données d'apprentissage dont on dispose autorisent à produire, par des algorithmes de type EM avec initialisation aléatoire des métaparamètres, des solutions localement maximales satisfaisantes, ce à condition d'itérer un nombre suffisant de fois. Les initialisations de métaparamètres ne sont considérées utiles qu'à accélérer la convergence vers ces maxima dans le champ d'optimisation. Certaines techniques de bruitage permettent à l'algorithme de sortir de l'attracteur en cours pour éventuellement poursuivre vers un nouveau point de meilleure maximisation ou de travailler sur une distribution de paramètres obtenus sur échantillons aléatoires (par exemple *Stochastic Expectation Algorithm SEM* (Celeux and Diebolt, 1985) (Celeux and Diebolt, 1986)). Elles sont a priori inutiles dans notre domaine, étant données les tailles d'échantillons d'apprentissage. Le fait qu'une initialisation n'accélère pas seulement la convergence, mais améliore la qualité du maximum obtenu, soulève certaines interrogations. Elle signifie qu'une part non négligeable des informations n'est pas capturée correctement par le modèle (ici par le cadre gaussien).

4.9 Bilan et conclusion

Les transformations que nous avons mises en place, sous forme de standardisations et normalisations de longueur successives, ont montré leur pertinence et leur efficacité dans le champ des i-vectors. Deux transformations dans le champ des i-vectors, appliquées dès leur extraction, ont été proposées : EFR (ou L$\mathbf{\Sigma}$) et SphN (ou L$\mathbf{W}$), la seconde étant spécifique à la PLDA. Elles visent au même objectif que les représentations par supervecteurs, décomposition en facteurs par FA-total var ou modèle génératif PLDA : optimiser la qualité de la modélisation en respectant des hypothèses et contraintes gaussiennes.

Nous montrons en effet, pour chacune de ces deux transformations, la convergence empirique des données transformées vers un modèle théorique. Ce modèle est chaque fois doté d'un certain nombre de propriétés qui justifient son opportunité dans l'objectif précédent. Nous recensons et démontrons ces propriétés puis les observons sur des données expérimentales. Les principales propriétés sont :

- EFR et SphN : gaussianisation, maximisation de la dispersion, mise au jour d'une base commune aux variabilités locuteur et nuisible (*eigen factors*), meilleure conformité des données de tests aux paramètres de l'apprentissage.
- EFR : optimalité déterministe de la séparation des variabilités en facteurs, équivalence des démarches déterministes LDA et NAP.
- SphN : transformation du modèle en modèle isotropique, quasi-équivalence de ce modèle déterministe aux modèles génératifs obtenus par maximum de vraisemblance.

La convergence des vecteurs transformés vers un modèle théorique plus cohérent et plus fiable entraîne une amélioration des performances des systèmes. Le gain moyen procuré par l'application de ces transformations, qu'il s'agisse d'EFR pour le modèle LDA-two-covariance ou de SphN pour le modèle PLDA-gaussienne, est de l'ordre de 50% en terme de diminution du taux d'erreur. Dans le dernier cas (SphN+PLDA-gaussienne), ce gain permet d'atteindre, par la voie du cadre probabiliste gaussien, les meilleurs taux de détection dans le domaine. Il est de l'ordre de 35% par rapport à l'approche LDA-WCNN-cosine scoring de (Dehak et al., 2011).

Les investigations et solutions qui ont été exposées rappellent que deux stratégies s'affrontent souvent dans le traitement des représentations vectorielles d'objets à discriminer :

La première considère que les versions numériques des objets, de par leur procédé d'extraction, suivent nécessairement une loi théorique qu'il s'agit de déterminer (statistique *confirmatoire*). Lorsque le phénomène vectorisé est formé d'une accumulation considérable de variables aléatoires indépendantes, sans qu'aucune ne soit dominante, le choix de la loi gaussienne paraît le plus pertinent. Eventuellement, une loi de la même famille (Student) peut s'accorder aux quelques anomalies constatées. Dans le cadre des modèles dits génératifs, tels que la PLDA par EM-ML, le vecteur représentant les observations n'est en aucun cas modifié, les facteurs à extraire étant estimés

par maximum de vraisemblance à posteriori de ces vecteurs. La collection d'apprentissage est ainsi soigneusement préservée tandis que les paramètres à estimer évoluent. La même démarche se retrouve dans les méthodes de réduction de dimensionnalité : elles conservent seulement la partie principale d'une représentation donnée.

La seconde stratégie applique des transformations non nécessairement affines, linéaires ou paramétriques aux représentations vectorielles. Cela revient à considérer ces versions numériques des objets comme des contenants potentiels d'informations explicatives, mais non ou insuffisamment distribués suivant une loi théorique. La question de la pertinence des transformations à appliquer se pose alors (statistique *exploratoire*). Les techniques de normalisation précédentes dans l'espace des i-vectors s'apparentent, par exemple, à celle du noyau de Fisher et du classifieur SVM : dans cette méthode, les données sont transportées dans un sur-espace par une fonction non-linéaire, afin de linéariser les cloisons entre classes. Les données ont été rendues compatibles avec des outils mathématiques à disposition : sous-espace linéaire, distributions usuelles. Le choix des transformations appliquées n'est souvent justifié qu'à posteriori, par la voie de l'empirisme.

L'utilisation de techniques de normalisation dans l'espace des i-vectors n'est ni concurrente ni incompatible avec les décompositions de la première stratégie, mais complémentaire. Le réalisme voudrait que ces deux outils soient toujours à disposition sur la table du chercheur, plutôt qu'en balance au nom d'*écoles* statistiques.

Chapitre 5

Fondements du paradigme

Les propriétés des vecteurs, après les transformations présentées dans le chapitre précédent, montrent l'opportunité de ces transformations dans le cadre de la reconnaissance du locuteur. Ces techniques de normalisation visent aux mêmes objectifs que les représentations par supervecteurs, décomposition en facteurs par FA-total var ou modèle génératif PLDA : optimiser la qualité de la modélisation en respectant des hypothèses et contraintes gaussiennes. Au delà des performances, les résultats expérimentaux confirment le rôle de ces transformations dans l'élaboration de représentations fiables des énoncés de voix pour la RAL.

Dans notre cadre d'étude, la réduction de dimensionnalité suivant la variabilité totale suivie d'une technique de normalisation telle qu'EFR ou SphN apparaît comme un facteur déterminant de réussite de la solution i-vector. Nous avançons l'hypothèse que la participation de cette approche aux objectifs de qualité de la modélisation est prépondérante.

La part jouée par chacun des modules de la chaîne de traitement dans la qualité doit donc être évaluée, pour mieux cerner leurs impacts respectifs dans la réussite de la démarche. Cette évaluation fait l'objet de ce chapitre. Le dépouillement de cette évaluation va également permettre d'isoler les facteurs-clés du concept : des étapes fondamentales -et non plus des méthodes précises- qui assurent la qualité d'un système de discrimination du locuteur basé sur ce concept.

5.1 Objectifs

La figure 5.1 affiche la chaîne de traitement i-vectors de notre étude, en y incluant maintenant la phase de normalisation. Une fois le signal vocal paramétrisé dans l'espace acoustique et les trames utiles sélectionnées par VAD, un GMM-UBM a été constitué pour structurer cet espace. A partir de ces étapes initiales, que nous ne remettons pas en question, l'ensemble des étapes du système emploie des méthodes (représentation des énoncés par statistiques d'ordre 0 et 1 du GMM-UBM, extraction des i-

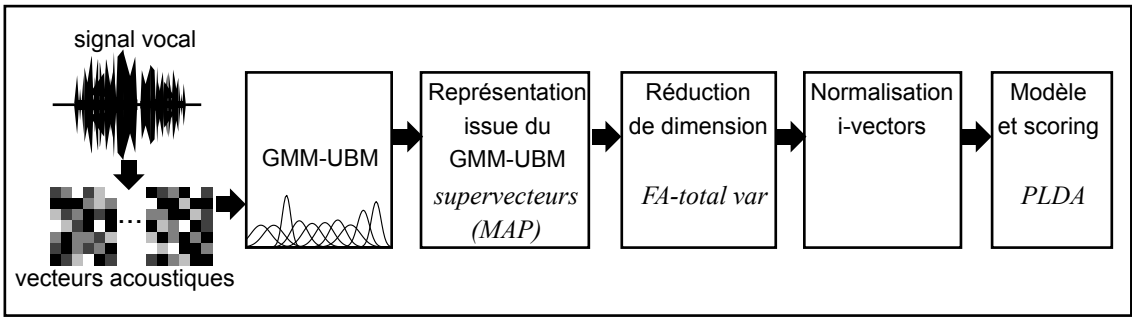

FIGURE 5.1 – *Les étapes d'un système de reconnaissance du locuteur basé sur le concept de i-vectors.*

vectors par FA-total-var, modélisation PLDA) qui tentent, en respectant des contraintes d'a priori gaussien, de modéliser et traiter au mieux la représentation compacte par i-vectors. Les phases successives de représentation par UBM et d'extraction doivent produire une représentation compacte qui puisse être considérée comme une observation d'un modèle génératif, de nature gaussienne. Cette gaussianité s'entend au sens le plus large du terme, le i-vector devant être décomposable en facteurs indépendants de variabilité eux-mêmes gaussiens.

Nous avons montré au chapitre précédent que les techniques de normalisation jouaient un rôle déterminant dans la qualité du système, en participant à ces objectifs de modélisation. Les expériences qui testaient plusieurs de ces techniques (avec standardisation suivant $\mathbf{\Sigma}$ ou $\mathbf{W}$) ont déjà conduit à comparer plusieurs modèles i-vectors. Si le modèle PLDA s'avère le plus performant, l'écart avec le modèle déterministe LDA-two-covariance est réduit, en particulier s'il est comparé à celui en l'absence de normalisation. Ces différents constats soulèvent plusieurs questions :

- quelle part joue chaque étape du système dans la qualité de la modélisation ? L'adéquation des i-vectors à un modèle génératif tient-elle à l'ensemble des méthodes, ou principalement à certaines d'entre elles ?
- peut-on quantifier ces impacts ? Les effets des différentes procédures, en terme de performance, peuvent-il être isolés ?
- pour orienter les recherches futures, une telle évaluation peut-elle dégager les facteurs-clés de la réussite du concept i-vectors ? Les étapes du système ont été élaborées sur une vingtaine d'années, au fur et à mesure de l'avancement du domaine. La pertinence des démarches successives dans le nouveau contexte i-vectors est-elle toujours effective ?

Notre objectif est d'évaluer la part, dans la réussite des systèmes i-vectors, des trois méthodes état-de-l'art :

- représentation par supervecteurs-MAP,
- réduction de dimensionnalité par FA-total variability,
- PLDA gaussienne

puis de les comparer à la part imputable aux procédures de normalisation. Pour procéder, nous proposons une alternative à chacune de ces trois méthodes, basée sur une approche déterministe ou non-paramétrique. Lorsque la comparaison porte sur la représentation issue du GMM-UBM, nous conserverons pour les vecteurs obtenus, par compression suivant la variabilité totale, la terminologie de *i-vector*, ce terme (*intermediate vectors*) nous paraissant toujours adapté.

Pour nous, la synthèse des résultats va permettre d'aller plus loin. Nous pensons qu'en mettant en avant des phases essentielles dans la qualité des systèmes (celles qui assurent des bons résultats, indépendamment des détails de ces composants), elle dégagera des facteurs-clés dans leur réussite. Il sera alors possible de définir de manière claire le "concept i-vector", en énumérant les seuls éléments et étapes indispensables à son efficacité, éclairant ainsi précisément les voies de recherche futures à explorer.

La section suivante présente une évaluation des étapes d'un système par i-vectors, en terme d'impact des méthodes et hypothèses sur la qualité de la modélisation et de la détection. La section 5.3 prolonge l'étude par une analyse complémentaire, autour du GMM-UBM et de la réduction de dimensionnalité, puis décrit les fondements du paradigme i-vectors que l'ensemble de ces investigations nous a permis de cerner et formaliser.

5.2 Evaluation des étapes d'un système basé sur les i-vectors

5.2.1 Protocole expérimental

Les expériences ont été menées à partir des i-vectors du LIA obtenus par la configuration décrite dans la partie 2.7 et en annexe A. La dimension des i-vectors est de 400.

Afin de comparer de la manière la plus objective possible les différents systèmes évalués, nous avons effectué pour chacun d'entre eux les évaluations sur trois conditions : NIST-SRE 2008 *det 7* et *det 6* et NIST-SRE-2010-det5Extended. Ces trois conditions étant actuellement très couramment utilisées dans les publications de la communauté, nous considérerons leurs performances comme des références fiables. Pour évaluer l'ensemble des systèmes obtenus par croisement des différentes options, nous avons testé sur la configuration LIA ces trois conditions, puis une mesure globale de performance d'un système a été produite par la moyenne de leurs EER [1]. Les résultats présentés dans ce chapitre concernent les fichiers d'évaluation hommes seulement, mesurés en terme d'EER. Les tableaux de résultats en terme de DCF minimale et pour les

1. Il peut être remarqué que la condition *det 7* de NIST-SRE 2008 est une partie de *det 6* de la même campagne. Nous avons choisi de faire participer les deux conditions à part égale dans l'EER moyen final. Il s'avère que det 7 est assez homogène pour tester un système mais réduit en nombre d'essais, alors que det 6, plus complète, inclut des tests de locuteurs très atypiques, en nombre assez important pour amortir les gains de performance. Il est difficile de faire progresser significativement l'EER avec cette dernière condition, même avec un système très performant.

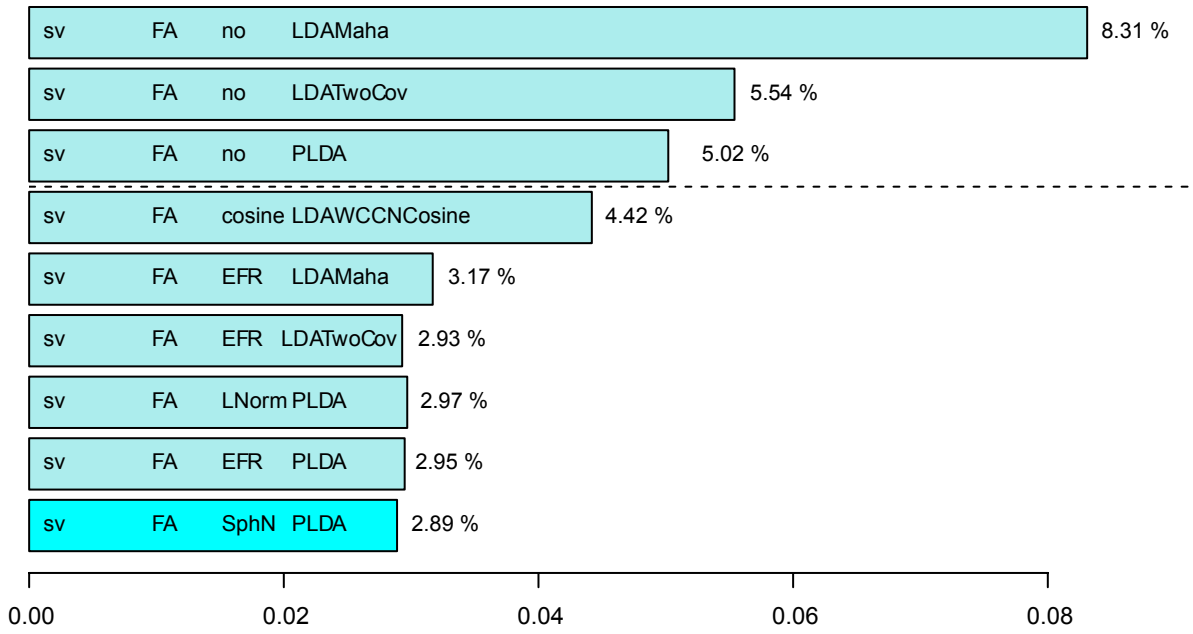

FIGURE 5.2 – *EER moyens de différents modèles et scorings i-vectors. Les méthodes sans normalisation sont affichées dans la partie supérieure, séparées par une ligne horizontale en pointillés.*

fichiers d'évaluation femmes sont présentés en annexe D. Ils confirment les conclusions de ce chapitre.

Les EER optimaux par condition sont systématiquement fournis. S'il s'agit d'expériences dépendant d'un rang de sous-espace (LDA ou PLDA), les rangs optimaux ne sont pas forcément identiques entre les méthodes. La hiérarchie de performance des méthodes que nous comparons ici n'a été infirmée par aucune des expériences que nous avons menées, ni par aucune publication dans le domaine.

5.2.2 Modèles et scorings i-vectors

Nous prolongeons l'étude réalisée au chapitre précédent, qui comparait des systèmes utilisant différents modèles et scorings, avec ou sans normalisation. Les modèles LDA-Mahalanobis et LDA-two-covariance sont des cas particuliers déterministes de la PLDA. Les métaparamètres matriciels des distribution locuteur et résiduelle y sont estimés empiriquement, sans faire appel à l'algorithme EM par maximum de vraisemblance sous hypothèses gaussiennes. Le scoring LDA-WCCN-cosinus inclut implicitement une normalisation et s'appuie également sur les estimations déterministes LDA et WCCN. Ces méthodes permettent donc de mesurer l'impact, en terme de performance, de la prise en compte de contraintes probabilistes par maximum de vraisemblance.

Le graphique 5.2 affiche les EER (%) moyens des systèmes comparés. La table 5.1 détaille les EER par condition. Chaque ligne du graphique et de la table indique la

EER (%)					2008		2010	
	repr	reduc dim	norm	scoring	det 7	det 6	det 5 Ext	**moy.**
1	sv	FA	non	LDAMaha	5.70	9.50	9.73	**8.31**
2	sv	FA	non	LDATwoCov	3.41	6.78	6.43	**5.54**
3	sv	FA	non	PLDA	3.18	6.41	5.48	**5.02**
4	sv	FA	cosine	LDA-WCCN-Cosine	3.30	6.29	3.68	**4.42**
5	sv	FA	EFR	LDAMaha	1.85	5.03	2.62	**3.17**
6	sv	FA	EFR	LDATwoCov	1.53	4.91	2.36	**2.93**
7	sv	FA	LNorm	PLDA	1.65	4.88	2.39	**2.97**
8	sv	FA	EFR	PLDA	1.59	4.92	2.33	**2.95**
9	sv	FA	SphN	PLDA	1.59	4.80	2.27	**2.89**

TABLE *5.1* – *Performances en terme d'EER de différents scorings avec ou sans normalisation, sur diverses conditions NIST-SRE.*

nature du système testé :

- **sv** : indique que la représentation utilisée est celle des supervecteurs adaptés par MAP. Cette configuration, commune à tous les systèmes, est indiquée pour homogénéïser la présentation des résultats.
- **FA** : indique qu'a été réalisée une extraction d'i-vectors par Factor analysis-total variability. Même remarque.
- **norm** : indique si une normalisation a été effectuée et de quel type : **Lnorm** pour une seule normalisation de longueur sur vecteurs centrés, **EFR**, **SphN**, ou **cosine** pour WCCN. Dans le cas indiqué **non**, précisons que les vecteurs d'apprentissage et de test sont tout de même centrés suivant la moyenne de l'apprentissage. Cette opération n'est qu'une translation, qui ne mérite pas le titre de "normalisation" mais évite un doute sur la validité de la démarche.
- **scoring** : il s'agit des différents modèles et scorings présentés précédemment.

Sur le graphique, les méthodes sans normalisation sont affichées dans la partie supérieure, séparées des autres méthodes par une ligne horizontale en pointillés.

Parmi les systèmes avec normalisation, celui utilisant le score LDA-WCCN-cosinus (ligne 4) est le moins performant, avec un EER moyen de 4.42%, mais il est significativement meilleur que tous les systèmes sans normalisation (lignes 1 à 3), qui ne descendent pas au-dessous de 5.02%.

Le meilleur système est celui enchaînant normalisation de type SphN et PLDA (dernière ligne), ce qui ne fait que confirmer l'étude réalisée au chapitre précédent. Son écart de performance significatif avec celui utilisant LDA-WCCN-cosinus (2.89% vs 4.42%) montre la pertinence des approches de décompositions en facteurs probabilistes et scoring par log-ratio de vraisemblance.

Mais le principal constat de ces expériences est la part prépondérante jouée par les techniques de normalisation dans la qualité de la détection. La normalisation en-

traîne en effet une réduction de l'ordre de 42% des taux d'erreur (2.89% vs 5.02% entre les meilleurs systèmes avec ou sans normalisation) Si la sophistication probabiliste de la PLDA reste indispensable pour atteindre les meilleurs résultats, elle n'y participe que faiblement (2.89% vs 2.93% pour le meilleur système déterministe, avec LDA-two-covariance, de la ligne 6). L'essentiel de la progression est imputable à la normalisation et induite par une meilleure adéquation à la modélisation gaussienne.

Ce constat est renforcé par le fait que le dernier et meilleur système est basé sur une PLDA initialisée de manière déterministe, qui n'a nécessité qu'un faible nombre d'itérations de l'algorithme EM pour converger (10 contre 100 pour l'avant-dernier système initialisé aléatoirement). La part de la modélisation PLDA par maximum de vraisemblance est minime lorsque l'espace des i-vectors a été préalablement conditionné par les techniques de normalisation visant à une modélisation de type gaussien. L'étude de conservation de l'orthogonalité des métaparamètres de la PLDA réalisée au chapitre précédent avait déjà attiré l'attention sur le fait, qu'une fois l'espace normalisé et les métaparamètres de la PLDA correctement initialisés, la méthode probabiliste n'affinait ces paramètres que dans des proportions modestes.

Un dernier point mérite l'attention : le score de Mahalanobis (systèmes des lignes 1 et 5) est moins performant que ceux des modèles LDA-two-covariance et PLDA. Mais s'il procure 10% d'erreur supplémentaire par rapport au meilleur système après normalisation (3.17% vs 2.89%), ce nombre atteint 66% sans normalisation (8.31% vs 5.02%). En absence de normalisation, l'écart est considérable et ne se retrouve pas sur les systèmes utilisant le modèle LDA-two-covariance. Le score de Mahalanobis ne s'appuie que sur l'estimation de la variabilité intra-locuteur. Il est donc très sensible à la précision de son estimation, les autres s'appuyant et se renforçant par la modélisation de la variabilité inter-locuteur. Ce fait montre donc que le manque de gaussianité des i-vectors en sortie de l'extracteur est essentiellement dû à la distribution intra-locuteur. Il confirme l'hypothèse, avancée dans (Dehak et al., 2011), que l'information non-locuteur affecte la norme des i-vectors et que l'ignorer augmente la robustesse du i-vector.

5.2.3 Réduction de dimension

Un i-vector d'un énoncé de voix est obtenu par extraction, depuis les paramètres GMM adaptés de cet énoncé, d'un facteur de variabilité totale suivant un algorithme EM-ML de Factor Analysis. Etudier l'impact de l'algorithme EM par maximum de vraisemblance dans la qualité du système peut être effectué en extrayant des vecteurs compressés par la plus usuelle des techniques vectorielles de réduction de dimension : l'analyse en composantes principales (PCA)(Jolliffe, 2002).

Des PCA sur les supervecteurs sont déjà réalisées par certains organismes comme extracteur de "*i-vectors*" (au sens élargi de vecteurs de taille réduite d'un énoncé de voix) (Yaman et al., 2011) (Campbell et al., 2012). Pour notre part, nous avons participé à une étude (Larcher et al., 2012b) sur les potentialités de cette méthode. Nous présentons ici la suite de nos investigations, qui s'appuie sur la problématique suivante : sur quelle représentation issue du GMM-UBM doit être effectuée la PCA pour la rendre comparable

avec la Factor Analysis-Total variability ?

L'extracteur de i-vectors par PCA

Soit F la dimension de l'espace acoustique et G le nombre de composantes de la mixture de gaussienne de l'UBM. Considérant la collection $\mathcal{X} = \{x_t\}_t$ des vecteurs acoustiques d'un segment donné, la statistique d'ordre 0 de ce segment est le vecteur de dimension G dont la $g^{\text{ème}}$ composante est :

$$\sum_t \gamma_g(t) \tag{5.1}$$

où $\gamma_g(t)$ est la probabilité d'occupation de la trame x_t pour la $g^{\text{ème}}$ composante.

Soit $N_{\mathcal{X}}$ la matrice de dimension $FG \times FG$ des statistiques d'ordre 0 de ce segment : $N_{\mathcal{X}}$ est la matrice diagonale par blocs dont les blocs-diagonaux $F \times F$ sont les $n_g I$ où I est la matrice identité $F \times F$, tel que, pour tout g de $[1, G]$:

$$n_g = \sum_t \gamma_g(t) \tag{5.2}$$

La valeur n_g est l'effectif de trames associées probabilistiquement à la gaussienne g. De même, la statistique centrée $S_{\mathcal{X}}$ d'ordre 1 du segment est le vecteur de dimension FG obtenu par concaténation des G vecteurs $S_{\mathcal{X},g}$ tels que :

$$S_{\mathcal{X},g} = \sum_t \gamma_g(t)\left(x_t - \mu_g\right) \tag{5.3}$$

où μ_g est la moyenne de l'UBM pour la $g^{\text{ème}}$ composante gaussienne.

Le supervecteur de ce segment est le vecteur s de dimension FG obtenu par :

$$S_{\mathcal{X}} = N_{\mathcal{X}}(s - \mu) \tag{5.4}$$

où μ est le supervecteur de l'UBM obtenu par concaténation des μ_g.

Une dimension de réduction p ayant été fixée, le i-vector w de ce segment est le vecteur de dimension p calculé à partir du postulat :

$$s = \mu + \mathbf{T}w \tag{5.5}$$

La matrice T, de dimension $FG \times p$, est estimée par un algorithme EM-ML pour répondre aux hypothèses de la factor analysis de normalité standard des i-vectors d'apprentissage et du résidu. La solution est :

$$w = \left(\mathbf{I} + \mathbf{T}^t\mathbf{\Sigma}^{-1}N_{\mathcal{X}}\mathbf{T}\right)^{-1}\mathbf{T}^t\mathbf{\Sigma}^{-1}S_{\mathcal{X}} \tag{5.6}$$

où Σ est la matrice de covariance du monde. La relation entre i-vector w et supervecteur s s'écrit :

$$w = \left(\mathbf{I} + \mathbf{T}^t\mathbf{\Sigma}^{-1}N_{\mathcal{X}}\mathbf{T}\right)^{-1}\mathbf{T}^t\mathbf{\Sigma}^{-1}N_{\mathcal{X}}\left(s-\mu\right) \tag{5.7}$$

Considérons maintenant la réduction de dimensionnalité amenée par une PCA effectuée sur les supervecteurs. La matrice $cov\left(s\right)$ de covariance totale des supervecteurs peut être décomposée en valeurs singulières $cov\left(s\right) = \mathbf{PDP}^t$ où $\mathbf{P}$ est la matrice orthogonale de vecteurs propres et $\mathbf{D}$ la matrice diagonale des valeurs propres. Notons $\mathbf{P}_{[a,b]}$ la matrice $FG \times p$ constituée des $a^{\text{ème}}$ à $b^{\text{ème}}$ colonnes (vecteurs-propres) de P. La PCA va produire un vecteur w de IR^p vérifiant :

$$w = \mathbf{P}^t_{[1,r]}\left(s-\mu\right) \tag{5.8}$$

Une méthode de réduction de dimensionnalité commet une erreur ε (dite en PCA "erreur de reconstruction"). Ici, le modèle PCA peut s'écrire $s = \mu + \mathbf{T}w + \varepsilon$ où :

$$\begin{cases} \mathbf{T} = \mathbf{P}_{[1,r]} \\ w = \mathbf{P}^t_{[1,r]}\left(s-\mu\right) \\ \varepsilon = \mathbf{P}_{[r+1,p]}\mathbf{P}^t_{[r+1,p]}\left(s-\mu\right) \end{cases} \tag{5.9}$$

La matrice de covariance théorique des vecteurs w obtenus est alors :

$$cov\left(w\right) = E\left[\mathbf{P}^t_{[1,r]}\left(s-\mu\right)\left(s-\mu\right)^t\mathbf{P}_{[1,r]}\right] = \mathbf{P}^t_{[1,r]}cov\left(s\right)\mathbf{P}_{[1,r]} \tag{5.10}$$

Afin d'assurer le caractère standard des vecteurs w, soit pour obtenir des vecteurs de basse dimension dont la moyenne soit nulle et la matrice de covariance égale à l'identité, le procédé le plus simple consiste à effectuer une PCA sur les supervecteurs modifiés $\Sigma^{-\frac{1}{2}}\left(s-\mu\right)$

On a alors :

$$cov\left(w\right) = \mathbf{\Sigma}^{-\frac{1}{2}}cov\left(s\right)\mathbf{\Sigma}^{-\frac{1}{2}} \tag{5.11}$$

qui doit normalement être égale à l'identité. Mais l'expérimentation pratique montre que la covariance des grands jeux de supervecteurs d'apprentissage utilisés pour la production d'i-vectors diffère significativement de celle, Σ, de l'UBM.

Campbell et al. (Campbell et al., 2012) proposent d'effectuer la réduction de dimensionnalité par la PCA suivante aux supervecteurs :

$$\mathbf{U}^t_{[1,r]}\left(\mathbf{\Sigma} + \mathbf{\Sigma}_n\right)^{-1}\left(s-\mu\right) \tag{5.12}$$

où $\mathbf{U}_{[1,r]}$ est la matrice des r premiers vecteurs propres de la matrice de covariance $\mathbf{\Sigma}$ (donc le sous-espace de rang r de la projection par PCA) et $\mathbf{\Sigma}_n$ est une matrice diagonale utilisée pour modéliser le bruit. L'équation 5.12 peut être interprétée comme la projection d'un supervecteur à filtre de Wiener dans un espace de i-vectors.

La PCA à appliquer demande de s'accorder sur le vecteur de haute dimension issu du GMM-UBM adopté comme représentation de la session de voix. Nous nous proposons ici de choisir celle sur laquelle est appliquée en réalité la Factor Analysis, rendant comparables les deux techniques d'extraction. Intéressons-nous pour cela au cas "extrême" de la PCA sur les supervecteurs, dans lequel la dimension finale p est égale à l'initiale FG. Le modèle devient :

$$\begin{cases} \mathbf{T} = \mathbf{\Sigma}^{-\frac{1}{2}}\mathbf{P} \\ w = \mathbf{P}^t\,(s-\mu) \\ \varepsilon = 0 \end{cases} \tag{5.13}$$

A fortiori, le cas -plus qu'extrême !- où $\mathbf{T} = \mathbf{I}$ (matrice identité $p \times p$) aboutit à :

$$\begin{cases} T = I \\ w = s - \mu \\ \varepsilon = 0 \end{cases} \tag{5.14}$$

ce qui produit, trivialement, un i-vector égal au supervecteur centré.

Reprenons maintenant ce résultat dans le cas de la Factor Analysis. Si la matrice $\mathbf{T}$ de variabilité totale est égale à l'identité $\mathbf{I}$, alors le modèle obtenu s'écrit :

$$\begin{cases} \mathbf{T} = \mathbf{I} \\ w = \left(\mathbf{I} + \mathbf{\Sigma}^{-1}N_{\mathcal{X}}\right)^{-1}\mathbf{\Sigma}^{-1}N_{\mathcal{X}}\,(s-\mu) = N_{\mathcal{X}}\,(\mathbf{\Sigma}+N_{\mathcal{X}})^{-1}\,(s-\mu) \\ \varepsilon = \mathbf{\Sigma}\,(\mathbf{\Sigma}+N_{\mathcal{X}})^{-1}\,(s-\mu) \end{cases} \tag{5.15}$$

Pour chaque gaussienne g, la coordonnée $w_{g,k}$ de la $k^{\text{ème}}$ dimension ($k \in [1, F]$) pour la gaussienne g du vecteur w vaut :

$$w_{g,k} = \frac{n_g}{n_g + \left(\mathbf{\Sigma}_g\right)_{k,k}}\,(s-\mu)_{g,k} \tag{5.16}$$

où $\left(\mathbf{\Sigma}_g\right)_{k,k}$ est la $k^{\text{ème}}$ valeur diagonale, pour la gaussienne g, de la matrice de covariance Σ de l'UBM.

et de même

$$\varepsilon_{g,k} = \frac{\left(\mathbf{\Sigma}_g\right)_{k,k}}{n_g + \left(\mathbf{\Sigma}_g\right)_{k,k}}\,(s-\mu)_{g,k} \tag{5.17}$$

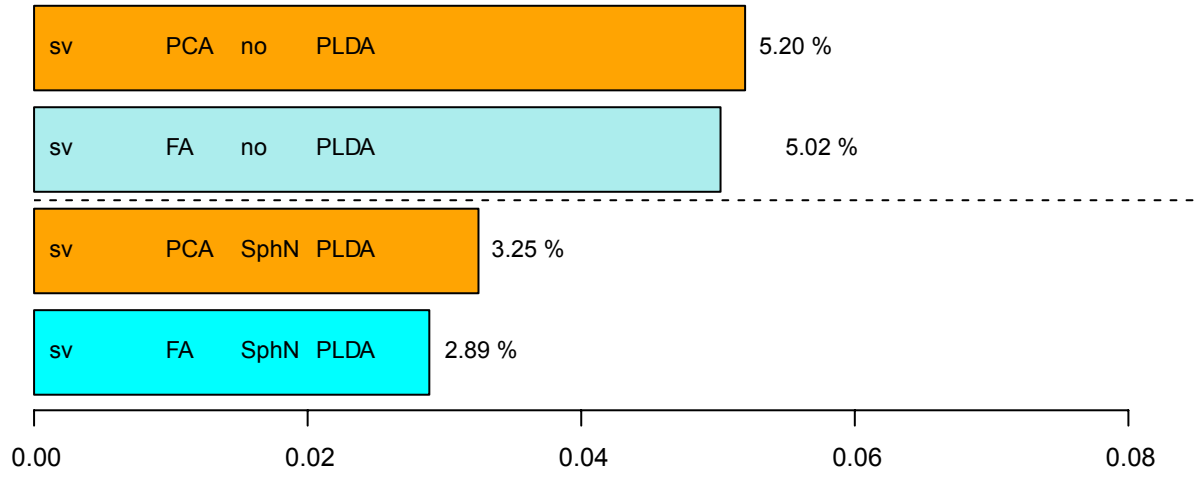

FIGURE 5.3 – *EER moyens suivant la méthode de réduction de dimension (FA-total var ou PCA) avec ou sans normalisation.*

Dans le cas trivial où $\mathbf{T} = \mathbf{I}$, la valeur de w n'est donc pas égale à celle du supervecteur centré, mais au supervecteur **adapté** par une procédure qui s'apparente à la technique MAP de maximum à posteriori. Les supervecteurs, donc les moyennes par gaussienne, sont translatés pour les rapprocher de l'UBM suivant une intensité inversement proportionnelle à la quantité d'information disponible. Mais, contrairement à la technique MAP usuelle, le facteur de pertinence (*relevant factor* τ) n'est pas constant entre les gaussiennes, mais égal à la variance par dimension. Nous remarquons l'analogie, ou la correspondance, entre *dispersion par dimension* et *quantité d'informations contenues dans cette dimension*.

Les i-vectors obtenus par la Factor analysis-total variability sont donc des réductions de dimension effectuées sur des supervecteurs préalablement adaptés par la formule précédente et non sur les supervecteurs d'origine. Pour comparer les performances d'une PCA à celles de la FA-Total Var, nous effectuerons donc la PCA sur ces supervecteurs adaptés. Le vecteur w obtenu vérifiera :

$$w = \mathbf{P}_{[1,r]}\widehat{s} \tag{5.18}$$

où $\mathbf{P}$ est la matrice de vecteurs propres obtenue par décomposition en valeurs singulières de la matrice de covariance des supervecteurs adaptés $\widehat{s}$, définis par :

$$\widehat{s} = N_{\mathcal{X}} \left(\mathbf{\Sigma} + N_{\mathcal{X}}\right)^{-1} (s - \mu) \tag{5.19}$$

Résultats

Le graphique 5.3 et la table 5.2 indiquent les résultats obtenus sur le protocole expérimental. Pour en clarifier le dépouillement, tous les systèmes comparés utilisent le

EER (%)					2008		2010	
	repr	reduc dim	norm	scoring	det 7	det 6	det 5 Ext	**moy**
1	sv	PCA	non	PLDA	3.18	6.63	5.80	**5.20**
2	sv	FA	non	PLDA	3.18	6.41	5.48	**5.02**
3	sv	PCA	SphN	PLDA	2.05	5.21	2.48	**3.25**
4	sv	FA	SphN	PLDA	1.59	4.80	2.27	**2.89**

TABLE 5.2 – *Performances, en terme d'EER, successivement à différentes techniques de réduction de dimension.*

type de normalisation et le modèle le plus performant (SphN et PLDA). Les extractions de "i-vectors" par PCA et FA-total variability, avec ou sans normalisation, sont testées.

La Factor Analysis-total variability est la méthode la plus performante. Elle procure un gain relatif d'EER de 11% (2.89% vs 3.25% pour la PCA). Mais cet écart est très réduit en comparaison de celui obtenu par la normalisation (les EER progressent en moyenne d'une valeur relative de 40%). Deux points principaux ressortent de ces expériences :

- La FA-total-var est une méthode de décomposition en facteurs probabiliste. Basée sur une fonction objective de vraisemblance qu'elle cherche à maximiser, elle tente ainsi de prendre en compte les contraintes de modélisation gaussienne nécessaires à la PLDA et que la PCA ignore. L'expérience précédente montre qu'elle prépare mieux les données que la PCA à la suite du traitement, mais l'apport de cette démarche probabiliste s'avère réduit, en terme de performance, s'il est comparé à celui induit par les fonctions de normalisation. Ces techniques participent donc principalement à modéliser les données suivant des contraintes gaussiennes.
- Nous avons appliqué la PCA à des supervecteurs adaptés de telle sorte que la concordance entre les deux approches soit la plus grande possible et donc leur comparaison. Les systèmes de certains laboratoires utilisent une PCA et non une FA-total-var sur leurs versions adaptées des supervecteurs (Campbell et al., 2012) pour produire des i-vectors, jugeant le raffinement probabiliste de cette dernière négligeable en terme de performance et la voie d'investigation de méthodes déterministes plus porteuse.
- l'extraction des i-vectors s'apparente moins à une décomposition en facteurs qu'à une réduction de dimensionnalité suivant la variabilité totale ou, plus brièvement, à une *compression* des représentations en haute dimension concaténant les informations locales du GMM-UBM.

Les vecteurs de moyenne adaptés de l'UBM peuvent donc être compressés pour produire des composantes principales de l'énoncé vocal. Par résumé des comportements corrélés de ces vecteurs suivant la variabilité totale, donc des composantes gaussiennes du GMM, se dégage un espace réduit dont les dimensions contiennent des informations acoustiques de haut niveau du signal de voix. Le i-vector est une représentation compacte de l'énoncé de voix complet (et non d'un ou plusieurs de ces facteurs dans un sous-espace de type *eigenvoices* ou *eigenchannels* comme en JFA), obtenue par une simple réduction de dimensionnalité (compression suivant la variabilité totale).

Le caractère indépendant des variables unidimensionnelles de cet espace permet le traitement pour scoring, mais à la condition que les vecteurs soient d'abord conditionnées par une opération de normalisation. Celle-ci, tout en accentuant cette indépendance, blanchit les données dont la magnitude a été affectée non-linéairement par la variabilité intrinsèque au locuteur.

5.2.4 Représentation issue du GMM-UBM

Le fait qu'une PCA exécutée en lieu et place de la FA-total-var ne dégrade pas significativement les performances ouvre la voie à de nouvelles investigations. Toute représentation vectorielle peut faire l'objet d'une réduction de dimensionnalité par PCA, cette technique ne nécessitant pas comme la Factor Analysis sur mixture de gaussiennes l'élaboration d'un module d'extraction par maximisation d'une fonction de vraisemblance et suivant une distribution a priori. Il est donc possible d'envisager l'extraction de i-vectors à partir d'autres représentations issues du GMM-UBM. Dans le cadre de cette évaluation, nous souhaitons comparer la représentation par supervecteurs adaptés par MAP à une représentation également issue du paradigme GMM-UBM (que nous n'évaluons pas dans cette étude et ne remettons pas en question), mais qui s'en écarte par une approche non-gaussienne voire non-paramétrique. D'autre part, pour permettre une comparaison objective, les mêmes traitements ultérieurs (normalisation, modélisation) seront appliqués aux versions compressées par PCA de cette nouvelle représentation.

La représentation en haute dimension utilisée est celle par vecteurs du modèle de clés binaires du locuteur. Ce modèle présente l'avantage de s'appuyer sur la structuration initiale de l'espace en mixture de gaussiennes, avant transformations non-paramétriques vers un espace binaire. Nous montrons dans la section ultérieure 8 qu'une PCA sur les vecteurs de clés binaires du locuteur est mathématiquement valide : dans le cas d'une variable qualitative binaire, les vecteurs propres et spectre d'une PCA sur les deux modalités codées 0 ou 1 équivalent à ceux de l'analyse des correspondance multiples (*Multiple Correspondance Analysis MCA*) [2] (Lebart et al., 2000).

Nous avons également testé la compression en i-vectors des vecteurs de "compte" du modèle binaire, produits en amont des clés binaires. La configuration utilisée pour produire l'ensemble de ces vecteurs est détaillée dans la section 8.

Le graphique 5.4 et la table 5.3 présentent les résultats de cette comparaison des méthodes de représentation issues du GMM-UBM. La représentation en vecteurs de comptes du modèle binaire est indiquée **BKcounts**, celle en clés binaires du locuteur **BKbinary**. Le modèle utilisé est celui de la PLDA, avec SphN ou sans normalisation. Pour comparer les représentations *toutes choses étant égales par ailleurs*, les résultats du système compressant par PCA les supervecteurs adaptés par MAP sont également rappelés.

Cette dernière représentation procure bien la meilleure performance (2.89%). Cette

2. Cette analyse dégage des composantes principales sur des variables catégorielles.

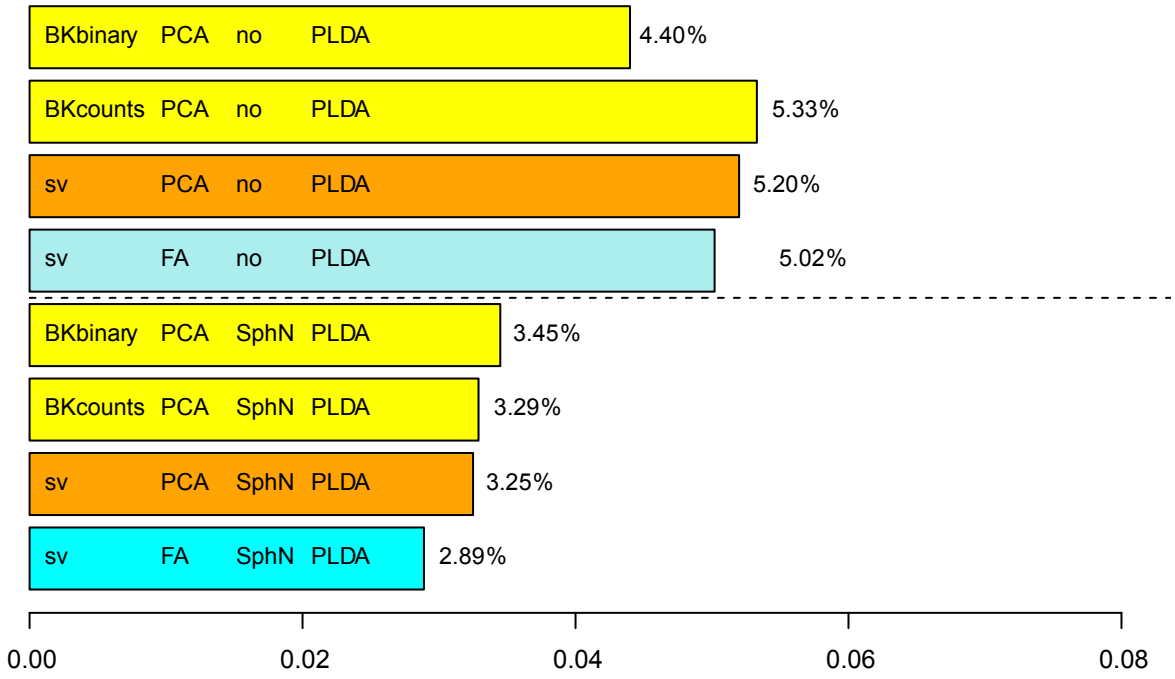

FIGURE 5.4 – *EER moyens suivant la méthode de représenation haute dimension, avec ou sans normalisation.*

EER (%)					2008		2010	
	repr	reduc dim	norm	scoring	det 7	det 6	det 5 Ext	**moy**
1	BKbinary	PCA	non	PLDA	2.94	5.83	4.42	**4.40**
2	BKcounts	PCA	non	PLDA	2.73	7.09	6.16	**5.33**
3	sv	PCA	non	PLDA	3.18	6.63	5.80	**5.20**
4	sv	FA	non	LDAMaha	5.70	9.50	9.73	**8.31**
5	BKbinary	PCA	SphN	PLDA	2.20	5.26	2.89	**3.45**
6	BKcounts	PCA	SphN	PLDA	1.82	5.14	2.91	**3.29**
7	sv	PCA	SphN	PLDA	2.05	5.21	2.48	**3.25**
8	sv	FA	SphN	PLDA	1.59	4.80	2.27	**2.89**

TABLE 5.3 – *Performances, en terme d'EER, successivement à différentes représentations issues du GMM-UBM et techniques de réduction.*

performance est à comparer à celle des représentations par comptes et clés du modèle binaire (3.45% et 3.29%). Le gain relatif d'EER procuré par la représentation en supervecteurs est donc de 16% et 12%.

Afin d'annuler l'effet de la réduction de dimension FA-total-var ou PCA, nous comparons les systèmes seulement basés sur la PCA : le gain relatif d'EER entre les représentations supervecteurs et issues du binaire n'est plus que de 1.2% et 5.8% (3.25% vs 3.45% et 3.29%). En tenant compte de la faible quantité d'informations d'un vecteur binaire (le ratio de taille -en octets- avec les statistiques MAP est de l'ordre de 32 [3]), ce gain limité à 5.8% mérite clairement l'attention. La représentation en haute dimension ne nécessite pas le volume d'informations souvent supposé. Si la diminution du nombre de composantes du GMM-UBM altère les performances, les informations d'un énoncé par gaussienne peuvent être efficacement résumées, suivant une démarche non-paramétrique et sous une forme binaire.

La comparaison des systèmes sans ou avec normalisation (lignes 1-4 vs 5-8) montre nettement la prépondérance de cette technique dans la bonne modélisation des données. La représentation par supervecteurs, basée sur une méthode respectant des contraintes probabilistes a priori, n'apporte qu'un gain limité s'il est comparé à celui induit par la normalisation (45%).

Ces expériences nous fournissent d'autres enseignements : la même procédure de normalisation procure également une nette amélioration des systèmes basés sur les nouvelles représentations. Après compression, le même algorithme conditionne les données pour modélisation et scoring. Un fait se confirme : la nécessité, sur des vecteurs compressés issus de représentations GMM-UBM et porteurs de variables acoustiques de haut niveau, d'homogénéïser les effets qu'elles relèvent (standardisation suivant une variabilité-cible) et de combattre un défaut de distribution gaussienne contenu dans leur magnitude.

Seul bémol à ce constat : le fait que la représentation en clés binaires procure les meilleures performances avant normalisation (4.36% vs 5.19% à 5.38%). Ce fait laisserait penser que la normalisation SphN n'est pas la mieux adaptée à cette représentation. Mais cette dernière contient une pré-normalisation (égalisation des appels par gaussienne, décrite en 8.2) avant binarisation, qui explique la meilleure adéquation des données initiales à la modélisation.

5.2.5 Synthèse

Le graphique 5.5 et la table 5.4 présentent l'ensemble des résultats des expériences précédentes. Ils permettent d'apprécier les effets des différentes alternatives durant les étapes du traitement.

3. Avec notre configuration de GMM-UBM à 512 composantes et un espace acoustique de dimension 50, la taille des statistiques d'ordre 0 et 1 est de 208.9 Ko. Avec 100 spécificités par gaussienne, celle d'une clé binaire est de 6.4 Ko.

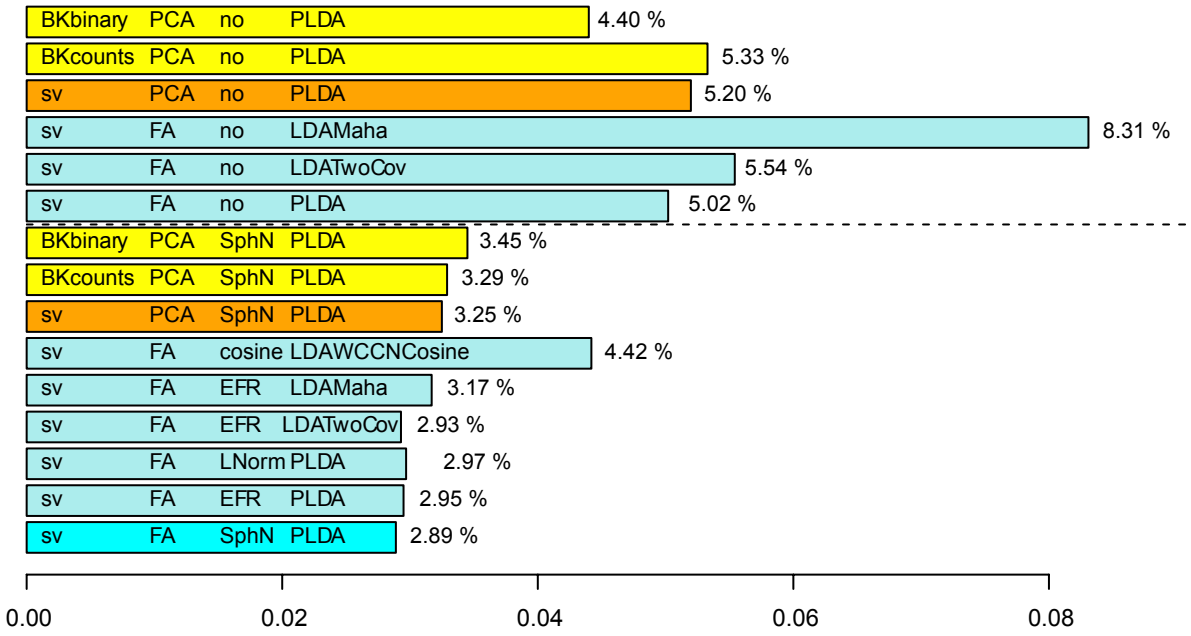

FIGURE 5.5 – *Synthèse des expériences précédentes.*

EER (%)					2008		2010	
	repr	reduc dim	norm	scoring	det 7	det 6	det 5 Ext	**moy**
1	BKbinary	PCA	non	PLDA	2.94	5.83	4.42	**4.40**
2	BKcounts	PCA	non	PLDA	2.73	7.09	6.16	**5.33**
3	sv	PCA	non	PLDA	3.18	6.63	5.80	**5.20**
4	sv	FA	non	LDAMaha	5.70	9.50	9.73	**8.31**
5	sv	FA	non	LDATwoCov	3.41	6.78	6.43	**5.54**
6	sv	FA	non	PLDA	3.18	6.41	5.48	**5.02**
7	BKbinary	PCA	SphN	PLDA	2.20	5.26	2.89	**3.45**
8	BKcounts	PCA	SphN	PLDA	1.82	5.14	2.91	**3.29**
9	sv	PCA	SphN	PLDA	2.05	5.21	2.48	**3.25**
10	sv	FA	cosine	LDA-WCCN-Cosine	3.30	6.29	3.68	**4.42**
11	sv	FA	EFR	LDAMaha	1.85	5.03	2.62	**3.17**
12	sv	FA	EFR	LDATwoCov	1.53	4.91	2.36	**2.93**
13	sv	FA	LNorm	PLDA	1.65	4.88	2.39	**2.97**
14	sv	FA	EFR	PLDA	1.59	4.92	2.33	**2.95**
15	sv	FA	SphN	PLDA	1.59	4.80	2.27	**2.89**

TABLE 5.4 – *Synthèse des performances obtenues par les différents systèmes précédents.*

EER (%)	det 7	det 6	moy.
Best avant i-vectors	2.72	6.29	**4.50**
BKbinary - PCA - EFR - LDAMaha	2.28	5.80	**4.04**
sv -FA - SphN - PLDA	1.59	4.80	**3.19**

TABLE 5.5 – *Comparaison de systèmes avec et sans solution i-vectors.*

Les résultats des systèmes sans procédure de normalisation sont affichés dans la partie supérieure du graphique 5.5, séparés par une ligne pointillée de ceux appliquant cette technique. Comme le montrent ces expériences, la pré-normalisation des i-vectors immédiatement en sortie de l'extracteur conduit à de meilleures performances absolues, quelles que soient les alternatives adoptées à chaque étape du système (EERs moyens dans $[2.89; 4.42]$ contre $[4.40; 8.31]$). La sophistication des étapes par prise en compte de contraintes probabilistes apporte un gain de performance, dans les deux cas. Le tout premier système, par PCA sur clés binaires et PLDA sans normalisation, est la seule exception à cette logique, un résultat qui mérite de futures investigations. Il se dégage de ces observations générales la prépondérance des techniques de normalisation dans le gain de performance, due à une plus forte participation à la bonne qualité de la modélisation probabiliste.

La compression des données en un seul vecteur, représentant l'intégralité de l'énoncé de voix initial, ne constitue une avancée dans le domaine qu'à la condition de transformer ces représentations compactes pour les rendre compatibles aux hypothèses usuelles de modélisation et scoring (la gaussianité pour l'instant). Les efforts réalisés pour compenser leur écarts aux lois usuelles, tels que la Heavy-tailed PLDA, sont évités par l'emploi de ces fonctions de normalisation. Le fait qu'elles puissent être non-linéaires et non-paramétriques doit être admis. Si l'on peut comprendre le scepticisme sur la rationalité de leur emploi tant qu'elles ne sont pas rattachées à une véritable théorie probabiliste, leur efficacité doit servir d'horizon aux futures recherches.

L'évaluation qui a été présentée ici permet de mieux cerner l'impact des différentes méthodes et hypothèses se succédant dans un système de reconnaissance par i-vectors. Elle permet également de mettre en lumière les étapes fondamentales de la démarche. La table 5.5 présente les résultats de trois systèmes, sur les conditions det 7 et 6 de NIST-SRE2008. Le premier est le meilleur système du LIA avant introduction des i-vectors. Il a déjà été présenté en 2.7. Nous rappelons qu'il est obtenu par fusion des scores de deux systèmes, basés sur la Factor Analysis et sur un SVM. Le troisième système est le plus performant parmi ceux de l'évaluation précédente sur les i-vectors. Le gain significatif apporté par la solution i-vectors peut être apprécié (la diminution des taux d'erreur est de l'ordre de 30%). Le second système est basé sur les i-vectors et utilise différentes des alternatives introduites durant l'évaluation précédente : la représentation issue de l'UBM par clés binaires, la réduction de dimensionnalité par PCA, la normalisation EFR et le modèle et scoring LDA-Mahalanobis. Ce système induit une diminution relative des EER de l'ordre de 10% par rapport au premier. Or, il s'agit d'un système individuel, basé sur une représentation qui , comme nous l'avons indiqué précédemment, est 32 fois plus légère en informations que celle par supervecteurs utilisée dans le pre-

mier système. D'autre part, la réduction de dimensionnalité s'opère par une PCA, sans contrainte probabiliste. Enfin, les LDA et matrice de Mahalanobis sont estimés de manière déterministe et le score ignore l'apprentissage de la variabilité locuteur ($\mathbf{B}^{-1} = 0$). Seule la normalisation EFR est employée pour conditionner les données aux hypothèses gaussiennes du score de Mahalanobis.

Si l'apport des sophistications probabilistes entre les deuxième et troisième systèmes est indéniable, la supériorité du second système sur le premier nous livre clairement les clés de la réussite de la solution i-vectors :

1. Le pouvoir structurant du GMM-UBM qui permet de constituer des représentations potentiellement discriminantes, ne se limitant pas aux statistiques d'ordre 0 et 1,
2. la réduction de dimensionnalité suivant la variabilité totale, sans égard pour la variable latente (*compression* plutôt que *décomposition en facteurs*),
3. la technique de normalisation comme principal outil de modélisation des représentations compressées.

Si les techniques de normalisation ont été étudiées en détail dans ce chapitre et les précédents, le pouvoir des représentations issues du GMM-UBM et celui de leur compression peuvent être analysés plus profondément. Ils seront mieux cernés par les investigations complémentaires de la section suivante, qui s'achève par une réunion de l'ensemble des observations et la description des étapes fondamentales du concept de i-vector.

5.3 Investigations complémentaires

Les représentations par statistiques ou clés binaires issues du GMM-UBM sont des concaténations de vecteurs par composante gaussienne. Nous estimons que le pouvoir de leur réduction de dimensionnalité suivant la variabilité totale, sans regard pour la variable locuteur, tient à deux points : d'une part la qualité de la structuration de l'espace acoustique par UBM permettant de dégager les caractéristiques essentielles d'un énoncé de voix, ce même en suivant diverses approches, d'autre part cette représentation présente l'opportunité de résumer les corrélations entre informations par gaussiennes. Le taux de compression drastique entre les représentations et leur version i-vector compressée (de 25600 à 400 par exemple, avec une dimension finale de 400 à 600 comparable au nombre de composantes de la mixture initiale de l'UBM) laisse penser que le résultat constitue un vecteur de composantes principales, incluant ces corrélations entre régions de l'espace acoustique et de là un vecteur de variables acoustiques indépendantes hautement informatives. La seule faiblesse de cette démarche est le manque de conformité de ces nouvelles variables acoustiques aux distributions usuelles des modules de détection. Elle peut cependant être contournée par l'emploi des techniques de normalisation, dont l'efficacité a été démontrée.

Concernant l'importance de la compression des données, plusieurs études extérieures viennent appuyer nos propos :

- Dans (Kenny, 2012), une version de la FA-total-var de complexité allégée est proposée. Celle-ci applique notamment une forme de standardisation aux statistiques d'ordre 1 **durant** les itérations de l'algorithme EM [4]. Cette version allégée lui a permis de produire par algorithme EM des i-vectors de dimension plus importante (jusqu'à 1600). Les performances par rapport au standard 400 se sont légèrement dégradées, à l'exception d'une valeur de DCF min. isolée.
- Dans (Jiang et al., 2012) , une PLDA gaussienne est mise en place sur des *i-supervectors* obtenus par transformation linéaire de plein rang des supervecteurs. Par rapport à leurs i-vectors de dimension 500, ces vecteurs de dimension 27648 ont dégradé les performances sur cinq conditions de NIST-SRE 2008 (perte de l'ordre de 5% en EER relatif) et les ont nettement améliorées sur la sixième, "det 5", mixant données téléphone et microphone.

Une réduction de dimensionnalité drastique apparaît donc comme décisive dans le succès de la démarche par i-vectors. En cherchant à limiter le taux de compression pour conserver une plus grande part de variabilité jugée informative, le système ne modélise pas suffisamment les composantes principales de la représentation issue du GMM-UBM, liées aux corrélations sur ses composantes. De plus, il fournit une représentation dont les dimensions supplémentaires s'accordent moins aux hypothèses probabilistes des modules de détection actuels.

Concernant l'efficacité du couple (représentation par vecteur concaténé d'informations par gaussienne $\rightarrow$ compression suivant la variabilité totale), nous présentons une étude complémentaire qui en approfondit la compréhension.

5.3.1 Analyse du concept i-vector par les accumulateurs GMM et modèle binaire

Le modèle par clé binaire présenté en 2.8 comptabilise les appels aux spécificités par trame dans un vecteur d'accumulation (*comptes*) avant binarisation. Si G est la taille de la mixture en composantes, la dimension du vecteur de comptes est $N = Gq$ où q est le nombre de spécificités par gaussienne. Ce vecteur de comptes constitue dans ce modèle une représentation intermédiaire de l'énoncé de voix.

La modélisation des énoncés par GMM-UBM produit également un vecteur d'*occupation* par gaussienne, qui accumule les probabilités des trames d'appartenance à une gaussienne. Etant donnée une trame x, la probabilité de la gaussienne g à posteriori de x est estimée par :

$$P(g|x) = \frac{P(g,x)}{P(x)} = \frac{P(x|g)\,P(g)}{P(x)} = \frac{\omega_g \mathcal{N}(x|\mu_g, \mathbf{\Sigma}_g)}{\sum_{l=1}^{G} \omega_l \mathcal{N}(x|\mu_l, \mathbf{\Sigma}_l)} \tag{5.20}$$

4. Le pouvoir des standardisations reste un sujet d'étude, ainsi que leurs positionnements dans la chaîne de traitement.

avec $\sum_g P(g|x) = 1$. La somme de ces probabilités d'occupation des G composantes gaussiennes est un vecteur de dimension G. Cette dimension est réduite, en comparaison des supervecteurs et plus généralement des quantités d'informations fournies par les modélisations GMM-UBM.

La dimension du vecteur de comptes N peut être limitée, pour coïncider avec celle du vecteur d'occupation. Considérons alors ces deux expressions vectorielles, issues des trames d'un segment, comme les uniques représentations vectorielles de l'énoncé de voix dont on dispose. Ces représentations vont permettre d'éclairer plusieurs aspects :

- ces vecteurs étant de faible taille, peut-on les considérer comme des i-vectors, capables de permettre la discrimination du locuteur ? Peut-on appliquer les mêmes techniques de normalisation et modélisation (PLDA, ...) sur ces nouvelles représentations compactes ?
- même si leur taille initiale est réduite (nous travaillerons lors de nos expérimentations sur des vecteurs de dimension 2048), gagne-t-on à respecter la phase de réduction suivant la variabilité totale ?
- quel lien existe-t-il entre la quantité d'informations initiale d'une représentation et la performance discriminative obtenue à partir de sa version i-vector ?

5.3.2 Evaluation des représentations par accumulateurs

Pour procéder, les configurations suivantes ont été mises en place :

- les vecteurs d'occupation sont issus d'un GMM-UBM à 2048 composantes produit par le LIA pour la campagne NIST-SRE 2010. Les détails de son contenu peuvent être consultés sur (Larcher et al., 2010).
- les vecteurs de comptes utilisent un GMM-UBM à 128 composantes basé sur la même configuration que celui à 512 composantes décrit en 2.7. Ce vecteur est égalisé dans son nombre d'appels par gaussienne, comme expliqué en 8. Pour chaque trame, les 3 gaussiennes les plus vraisemblables sont sélectionnées, puis les 8 spécificités les plus vraisemblables de chacune d'entre elles allument leur composant correspondant dans la clé binaire. Le nombre de spécificités par gaussienne est fixé à $q = 16$. Le vecteur de comptes obtenu est donc de dimension $128 \times 16 = 2048$.

L'expérience a été menée sur les conditions téléphone-téléphone det 7 et det 6 de NIST-SRE 2008 décrites au paragraphe 2.7. Huit systèmes sont comparés, dont la table 5.6 affiche les résultats en terme d'EER :

- **repr** indique la représentation choisie (par vecteurs d'occupation ω **occup.** ou de comptes du modèle binaire **BKcounts**)
- **dim** indique la dimension des représentations
- **reduc dim** indique si la réduction de dimensionnalité a été effectuée ou pas (**non** ou **PCA**)

syst	repr	dim	reduc dim	dim i-vect	norm	det 7 EER %	det 6 EER %
1	ω occup.	2048	non	2048	non	9.83	14.14
2	ω occup.	2048	non	2048	SphN	7.29	11.30
3	ω occup.	2048	PCA	512	non	7.25	10.63
4	ω occup.	2048	PCA	512	SphN	5.39	8.98
5	BKcounts	$128 \times 16 = 2048$	no	2048	non	4.77	9.25
6	BKcounts	$128 \times 16 = 2048$	no	2048	SphN	3.83	8.10
7	BKcounts	$128 \times 16 = 2048$	PCA	512	non	4.12	8.70
8	BKcounts	$128 \times 16 = 2048$	PCA	512	SphN	3.14	7.11

TABLE 5.6 – *Performances, en terme d'EER, de différents systèmes basés sur des représentations par accumulateurs.*

– **norm** indique si la procédure de normalisation a été effectuée (**non** ou **SphN**) avant PLDA

Les systèmes n°1, 2, 5, 6 utilisent directement les représentations issues du GMM-UBM comme i-vectors, alors que les systèmes n°3, 4, 7, 8 les extraient par réduction de dimensionnalité.

La table 5.6 affiche les résultats de cette évaluation en terme d'EER, sur les conditions det 7 et 6 de NIST-SRE 2008. Deux constats s'en dégagent :

– les vecteurs de compte égalisés dépassent en performance ceux d'occupation (ils procurent une diminution des taux d'erreur de l'ordre de 40 à 50%). Cette capacité confirme la pertinence du modèle binaire, malgré le caractère artificiel de la comparaison (ces vecteurs ne sont pas conçus pour un traitements direct i-vectors). Développer sur ce point n'entre pas dans le cadre traité ici.
– L'algorithme de normalisation des i-vectors s'avère généralisable à des représentations par accumulateurs. Le gain de performance est de l'ordre de 25% en écart relatif d'EER.
– Malgré la faible taille des vecteurs initiaux, l'élimination de résidus vis à vis de la variabilité totale s'avère pertinente. La représentation compacte en 512 dimensions améliore les performances, là aussi d'un facteur de l'ordre de 25% en écart relatif d'EER.
– Le huitième système fournit la meilleure performance : 3.14% pour det 7 et 7.11% pour det 6. Ces performances peuvent être comparées à celle du meilleur système de l'évaluation générale : 1.53% et 4.80%. Si les taux d'erreur sont clairement plus élevés, ils restent bas en regard de la taille de la représentation initiale. Un chiffre peut permettre de mieux apprécier la capacité de ce $8^{\text{ème}}$ système : hors i-vectors, la meilleure performance, en terme d'EER, d'un système unique au LIA (il s'agissait d'une décomposition en facteur de type FA $m + \mathbf{U}x + \mathbf{D}z$), sans normalisation des scores, était de 3.87% pour det 7.

La représentation des énoncés de voix par concaténation d'informations par gaussiennes (ici une seule valeur numérique pour le vecteur de probabilité d'occupation)

ou par spécificités apparaît une nouvelle fois efficiente pour la reconnaissance du locuteur. Rappelons que ces représentations constituent des vecteurs d'un méta-espace de coefficients, issus de l'espace acoustique. La qualité de la structuration de l'espace acoustique du GMM-UBM s'avère une nouvelle fois déterminante.

La réduction de dimensionnalité de ces vecteurs suivant la variabilité totale génère des variables contenant les informations de corrélations entre gaussiennes, qui s'avèrent également déterminantes dans la modélisation statistique des constituants du signal vocal. L'étude du triplet de valeurs (dimension initiale, dimension réduite, performance) mérite une analyse particulière, réalisée ci-dessous.

5.3.3 Dimension initiale et performance en discrimination

La représentation par vecteurs de comptes peut être mise à profit pour étudier la relation entre la dimension de l'espace de représentation initial et la performance discriminative obtenue. Etant donné un GMM-UBM, ici celui à 128 composantes de l'expérience précédente, l'augmentation du nombre q de spécificités par gaussienne permet de former des représentations de taille croissante, puis de leur appliquer la méthode "i-vector" précédente : projection par PCA en dimension 512 puis normalisation SphN, modélisation et scoring PLDA.

L'ensemble $\{4, 8, 16, 256\}$de valeurs de q a été testé, aboutissant à des dimensions $\{512, 1024, 2048, 32768\}$. Le recouvrement en spécificités a été effectué à partir de l'algorithme $\mathbf{\Sigma}^{-1}$-*KNN-kmeans* expliqué en 8

Les graphiques de la figure 5.6, partie 1, affichent les dimensions initiales en abscisse et les EER (%) en ordonnées des conditions det 7 et 6. La dimension 51200 adoptée dans la partie 8 et son EER ont été ajoutés au graphique, pour prolonger la courbe. La ligne horizontale indique les performances optimales obtenues lors des évaluations expérimentales de la section précédente. La décroissance attendue des quantités d'erreur est bien vérifiée. Son caractère asymptotique à l'infini laisse envisager un plafonnement de la performance à la valeur optimale précédente.

La partie 2 du graphique 5.6 présente la même expérience, mais cette fois les abscisses sont les logarithmes en base 2 de la dimension initiale $\ln(\dim) / \ln 2$. La courbe de décroissance du taux d'erreur présente un degré de linéarité plus important. Une corrélation nette apparaît entre les deux grandeurs (de l'ordre de -0.97 et -0.98 pour les deux conditions), ce en gardant à l'esprit qu'il s'agit chaque fois de dimension initiale compressée en dimension 512 (hormis bien sûr le premier modèle à 512 spécificités, immédiatement exploitable). Tôt ou tard, l'épuisement des données d'apprentissage doit conduire à une stagnation asymptotique de l'EER. Le caractère linéaire de la relation entre le logarithme en base 2 de la dimension initiale (égal à p si la dimension est 2^p) et la performance rattache l'étude du modèle binaire à un problème d'entropie en théorie de l'information binaire.

La table 5.7 présente un échantillon intéressant des résultats précédents. Les colonnes 1 et 2 indiquent respectivement la dimension initiale des vecteurs de comptes

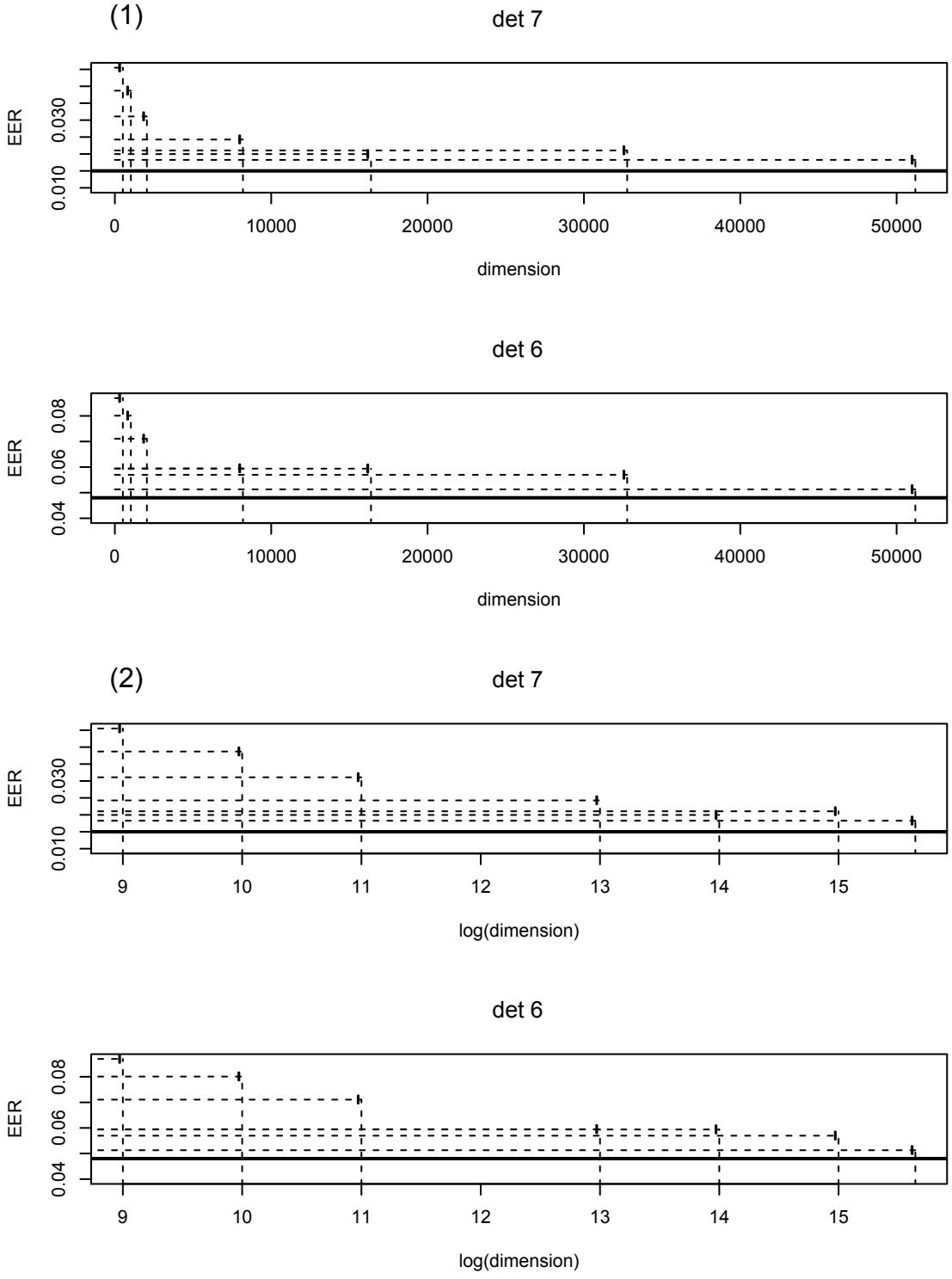

FIGURE 5.6 – *1. Dimensions initiales et EER des systèmes i-vectors basés sur la représentation haute dimension par vecteurs de compte du modèle binaire. 2. Même graphique que précédemment avec les logarithmes en base 2 des dimensions en abscisse.*

dimension		EER	
haute	basse	det 7	det 6
512	*id.*	4.54	8.69
2048	*id.*	3.83	8.10
2048	512	3.14	7.11

TABLE 5.7 – *Comparaison de performances de systèmes basés sur les accumulateurs binaires, en fonction des dimensions initiale et finale.*

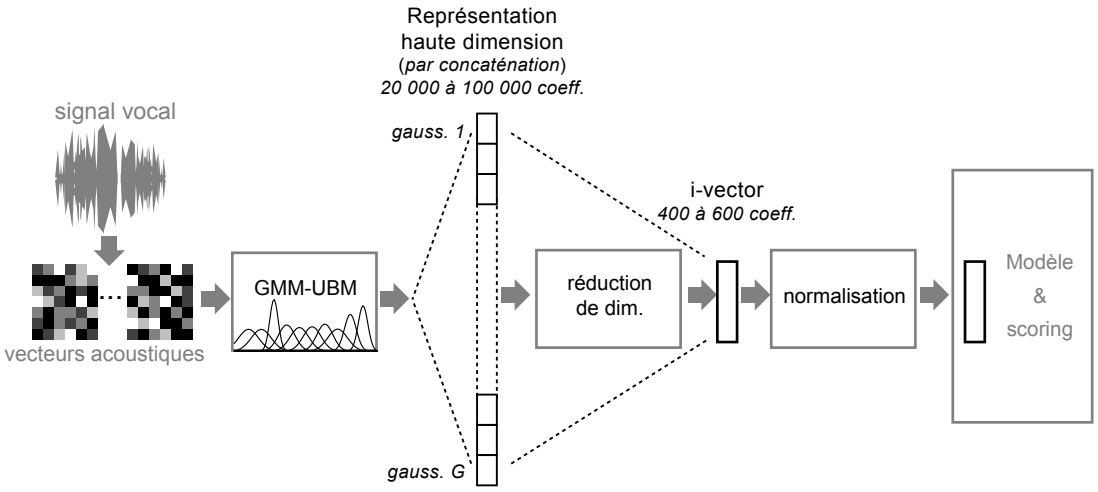

FIGURE 5.7 – *Les étapes du système de reconnaissance du locuteur par i-vectors, reconsidéré à posteriori des analyses précédentes.*

et celle des i-vectors résultants : la PCA n'est effectuée que sur le troisième système. Toutes les performances s'améliorent d'une ligne à l'autre. Les vecteurs 2048 compressés en 512 (3$^{\text{ème}}$ ligne) sont plus discriminants que des vecteurs 512 (1$^{\text{ère}}$ ligne) et que les mêmes vecteurs 2048 directement traités sans compression. C'est donc une nouvelle fois la compression suivant la variabilité totale, aveugle à toute variable latente, qui confirme sa pertinence dans le champ de la discrimination du locuteur. Si la "malédiction de la dimension" a été avancée depuis longtemps dans le domaine de l'apprentissage automatique (par épuisement du contenu informatif de l'apprentissage), le passage par une modélisation en grande dimension reste indispensable, s'il est suivi d'une compression sur la variabilité totale et d'une normalisation non-linéaire [5].

5.4 Bilan et conclusion

L'ensemble des études conduites dans ce chapitre n'a pas amené à reconsidérer la chaîne de traitement d'un système de reconnaissance du locuteur basé sur les i-vectors, ni dans ses étapes, ni dans les méthodes état-de-l'art généralement mises en oeuvre. Mais elles ont permis de mettre en avant le pouvoir prépondérant de certaines étapes et de dégager un concept. La figure 5.7 reprend la figure 5.1 de la chaîne de traitement d'un système i-vector, en la décomposant cette fois en ses véritables étapes fondamentales. Une fois produites les trames acoustiques d'un segment de voix :

1. Une représentation vectorielle de haute dimension est réalisée par concaténation d'informations des composante gaussiennes. Le GMM-UBM est capable de structurer l'espace d'une manière suffisamment cohérente pour former des représentations performantes, ce suivant différentes approches et confirme ainsi son rôle de paradigme.
2. L'extraction d'un i-vector à partir de cette représentation s'apparente plus à une simple réduction de dimensionnalité ("*éliminer la variance globale résiduelle, sans égard pour aucune variable latente*") qu'à une décomposition en facteurs. Les nombreux travaux antérieurs réalisés dans le domaine (FA session, JFA) ont généralisé cette dénomination, mais elle s'entend plutôt lorsqu'un vecteur est décomposé additivement en termes spécifiques, éventuellement réduits, ou lorsque l'unique facteur à produire est dévolu à contenir une variabilité spécifique : locuteur, nuisance. C'est bien l'énoncé de voix complet dont on produit la version vectorielle en faible dimension et non des facteurs associés à certaines de ses variabilités. Cette compression est aveugle à toute variable latente, en particulier la variable locuteur et présente un caractère drastique : de plus de 25000 dimensions à moins de 600-. Le caractère global de la compression (suivant la variabilité totale) et l'ampleur inhabituelle du taux de compression s'expliquent par l'opportunité de résumer les comportements corrélés des informations par gaussiennes, produisant ainsi des **composantes principales** du signal de voix. Ces composantes principales constituent les variables statistiques de nouvelles informations acoustiques de haut niveau.
3. La représentation compressée est alors transformée (post-normalisation) par des procédures qui ne sont actuellement ni linéaires ni issues de fonctions probabilistes (la standardisation suivant une variabilité-cible est suivie d'une uniformisation des longueurs). Ces procédures s'avèrent former la solution la plus efficace pour mettre les données à conformité de modèles et scorings à hypothèses probabilistes gaussiennes.
4. Ces procédures sont indissociables. On ne peut parler d'efficacité de la solution i-vector si l'un de ces principes : (1) vecteur de représentation par concaténation multi-gaussienne, (2) compression suivant la variabilité totale, (3) normalisation non-linéaire, n'est pas respecté.

5. Ceci en gardant à l'esprit le nombre de classes-locuteurs d'apprentissage, de 1000 à 2000 actuellement, utilisés pour estimer la variabilité inter-locuteurs.

Le paradigme i-vector, défini par ces étapes fondamentales, est une **représentation compacte et normalisée des énoncés complets** et non de facteurs ciblés de variabilité, issue d'une **concaténation vectorielle d'informations locales du GMM-UBM**. Le résultat est un vecteur de variables acoustiques de très haut niveau informatif. Ces supervariables quantifient les causes indépendantes de variabilité du signal de voix, dont celles, propres au locuteur, qui nous intéressent. D'autre part, elles obéissent suffisamment à des hypothèses probabilistes pour leur appliquer des métriques robustes. Les représentations vectorielles des énoncés de voix à partir de ces supervariables répondent ainsi efficacement aux problématiques de la reconnaissance automatique du locuteur.

Chapitre 6

Limites de la modélisation linéaire gaussienne

6.1 Introduction

La solution i-vectors, que nous généralisons et identifions aux étapes fondamentales énoncées au chapitre précédent [1], est en mesure de gérer avec une grande efficacité des tâches de reconnaissance de locuteur. Encore récente, cette solution fait l'objet d'un nombre considérable d'études, destinées autant à étudier sa mise en oeuvre dans des contextes particuliers, voire extérieurs à la reconnaissance du locuteur (reconnaissance du langage) qu'à en affiner les étapes ou en simplifier le fonctionnement, essentiellement en terme de complexité de calcul.

Si une large partie de ce document a été consacrée à relever et circonscrire des propriétés de cette approche, la poursuite de son analyse nous paraît une priorité, étant donné le caractère encore neuf -et en partie exploratoire- de ce nouveau paradigme. En particulier, ses limites doivent être identifiées.

Lors du chapitre consacré à notre nouvelle approche de normalisation des i-vectors, *Spherical Nuisance* (au paragraphe 4.6.1), nous avions signalé la difficulté à estimer un métaparamètre linéaire universel de variabilité intra-locuteur lorsque les données avaient été transportées sur la surface d'une hypersphère pour gaussianisation. Cette surface est non-linéaire, ôtant toute validité a priori à une hypothèse d'homoscédasticité. Nous développons sur ce point dans cette section : l'utilisation de transformations non-linéaires durant la chaîne de traitements des vecteurs méritait qu'une mesure plus précise d'adéquation des métaparamètres linéaires aux hypothèses finales de la modélisation soit effectuée. Nous proposons donc une analyse supplémentaire qui vise à éprouver les limites du concept.

1. quelle que soit la représentation par concaténation issue du GMM-UBM, ainsi que la méthode de réduction de dimensionnalité.

6.2 Anisotropie dans l'espace des i-vectors normalisés

6.2.1 Présentation

Un système de reconnaissance du locuteur basé sur les i-vectors exécute les étapes-clés que nous avons énumérées dans le chapitre précédent :

- représentation issue du GMM-UBM par concaténation vectorielle d'informations par gaussiennes,
- compression sans prise en compte d'aucune variable latente,
- normalisation par standardisation suivant une variabilité puis uniformisation des longueurs
- modélisation et scoring par des approches usuelles à hypothèses gaussiennes.

Observant cette chaîne de traitement, la partie la plus fragile nous a semblé être celle d'estimation des métaparamètres globaux de variabilité explicative et résiduelle des modèles génératifs de i-vectors : les modèles LDA-two-covariance comme PLDA (Gaussienne ou Heavy-tailed) font l'hypothèse, peu citée ou rappelée, que les matrices de covariance ($\mathbf{B}$ et $\mathbf{W}$ pour LDA-two-covariance, $\mathbf{\Phi}$ et $\mathbf{\Gamma}$ pour la PLDA) entraînées sur de grands jeux d'apprentissage multi-locuteur et multi-sessions, sont applicables partout dans l'espace. Cette hypothèse implique que la classe des vecteurs d'un locuteur donné, où qu'elle soit localisée, suit la même loi avec les mêmes paramètres. Nous parlerons alors d'*isotropie* pour cette invariance des métaparamètres dans l'espace.

L'*homoscédasticité* n'entend pas que le métaparamètre global soit forcément exact partout, mais qu'il constitue en n'importe quel point un estimateur statistique robuste de la variabilité locale. Sous cette hypothèse, les écarts entre métaparamètre et paramètres empiriques locaux (estimés depuis des données d'apprentissage) sont seulement attribuables à la fluctuation d'échantillonnage. Les transformations à image sphérique que nous avons utilisées pour normaliser les i-vectors rendent peu crédible l'hypothèse d'invariance de tels métaparamètres, ceux-ci étant calculés par moyenne d'axes linéaires. Nous avons présenté lors de la justification de l'algorithme de normalisation *Spherical Nuisance* (4.6.2) le problème de l'estimation linéaire d'une matrice de covariance intra-locuteur quand les données s'étendent sur une surface sphérique. La prise en compte d'une *anisotropie* de l'espace, c'est à dire de modulation des métaparamètres suivant la localisation, amène à envisager des modèles génératifs qui s'appuient sur la non-linéarité de l'espace des observations :

- basés par exemple sur des distributions radiales (loi gaussienne radiale de Von Moses par exemple),

- envoyant les données dans un nouvel espace dont les variabilités seraient linéaires (par une fonction non-linéaire telle qu'un noyau de Fisher employé par Vapnik dans la méthodologie des SVM),

- évitant le problème d'estimations paramétriques linéaires par une cartographie (*Laplacian Eigen Maps* (Belkin and Niyogi, 2003), *Similarity Embedding* (Lee and Verleysen, 2009), *ISOMAP* (Tenenbaum et al., 2002)) ou une analyse en composantes indé-

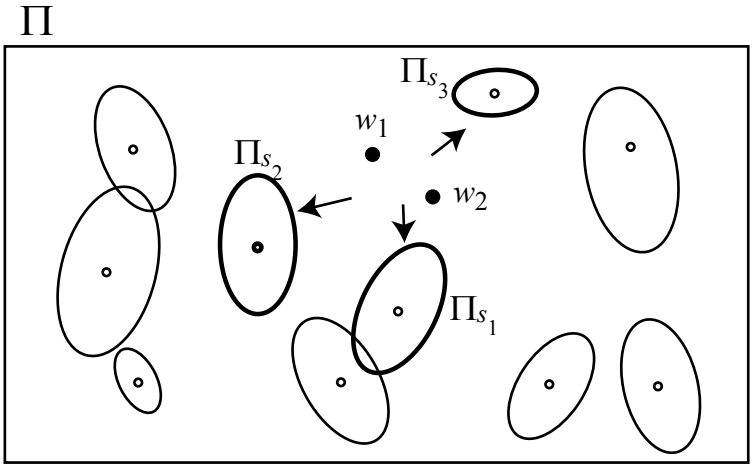

FIGURE 6.1 – *Adaptation à posteriori d'un métaparamètre global aux paramètres locaux.*

pendantes (*Independant Component Analysis ICA* (Hyvärinen et al., 2001), capables de modéliser des classes non-linéairement réparties, telles celles du défi de la variété *swiss roll* [2] (Weinberger and Saul, 2006).

Les investigations que nous avons menées sur ces méthodes n'ont pas abouti à l'heure actuelle à un résultat satisfaisant, ni même encourageant. Pour pallier la carence de modèle adapté, il est possible de traiter l'anisotropie par une prise en compte de l'information locale : à un métaparamètre global calculé sur la totalité d'un jeu d'apprentissage, est substitué une famille de métaparamètres locaux. On s'approche de méthodes telles que la HLDA (Kumar and Andreou, 1998) qui tient compte d'une hétéroscédasticité, l'hypothèse d'égalité des covariances par classes n'étant plus avancée. Mais le problème de vérification du locuteur est un problème inférentiel : à partir de classes d'apprentissage, doivent être évalués des paramètres de classes inconnues (celles des nouvelles observations à évaluer). Les méthodes hétéroscédastiques s'avèrent donc insuffisantes.

Les tentatives intuitives d'adaptation peuvent consister, par exemple, à déterminer dans le jeu de classes d'apprentissage à disposition les plus proches voisines des nouvelles observations à évaluer et : (i) soit de leur appliquer la moyenne des paramètres de ces classes voisines, (ii) soit d'adapter à posteriori un métaparamètre global à ces paramètres locaux.

La figure 6.1 décrit ces deux cas : un métaparamètre global $\mathbf{\Pi}$ a été estimé ainsi que des métaparamètres locaux $\mathbf{\Pi_s}$ pour chaque locuteur d'apprentissage s. A partir des classes plus proches voisines de deux nouvelles observations w_1 et w_2 à comparer (dans la figure les classes-locuteur s_1, s_2, s_3), l'estimation $\widehat{\mathbf{\Pi}}$ du métaparamètre à adopter pour w_1 et w_2 peut être obtenue suivant deux stratégies :

2. http ://www.convexoptimization.com/dattorro/manifold_learning.html

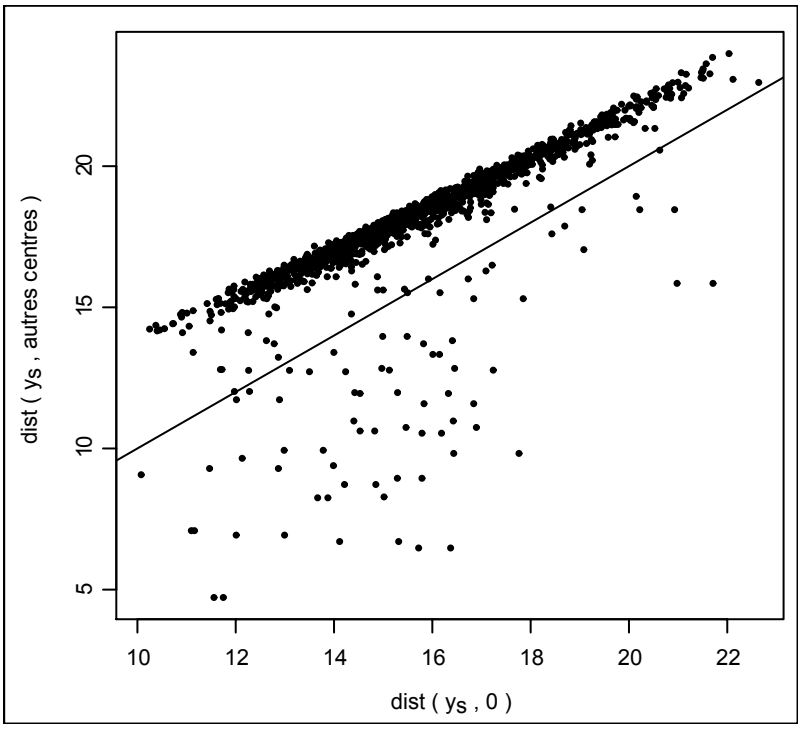

FIGURE 6.2 – *L'étrangeté des espaces de grande dimension. Distances à l'origine (moyenne globale des données) et au plus proche voisin des centres de classes-locuteurs.*

$$\left|\begin{array}{l} \text{cas } (i) : \widehat{\mathbf{\Pi}} = \mathbf{E}_{kNN}\left[\mathbf{\Pi_s}\right] \\ \text{cas } (ii) : \widehat{\mathbf{\Pi}} = \alpha\mathbf{\Pi} + (1-\alpha)\,\mathbf{E}_{kNN}\left[\mathbf{\Pi_s}\right] \end{array}\right.$$

où $\mathbf{E}_{kNN}\left[\mathbf{\Pi_s}\right]$ est une combinaison à définir des métaparamètres des plus proches voisins et α un paramètre d'adaptation. La stratégie (i) ne se base que sur les estimations locales : elle présente l'avantage de mieux estimer les caractéristiques locales, mais l'inconvénient de s'appuyer au final sur une faible quantité d'informations. La stratégie (ii) relativise le métaparamètre global $\mathbf{\Pi}$ à l'anisotropie locale. Mais dans les deux cas, l'estimation $\widehat{\mathbf{\Pi}}$ suppose de disposer de "plus proches voisins". Or, la notion de *voisinage* s'avère délicate à mettre en place sur les données de reconnaissance vocale.

La figure 6.2 affiche un graphique qui rappelle l'étrangeté des propriétés des espaces de grande dimension et leur faculté à résister à l'intuition. Pour élaborer ce graphique, nous avons calculé les vecteurs centres de classes y_s des 1444 locuteurs du jeu de données d'apprentissage BUT-hommes [3], après deux itérations de l'algorithme EFR. Pour chaque classe s, le centre y_s est sa moyenne $\dfrac{1}{n_s}\displaystyle\sum_{w\in s} w$, où n_s est l'effectif de la classe s. Puis nous avons affiché :

– en abscisse, les normes $\|y_s\|$, qui sont les distances à l'origine (donc à la moyenne globale μ des données),

3. Le fichier compte 1575 locuteurs mais seules les classes à deux observations au moins sont affichées.

– en ordonnée, la distance minimale de y_s aux centres des autres classes, donc la distance à la classe la plus proche au sens de la distance euclidienne entre les centroïdes. Nous rappelons que les i-vectors s'étendent sur une surface sphérique de rayon $\sqrt{p} = \sqrt{600} = 24.49$ centrée en 0 et se trouvent donc à $\sqrt{p}$ de l'origine.

La bissectrice $y = x$ nous permet d'apprécier les deux grandeurs : les points situés **en-dessous** cette droite correspondent à des classes pour lesquelles il existe une classe-locuteur voisine plus proche de s que s ne l'est de l'origine. Seules 73 classes locuteur sur 1444 sont dans ce cas. Les classes situées **au-dessus** de cette droite n'ont pas de classe voisine plus proche d'elles que de l'origine. Par exemple, la classe de coordonnées $(16.49, 19.05)$ sur la figure 6.2 a un centre y_s situé à 16.49 de l'origine 0, mais pas de classes voisines dont le centre soit à moins de 19.05. Sur les 1444 classes locuteurs, 1371 se retrouvent dans un tel cas.

Le bilan de ce graphique est flagrant : dans l'ensemble, les classes sont plus proches du centre 0 que de toute autre classe. Le centre 0 est plus proche voisin de chacune de ces classes que de n'importe laquelle des 1443 classes restantes. Pour résumer, les i-vectors sont plus proches du centre de sphère que des i-vectors "imposteurs". Cela tient à la dimension élevée de l'espace en regard du nombre de classes. Ces classes d'apprentissage représentent un recouvrement très clairsemé de la surface sphérique. Ayant à comparer deux nouvelles observations, celles-ci se retrouveront dans une zone "déserte" de la surface : leurs plus proches voisins parmi les classes d'apprentissage en seront en général trop éloignées pour fournir des informations locales fiables.

La stratégie de prise en compte de l'anisotropie par apprentissage local se heurte à deux obstacles majeurs :

- si les fichiers d'apprentissage actuels peuvent contenir plus de 1000 locuteurs, le recouvrement de l'espace des i-vectors (dimensions habituelles 400 à 600) par les classes de ces locuteurs est trop "clairsemé" : les classes locuteur plus proches voisines des observations à comparer sont souvent trop éloignées de ces observations pour fournir une estimation fiable,

- si l'effectif total de sessions est considérable, celui de sessions par classe-locuteur varie dans des ordres de grandeur de 2 à 80, la plupart ne dépassant pas 20. Ces classes prises individuellement ne constituent que des estimations très pauvres de la variabilité locale.

Nous nous retrouvons confrontés ici à un problème qui n'est pas nouveau : vouloir remplacer une estimation globale, réductrice mais robuste (car calculée sur une grande masse d'informations) par des estimations locales souffrant d'un manque d'informations. Des expérimentations menées sur nos données (application des méthodes décrites sur la figure 6.1) n'ont fait que confirmer cet état de fait.

Par contre, la prise en compte de l'anisotropie d'un espace de représentation, si elle est avérée, est possible si une relation logique se dégage de l'analyse des données. Toute estimation locale doit différer de son estimateur global par un écart aléatoire. Si tel n'est pas le cas, cela signifie qu'existe une contrainte entre la localisation dans l'espace et la valeur adaptée du paramètre. Reste alors à exhiber une expression mathématique de

cette relation. La chose n'est malheureusement pas aisée, la fonction prenant souvent une allure non-linéaire. Nous proposons dans cette section de montrer l'anisotropie de l'espace des i-vectors vis à vis de ses métaparamètres de variabilité explicative et résiduelle, puis d'ouvrir quelques perspectives déduites du caractère non-aléatoire de cette anisotropie.

6.2.2 Mesure de l'anisotropie

Soit $\mathcal{X}$ un jeu de i-vectors d'apprentissage et ses métaparamètres matriciels **B** et **W** de covariance inter- et intra-locuteur.

Le manque de fiabilité d'un métaparamètre matriciel tel que **W** peut tenir à trois causes : une mauvaise qualité d'estimation de son volume, de sa forme ou de son orientation. **W** étant la moyenne pondérée (par l'effectif n_s) des matrices de covariance $\mathbf{W}_s$ de chaque classe-locuteur s, nous étudions ces caractéristiques par le biais de trois séries :

Orientation : on construit sur l'ensemble des classes locuteurs s d'apprentissage la série :

$$\cos\left(\mathbf{W}, \mathbf{W}_s\right) = \frac{Tr\left(\mathbf{W}^t \mathbf{W}_s\right)}{\sqrt{Tr\left(\mathbf{W}^t \mathbf{W}\right)}\sqrt{Tr\left(\mathbf{W}_s^t \mathbf{W}_s\right)}} \tag{6.1}$$

qui compare les axes principaux des matrices de covariance locales à la globale **W**.

Une valeur maximale de 1 indique une coïncidence parfaite des axes principaux des deux matrices.

Forme : elle peut être évaluée par l'énergie du spectre (pente de la série des valeurs propres). On construit la série sur l'ensemble des classes locuteurs s d'apprentissage :

$$\log\det\left(\frac{\mathbf{W}_s}{\sqrt{Tr\left(\mathbf{W}_s^t \mathbf{W}_s\right)}}\right) - \log\det\left(\frac{\mathbf{W}}{\sqrt{Tr\left(\mathbf{W}^t \mathbf{W}\right)}}\right) \tag{6.2}$$

En effet, on a, en notant $\left\{\lambda_j\right\}_j$ le spectre des valeurs propres d'une matrice $p \times p$ définie positive **M** :

$$\log\det\left(\frac{\mathbf{M}}{\sqrt{Tr\left(\mathbf{M}^t \mathbf{M}\right)}}\right) = \sum_j \log\frac{\lambda_j}{\sum_k \lambda_k} \tag{6.3}$$

qui est maximal et égal à à $-p\log p$, lorsque ce spectre est plat.

Volume : on construit la série sur l'ensemble des classes locuteurs s d'apprentissage :

$$\frac{Tr\left(\mathbf{W}_s\right)}{Tr\left(\mathbf{W}\right)} \tag{6.4}$$

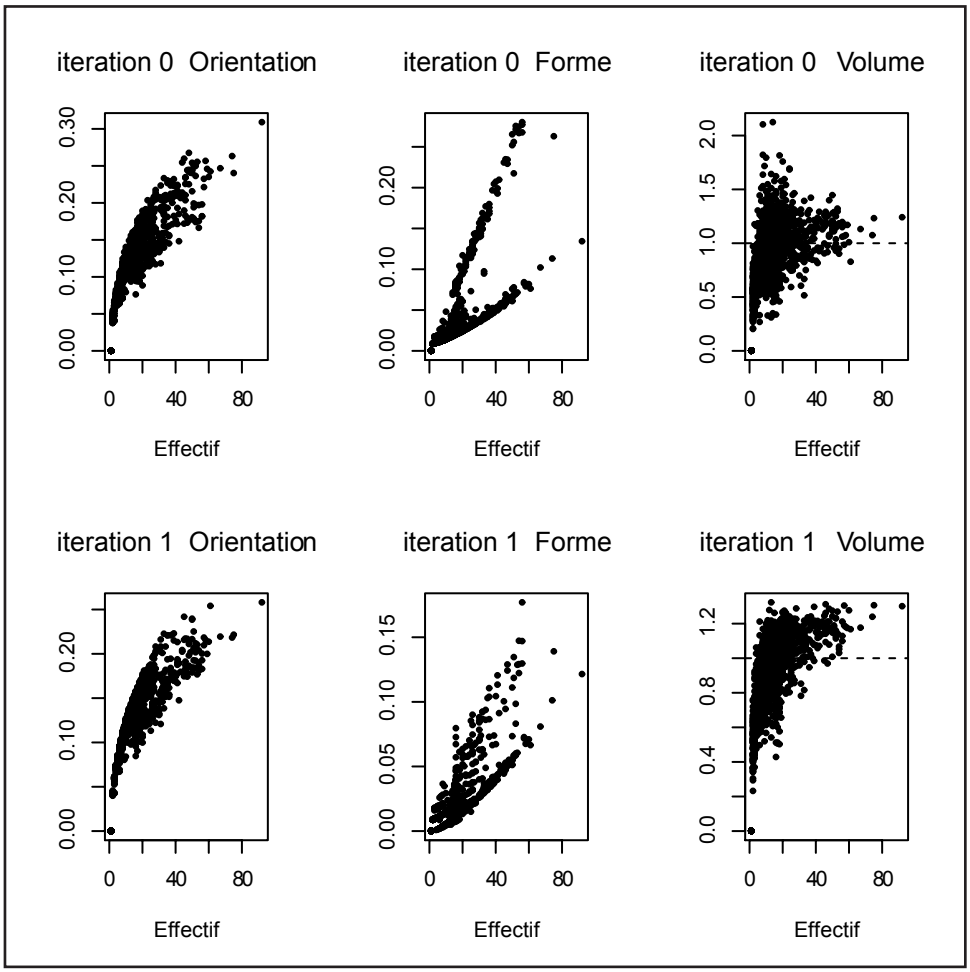

FIGURE 6.3 – *Anisotropie du métaparamètre* **W** *avant et après normalisation EFR, en termes d'orientation, forme et volume.*

qui est le ratio des indicateurs de volume de covariance, entre celles d'une classe s et du métaparamètre.

La figure 6.3 permet d'étudier l'anisotropie du métaparamètre **W** avant et après normalisation EFR, pour le jeu BUT-hommes. Sur chaque graphique, l'abscisse est l'effectif de la classe-locuteur d'apprentissage. Les ordonnées successives sont les indicateurs d'orientation, forme et volume. La première ligne de graphiques correspond aux données initiales, telles que produites par l'extracteur d'i-vectors. La seconde correspond aux données après 1 itération de l'algorithme EFR. Il est à noter dans toute la suite que, d'une part, les graphiques n'évoluent plus -ou de manière négligeable- aux itérations suivantes, d'autre part que les mêmes allures de graphique se retrouvent sur les autres jeux de données testés.

Orientation : l'indicateur étant un cosinus, sa valeur est dans $[-1, 1]$. On voit qu'avant comme après normalisation, l'orientation du métaparamètre **W** coïncide le mieux avec

celles des classes de plus grand effectif. Ce fait est normal, celles-ci participant avec le plus d'intensité à la construction des axes principaux. Notons que les faibles valeurs des cosinus (borne supérieure à 0.25 pour un maximum de 1) tiennent aux faibles effectifs, donc d'axes principaux, en regard de la dimension $p = 600$.

Forme : de la même manière, avant comme après normalisation, la coïncidence de forme entre le métaparamètre et les paramètres locaux est fonction -approximative- de l'effectif des classes, donc de leur richesse en terme d'informations. La conclusion est donc la même que précédemment.

Volume : avant normalisation, le ratio de volume tend vers 1 en fonction de l'effectif classe. On aboutit une nouvelle fois à la même conclusion. Mais après normalisation, on peut voir (graphique en bas à droite) que le ratio de volume croît avec l'effectif. Cela signifie que le volume des classes les mieux renseignées (quantité d'information maximale) est sous-estimé par le métaparamètre **W**. Ce fait a attiré notre attention : lorsque l'effectif de classe augmente, la dispersion empirique ne tend pas nécessairement à s'accroître : celle-ci, mesurée par la trace qui n'est autre que la variance de la classe, est proportionnelle à la moyenne -et non à la somme- des distances mutuelles entre ces éléments. En ce sens, une relation entre le nombre de points d'une classe et sa variance ne doit pas être constatée lors de l'analyse de données empiriques.

Nous poursuivons cette étude par l'analyse des vraisemblances des métaparamètres **B** et **W**. Considérons deux mesures de vraisemblances de **B** et **W** pour la classe d'un locuteur s de $\mathcal{X}$:

$$\mathcal{L}_{\mathbf{B}}(s) = \log\left(\left(\prod_{w\in s}\mathcal{N}(y_s|\mu,\mathbf{B})\right)^{\frac{1}{n_s}}\right) = \log\left(\mathcal{N}(y_s|\mu,\mathbf{B})\right) \tag{6.5}$$

et

$$\mathcal{L}_{\mathbf{W}}(s) = \log\left(\left(\prod_{w\in s}\mathcal{N}(w|y_s,\mathbf{W})\right)^{\frac{1}{n_s}}\right) \tag{6.6}$$

La première est la log-vraisemblance de la moyenne y_s des sessions de s, sous l'hypothèse de normalité des moyennes-locuteur suivant une loi de moyenne μ et de covariance **B**. La seconde est la log-vraisemblance de l'échantillon des sessions de s, d'effectif n_s, sous l'hypothèse de leur normalité suivant une moyenne y_s et une covariance **W**.

Afin de comparer seulement ses mesures entre locuteurs, on ignore le terme constant et les vraisemblances se réécrivent :

$$\mathcal{L}_{\mathbf{B}}(s) = -\frac{1}{2}(y_s-\mu)^t\,\mathbf{B}^{-1}(y_s-\mu) \tag{6.7}$$

$$\mathcal{L}_{\mathbf{W}}(s) = -\frac{1}{2}\frac{1}{n_s}\sum_{w\in s}(w-y_s)^t\,\mathbf{W}^{-1}(w-y_s) \tag{6.8}$$

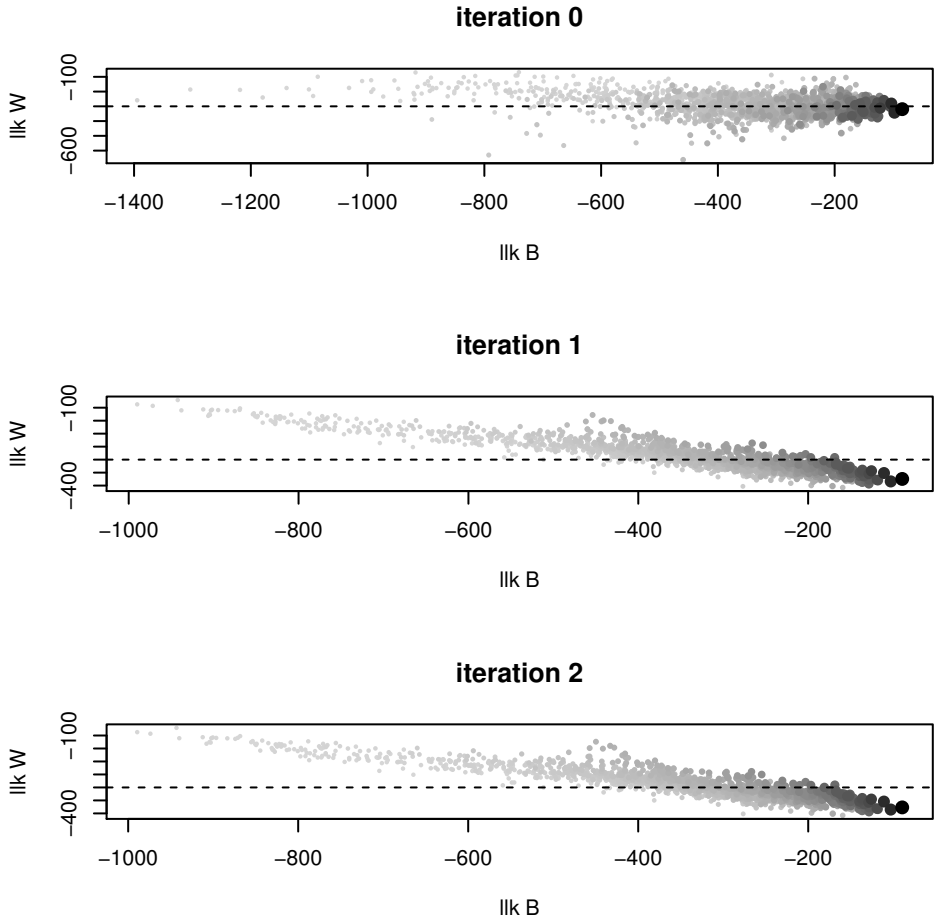

FIGURE 6.4 – *Vraisemblances $\mathcal{L}_{\mathbf{B}}(s)$ et $\mathcal{L}_{\mathbf{W}}(s)$ des métaparamètres* **B** *et* **W** *d'un ensemble de classes-locuteurs s, initialement puis après* 1 *et* 2 *itérations de la normalisation EFR.*

$\mathcal{L}_W(s)$ peut s'écrire plus simplement :

$$\mathcal{L}_{\mathbf{W}}(s) = -\frac{1}{2} Tr\left(\mathbf{W}^{-1} \frac{1}{n_s} \sum_{w \in s} (w - y_s)(w - y_s)^t\right) = -\frac{1}{2} Tr\left(\mathbf{W}^{-1} \mathbf{W}_s\right) \tag{6.9}$$

Cette formulation rappelle que $\mathcal{L}_{\mathbf{W}}(s)$ mesure l'écart entre l'estimation globale **W** et la locale $\mathbf{W}_s$.

La figure 6.4 affiche pour tous les locuteurs du jeu de données d'apprentissage leurs vraisemblances $\mathcal{L}_{\mathbf{B}}(s)$ en abscisse et $\mathcal{L}_{\mathbf{W}}(s)$ en ordonnée. En haut, elle est calculée sur les i-vectors initiaux (tels que produits par l'extracteur), puis après 1 puis 2 itérations de la normalisation EFR. La taille du point et son niveau de gris sont proportionnels à l'effectif dont on disposait pour ce locuteur. D'autre part, dans le cas idéal où la covariance des vecteurs d'apprentissage d'une classe s correspondrait exactement à **W**, la

vraisemblance $\mathcal{L}_{\mathbf{W}}(s)$ est égale à $-\frac{1}{2}Tr\left(\mathbf{W}^{-1}\mathbf{W}\right) = -\frac{p}{2}$. Cette valeur est indiquée par la ligne horizontale en pointillés.

Avant normalisation, il n'existe visiblement pas de relation entre les deux vraisemblances. Les variations peuvent être attribuées à la fluctuation d'échantillonnage, importante sur des classes de taille aussi réduite en regard de la dimension.

Dès la première itération de l'algorithme EFR, une relation linéaire forte apparaît entre les vraisemblances des données, qui s'étendent maintenant sur une surface sphérique. La relation se stabilise immédiatement, comme le montre la seconde itération. Le même phénomène a été observé sur les jeux de données BUT-femmes et LIA-hommes. Pour le jeu dont la figure 6.4 affiche les vraisemblances $\mathcal{L}_{\mathbf{B}}$ et $\mathcal{L}_{\mathbf{W}}$, une régression linéaire entre celles-ci aboutit à l'approximation :

$$\mathcal{L}_{\mathbf{W}} \approx \beta_1 \mathcal{L}_{\mathbf{B}} + \beta_0 \tag{6.10}$$

où $\beta_1 = -0.312$ et $\beta_0 = -397.18$, avec un coefficient de détermination $R^2 = 0.932$ proche de 1 indiquant une excellente qualité d'approximation linéaire.

Cette relation entre $\mathcal{L}_{\mathbf{B}}$ et $\mathcal{L}_{\mathbf{W}}$ induit celle entre les probabilités suivantes de chaque classe :

$$E_{w \in s}\left[P\left(w|y_s, \mathbf{W}\right)\right] \approx e^{\beta_0} P\left(y_s|\mathbf{B}\right)^{\beta_1} \tag{6.11}$$

où β_0 est une constante, $P\left(y_s|\mathbf{B}\right) = \mathcal{N}\left(y_s|\mu, \mathbf{B}\right)$ est la probabilité de la classe s à posteriori de $\mathbf{B}$ et

$$E_{w \in s}\left[P\left(w|y_s, \mathbf{W}\right)\right] = \left(\prod_{w \in s} \mathcal{N}\left(w|y_s, \mathbf{W}\right)\right)^{\frac{1}{n_s}} \tag{6.12}$$

est la moyenne géométrique des probabilités des observations w de s à posteriori de $\mathbf{W}$.

Cette relation, constatée systématiquement sur des jeux d'apprentissage issus de différentes paramétrisations et extractions en amont, amène une remarque : après normalisation, les estimations globales de covariance par des métaparamètres linéaires ne sont pas adaptées. Les quantités importantes de sessions utilisées pour leur apprentissage expliquent leur efficacité globale, mais ignorent en partie la richesse de l'information locale. La relation entre vraisemblance des sessions d'une classe et position locale (ici position de leur centre vis à vis des ellipsoïdes de distribution gaussienne interlocuteur) interdit d'attribuer ces spécificités à la fluctuation aléatoire d'échantillonnage.

L'inquiétude principale à la vue de ce graphique tient au fait que les classes de plus faible vraisemblance $\mathcal{L}_{\mathbf{W}}$ (donc d'estimation de la covariance intra-classe par $\mathbf{W}$ la plus médiocre) sont celles de plus grande vraisemblance de leur centre ($\mathcal{L}_{\mathbf{B}}$) et souvent de plus grand effectif.

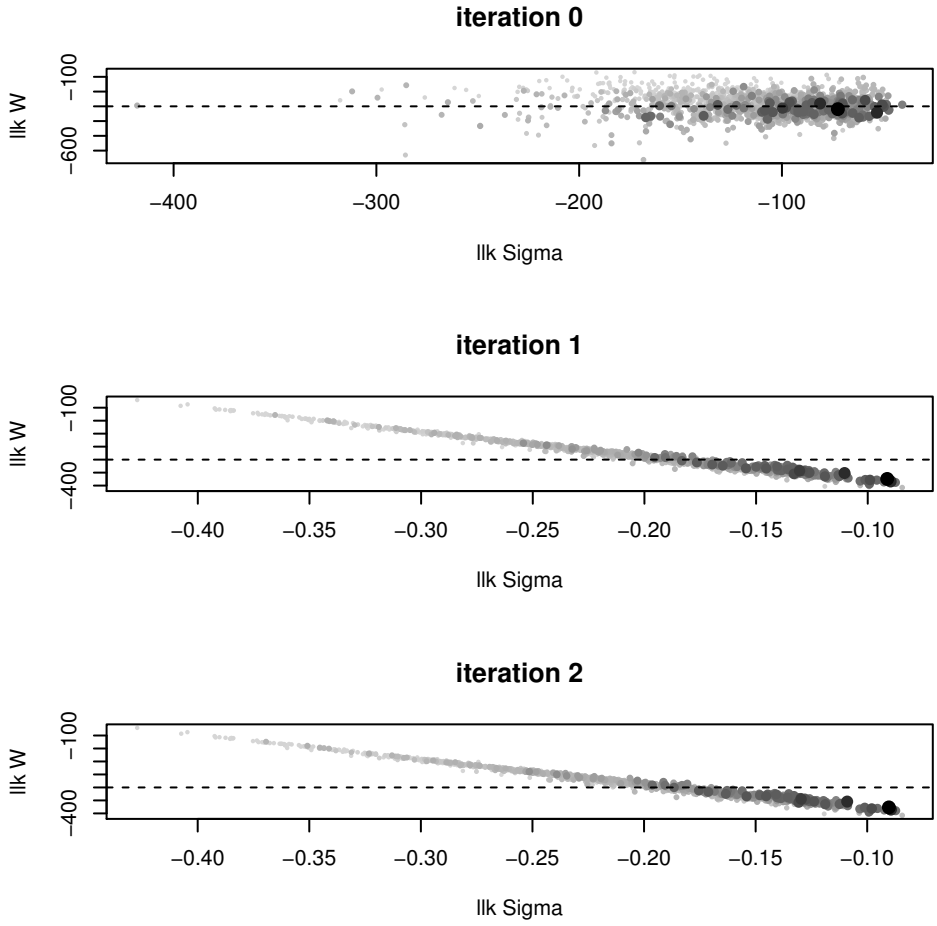

FIGURE 6.5 – *Vraisemblances - en abscisse $\mathcal{L}_{\mathbf{\Sigma}}(s)$ du centre de classe y_s en regard de la variabilité totale, - en ordonnée $\mathcal{L}_{\mathbf{W}}(s)$ du métaparamètre* $\mathbf{W}$*, initialement puis après* 1 *et* 2 *itérations de la normalisation EFR.*

Ce phénomène était prévisible, au vu de l'analyse précédente (figure 6.3 : le défaut d'estimation du volume de classe découlant d'une dilatation de la matrice $\mathbf{W}_s$ avec l'effectif).

La figure 6.5 affiche, pour les mêmes données, les vraisemblances de classe $\mathcal{L}_{\mathbf{W}}(s)$ en ordonnées, mais cette fois en abscisse la vraisemblance $\mathcal{L}_{\Sigma}(s)$ du centre de classe y_s en regard de la variabilité totale :

$$\mathcal{L}_{\mathbf{\Sigma}}(s) = \log\left(\mathcal{N}\left(y_s|\mu,\mathbf{\Sigma}\right)\right) \propto -\frac{1}{2}\left(y_s-\mu\right)^t \mathbf{\Sigma}^{-1}\left(y_s-\mu\right) \tag{6.13}$$

Utilisant la quasi-égalité entre $\mathbf{\Sigma}_i$ et $\mathbf{I}$, cette vraisemblance est remplacée en abscisse de la figure par son approximation (μ après EFR tend vers 0 et $\mathbf{\Sigma}$ vers $\mathbf{I}$) :

$$\widetilde{\mathcal{L}_{\mathbf{\Sigma}}}(s) = -\frac{1}{2}\|y_s\|^2 \tag{6.14}$$

La relation linéaire entre $\widetilde{\mathcal{L}_{\mathbf{\Sigma}}}(s)$ et $\mathcal{L}_{\mathbf{W}}(s)$ est plus flagrante encore. Le R^2 de la régression linéaire est de 0.995.

Ce résultat était en partie prévisible : l'indicateur de forme vérifie :

$$\begin{aligned} Tr(\mathbf{W}_s) &= \frac{1}{n_s}\sum_{w\in s}(w-y_s)^t(w-y_s) \\ &= \frac{1}{n_s}\sum_{w\in s}\|w\|^2 - \|y_s\|^2 \\ &= 1 - \|y_s\|^2 \\ &= 1 - 2\mathcal{L}_{\mathbf{\Sigma}}(s) \end{aligned} \tag{6.15}$$

ce qui induit une corrélation négative forte entre $\mathcal{L}_{\mathbf{\Sigma}}(s)$ et $\mathcal{L}_{\mathbf{W}}$.

Pour mieux cerner ce phénomène, nous l'avons étudié sur un fichier obtenu par simulation : nous générons un ensemble de classes d'effectifs aléatoires et suivant des lois normales de paramètres distincts.

La figure 6.6 affiche les points $(\mathcal{L}_{\mathbf{B}}, \mathcal{L}_{\mathbf{W}})$ de ces données d'apprentissage simulées, avant puis après les 1ère à 5ème itérations de l'algorithme EFR. Le même phénomène se produit : une corrélation linéaire négative forte entre les deux indicateurs de vraisemblance. Ici, les classes de grand effectif ne présentent pas de positionnement excentré, comme c'était le cas sur les jeux de données réelles.

Enfin, nous avons réalisé une analyse d'adéquation de données d'évaluation aux paramètres d'apprentissage. Pour cela, à partir du jeu de données de i-vectors de tests hommes de NIST-SRE 2008 (3798 exemples) produit par le LIA, un fichier multi-sessions a été constitué (3689 sessions de 397 locuteurs). Sur les i-vectors initiaux puis après chaque itération de l'algorithme EFR, ont été calculés les indicateurs d'orientation, forme et volume ainsi que les vraisemblances $\mathcal{L}_{\mathbf{B}}$ et $\mathcal{L}_{\mathbf{W}}$ de chaque classe de ce fichier, ce à partir des métaparamètres calculés sur le fichier d'apprentissage précédent. Les figures 6.7, 6.8 et 6.9 affichent les résultats de cette procédure d'analyse. On voit sur la figure 6.7 que le métaparamètre **W** issu de l'apprentissage coïncide avec les classes des tests en proportion de leur effectif, vis à vis de leur orientation et forme. Mais l'indicateur de volume pose toujours problème : ratio de volume croissant avec l'effectif, volume des classes les mieux renseignées sous-estimé par le métaparamètre **W**. Toutefois, le phénomène est déjà apparent avant normalisation (itération 0) et son accroissement est réel mais minime.

La figure 6.8 (resp. 6.9) qui affiche les couples de vraisemblance $\mathcal{L}_{\mathbf{B}}, \mathcal{L}_{\mathbf{W}}$ (resp. $\mathcal{L}_{\mathbf{\Sigma}}, \mathcal{L}_{\mathbf{W}}$) confirme, sur des données d'évaluation, le constat de l'apprentissage : de fortes relations linéaires existent entre les vraisemblances. Mais ici, les classes de plus grand

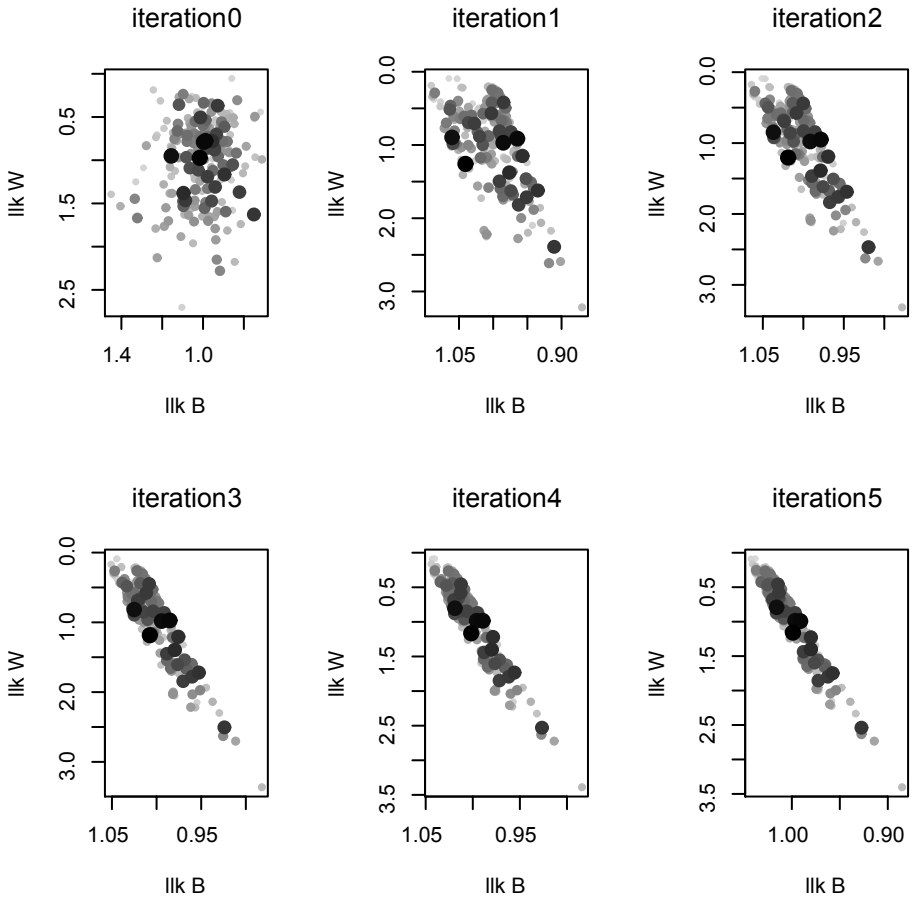

FIGURE 6.6 – *Points* $(\mathcal{L}_{\mathbf{B}}, \mathcal{L}_{\mathbf{W}})$ *de données d'apprentissage simulées, avant puis après les* $1^{ère}$ *à* $5^{ème}$ *itérations de l'algorithme EFR*

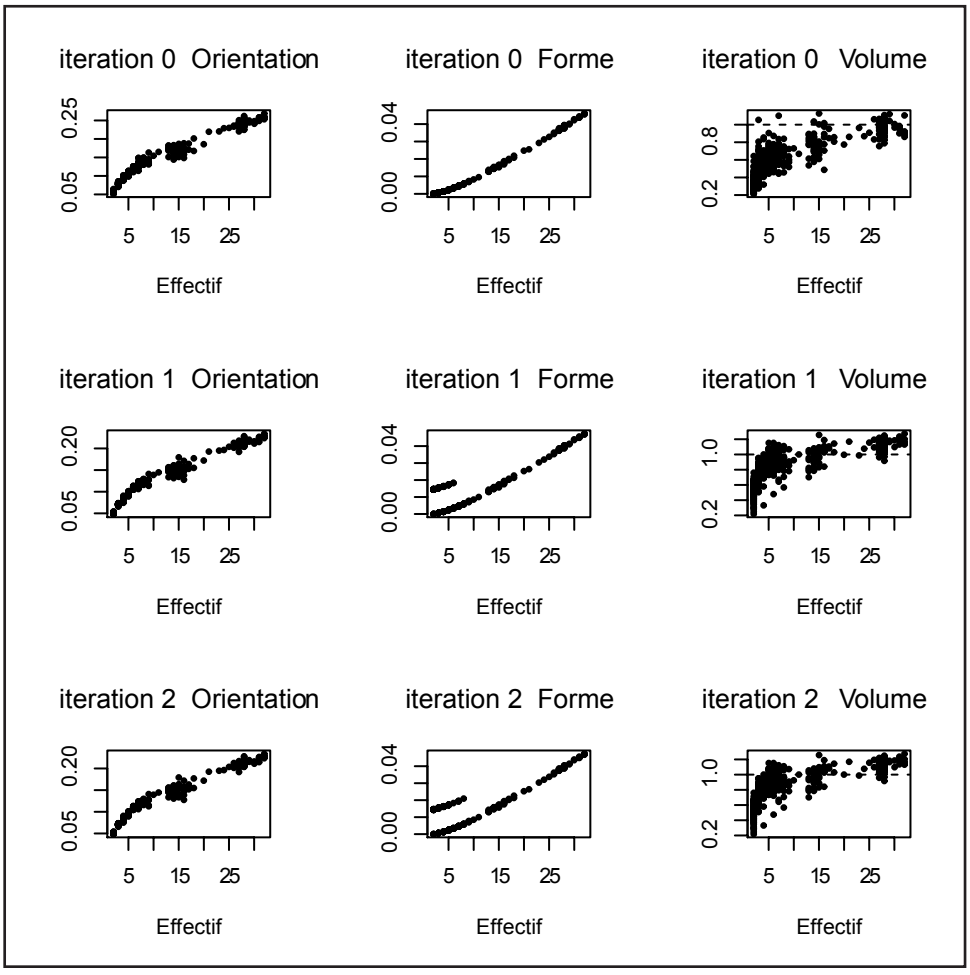

FIGURE 6.7 – *Adéquation de données d'évaluation au métaparamètre d'apprentissage* **W**.

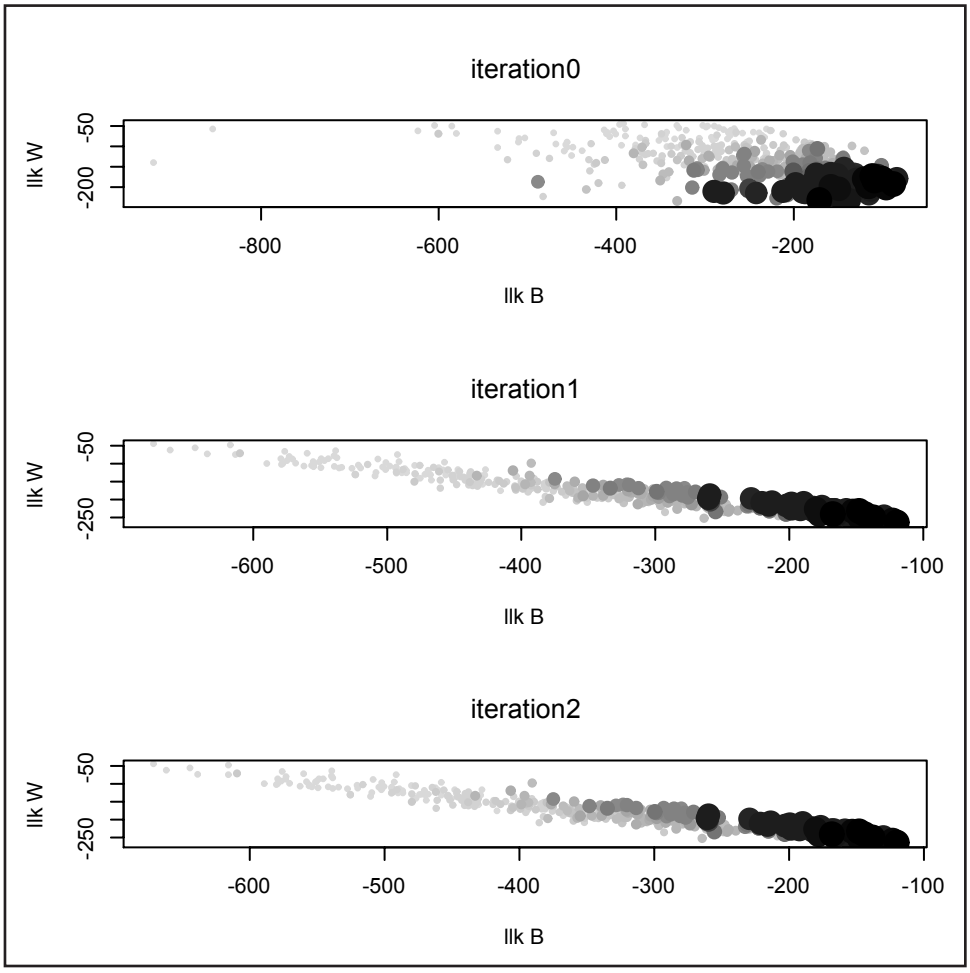

FIGURE 6.8 – *Couples de vraisemblance* $(\mathcal{L}_{\mathbf{B}}, \mathcal{L}_{\mathbf{W}})$ *sur des données d'évaluation,*

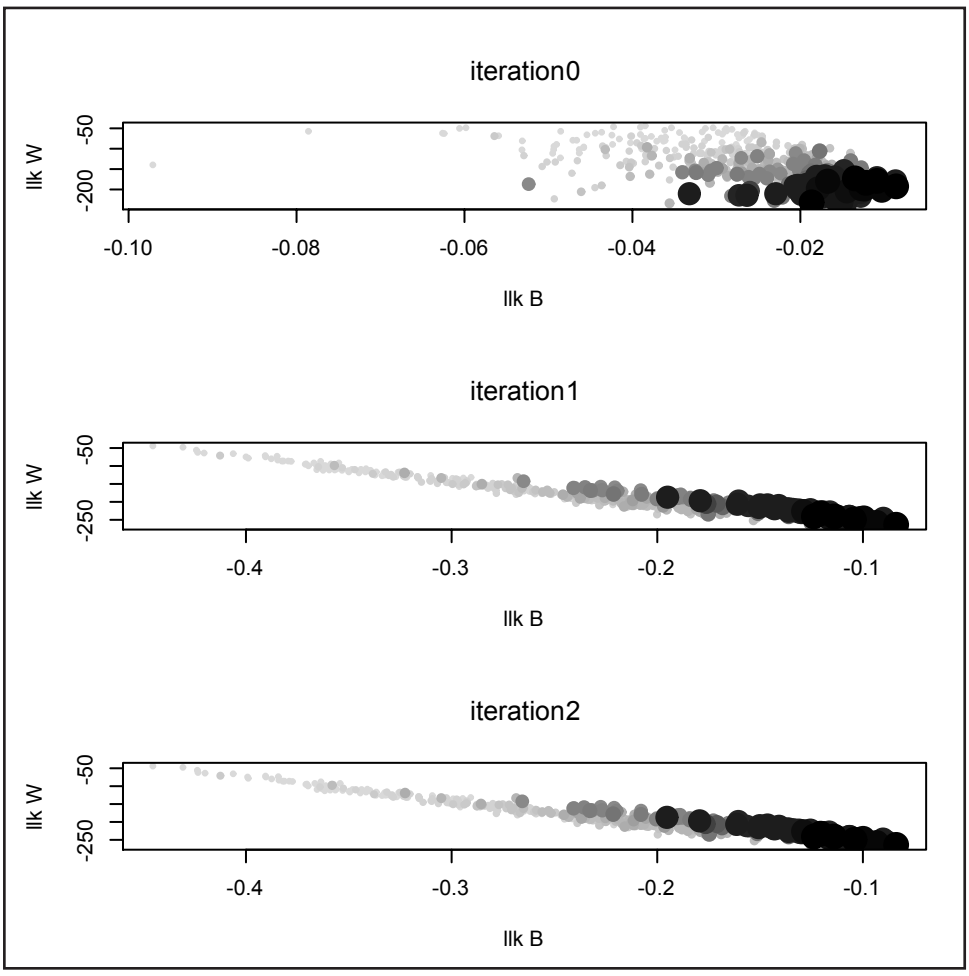

FIGURE 6.9 – *Couples de vraisemblance* $(\mathcal{L}_{\mathbf{\Sigma}}, \mathcal{L}_{\mathbf{W}})$ *sur des données d'évaluation.*

effectif possèdent déjà avant itération les vraisemblances intra-classes $\mathcal{L}_{\mathbf{W}}$ les plus réduites, sans toutefois qu'elles soient corrélées linéairement aux autres vraisemblances.

6.3 Conclusion

Pour résumer l'ensemble de ces observations, l'utilisation des métaparamètres linéaires globaux après sphérisation laisse apparaître une anisotropie de l'espace de représentation, qui est une surface sphérique. Le métaparamètre matriciel $\mathbf{W}$ n'est pas dégradé par l'opération de normalisation en terme d'orientation et de forme, ce qui signifie seulement qu'il n'est pas pire a priori qu'avant normalisation, mais il apparaît après celle-ci une difficulté dans l'estimation de son volume. Surtout, la variabilité de la forme n'est pas aléatoire, présentant une corrélation forte avec la position du centre de classe.

Les formules de scoring des modèles LDA-two-covariance et PLDA calculent en fait une vraisemblance de l'hypothèse "même locuteur" par accumulation de vraisemblances à posteriori d'une position de centre de classe présumée. Cette position est variée suivant la distribution présumée des classes-locuteur. Si, à une position de centre de classe, ne correspond pas une unique distribution intra-classe (à un facteur négligeable près), l'estimation de l'hypothèse sera nécessairement imprécise. Ce caractère non aléatoire est donc préoccupant et implique qu'une prise en compte de cette anisotropie doit être envisagée. L'existence d'une relation mathématique quasi-exacte sur celle-ci présente l'avantage de réduire le nombre de degrés de liberté d'une procédure d'adaptation locale, qui serait impraticable ou trop imprécise sans connaissances a priori.

Nous présentons en annexe H une tentative de prise en compte de l'anisotropie que nous venons de révéler. Il s'agit d'une adaptation du modèle two-covariance à des hypothèses sous-jacentes (des relations mathématiques explicites entre ces métaparamètres et la position dans l'espace de représentation). Ces travaux ont abouti à des résultats intéressants mais encore peu significatifs en terme de performance.

La sphérisation des données est aujourd'hui effectuée dans la très grande majorité des systèmes basés sur les i-vectors. Il s'agit donc de poursuivre la tâche de blanchiment (*whitening*) pour mise à conformité des données, mais en prenant garde que les transformations actuellement les plus efficaces portent en elles leurs propres limites : leurs images sont non-linéaires, limitant l'isotropie et donc la validité des modélisations linéaires en aval.

Les chapitre suivants présentent diverses contributions : au modèle de clés binaires, puis l'application de la solution i-vectors à des contextes de reconnaissance particuliers.

Chapitre 7

Conclusion et perspectives

La majorité des systèmes de reconnaissance du locuteur s'appuient depuis plus d'une dizaine d'années sur l'approche probabiliste par mixtures de gaussiennes. Cette approche est basée sur la modélisation probabiliste de l'espace acoustique par un modèle unique (GMM-UBM) et la représentation sous forme vectorielle des énoncés de voix, les supervecteurs, constitués par les statistiques adaptées du GMM-UBM. Ces objets vectoriels sont à la source de nombreux travaux destinés à séparer les facteurs de variabilité, intrinsèque ou extrinsèque au locuteur et à mesurer la proximité entre des énoncés vis à vis de la variable locuteur.

Dans ce cadre de recherche, des avancées récentes ont révélé le pouvoir discriminant d'une nouvelle représentation vectorielle des énoncés vocaux, les i-vectors. Cette représentation extrait de la précédente un seul facteur de variabilité totale, de dimension réduite (400 à 600). Elle s'avère plus performante que les solutions précédentes issues du GMM-UBM dans les tâches de reconnaissance.

Contributions

Nous avons consacré cette thèse à l'analyse de ce nouveau paradigme, à la présentation de solutions pour en améliorer la qualité et l'efficacité, à la justification de ces solutions et, enfin, à dégager les concepts-clés qui assurent la qualité et la fiabilité de cette approche, par évaluation des composantes du système. Certaines limites théoriques du concept ont été également mises au jour.

Plusieurs de ces travaux se sont appuyés sur le modèle de clés binaires du locuteur (Bonastre et al., 2011b) (Bonastre et al., 2011a), utilisé ici pour l'évaluation des composantes du système. Nous avons également présenté dans ce document quelques unes de nos contributions dans ce domaine. D'autre part, l'étude de fiabilité du concept de i-vector nous a amenés à étudier son application dans des contextes de reconnaissance particuliers.

Le concept i-vector

Notre premier travail a consisté en une analyse statistique et spatiale de cette nouvelle représentation, pour en expliquer les capacités et en améliorer l'efficacité. Le concept i-vector nécessitait en effet un approfondissement de ses propriétés théoriques, en particulier vis à vis de la variable-cible locuteur. Les points suivants ont attiré notre attention, motivant et guidant l'analyse que nous avons menée :

- le scoring proposé par les initiateurs du concept est un cosinus, donc un produit scalaire sur des vecteurs dont l'information de longueur a été ignorée. De plus, une forme de standardisation y est effectuée (WCCN) qui annule des écarts de variance suivant une métrique intra-locuteur. Ces faits sont surprenants : les bonnes performances de ce scoring montrent qu'une part des informations contenues dans le i-vector et donc dans la représentation par supervecteurs d'origine gagne à être ignorée ou redressée. La perte ou la correction d'informations sur une représentation déjà réduite ne peut qu'interroger et susciter une enquête.
- l'application de décompositions en facteurs telles que la PLDA dans ce nouvel espace compact de représentation s'est heurtée à une non-conformité partielle des données au modèle théorique gaussien sous-jacent. Si la solution proposée (Kenny, 2010), basée sur des lois de Student en lieu et place des lois gaussiennes, a fourni un cadre probabiliste plus probant aux systèmes basés sur les i-vectors, la question des anomalies au postulat gaussien nous a paru mériter des analyses complémentaires.
- Ceci amène la question suivante : considérant les i-vectors comme les images d'une représentation issue du GMM-UBM, la question se pose de leur nature : peut-on, ignorant leur processus d'extraction, les considérer comme des observations issues d'un modèle génératif probabiliste ? Ou bien faut-il d'abord les transformer, pour redresser certaines anomalies et les rendre compatibles aux hypothèses probabilistes d'un modèle de décomposition gaussien ?

Les diverses caractéristiques statistiques et spatiales que nous avons étudiées ont montré certains défauts des i-vectors produits par la procédure FA-total-var, en terme de conformité aux hypothèses précédant ou suivant leur extraction. Ainsi, leur normalité et standardité théoriques ne sont pas acquises. En lieu et place d'une modélisation basée sur des postulats non-gaussiens, tels que ceux de la Heavy-tailed PLDA utilisant la loi de Student, des transformations sont envisageables pour corriger ces défauts, tels que la standardisation et la normalisation de longueur. La prise en compte des conclusions de cette analyse nous ont conduits à mettre en place une famille de transformations des i-vectors, préliminaire à la phase de modélisation et scoring.

Nous avons proposé deux techniques de normalisation, *Eigen Factors Radial* (*EFR*) et *Spherical Nuisance* (*SphN*) [1]. La seconde est spécifique à la modélisation PLDA-gaussienne. L'objectif fondamental de ces transformations est d'obtenir des vecteurs qui puissent constituer des observations d'un modèle génératif probabiliste. Nous avons montré

1. aussi notées dans certains articles $\mathbf{L\Sigma}$ et $\mathbf{LW}$ pour *length-normalization* et standardisation suivant les matrice de covariance totale $\mathbf{\Sigma}$ ou intra-locuteur $\mathbf{W}$.

comment il est possible de traiter les données préalablement pour fournir à des modèles des données plus proches de leurs hypothèses et ainsi mieux optimiser les critères qu'ils se sont fixés.

Les transformations proposées, sous forme de standardisations et normalisation de longueur itératives, montrent leur pertinence et leur efficacité dans le champ des i-vectors. Etapes préliminaires avant l'élaboration d'un modèle génératif dans l'espace des i-vectors, elles rapprochent les données des critères et hypothèses fixés par ces modèles. Leur particularité est de n'être ni linéaires, ni paramétriques. Leur opportunité est validée par le gain de performance qu'elles induisent et par la meilleure conformité aux modèles, garante du niveau de fiabilité d'un système. Les analyses réalisées au chapitre 3, reprises sur les i-vectors après application de ces diverses transformations au chapitre 4, justifient leur pertinence. D'autre part, un certain nombre de propriétés des vecteurs transformés sont démontrées, qui éclairent plus précisément le rôle de ces procédures dans la qualité de la modélisation : après transformation, les i-vectors tendent vers un modèle théorique, dans lequel les variabilités locuteur et nuisible sont plus facilement séparables :

- ou bien une base commune aux deux variabilités (*eigen factors*) optimise les critères de séparabilité et leur adéquation à un modèle déterministe,
- ou bien une base commune de voix propres (*eigenvoices*) à résidu isotropique initialise de manière robuste la décomposition probabiliste PLDA.

Dans les deux cas, la sphérisation de l'espace de représentation, qui devient une variété finie, distribue les vecteurs suivant une dispersion maximale. Avec la standardisation, elle contribue également à mettre en conformité de nouvelles observations de test à la distribution des données d'apprentissage, combattant ainsi des biais éventuels. L'EFR L$\mathbf{\Sigma}$ rend équivalentes et optimales les démarches déterministes LDA et NAP. La SphN L$\mathbf{W}$ permet, par l'initialisation des paramètres que nous proposons au chapitre 4, de rendre quasi-équivalent le modèle déterministe aux modèles génératifs obtenus par maximum de vraisemblance.

Sur le plan des performances, le gain moyen procuré par ces transformations dans le cadre de la PLDA gaussienne est de l'ordre de 50% en terme de diminution du taux d'erreur. Cette configuration permet d'atteindre, par la voie du cadre probabiliste gaussien, les meilleurs taux de détection dans le domaine. Le gain est de l'ordre de 35% par rapport à l'approche LDA-WCNN-cosine scoring de (Dehak et al., 2011).

Les techniques de normalisation (telles que WCCN-cosinus, EFR, SphN) apparaissent donc comme des facteurs déterminants de réussite de la solution i-vector. Elles visent à optimiser la qualité de la modélisation en lui fournissant des vecteurs répondant mieux à ses hypothèses gaussiennes. Les différentes étapes d'un système i-vector (représentations par supervecteurs, décomposition en facteurs par FA-total var, modèle génératif PLDA) respectent des contraintes de gaussianité et prennent donc également en compte cet objectif. Les gains conséquents de performance induits par les normalisations nous ont conduits à penser que la participation de ces mises en conformité des vecteurs était déterminante, voire prépondérante.

La part jouée par chacun des modules de la chaîne de traitement dans la qualité de la modélisation a donc fait l'objet d'une évaluation, destinée à mieux cerner leur impact dans la réussite de la démarche. Plus généralement, cette évaluation devait nous renseigner sur les constituants principaux de ces systèmes, mieux les maîtriser et ainsi mieux les cerner en terme de fiabilité. Son dépouillement nous a en effet permis d'isoler les facteurs-clés du concept : des étapes fondamentales, et non des méthodes précises. L'évaluation a été orientée autour de plusieurs questionnements, que nous rappelons ici :

- quelle part joue chaque étape du système dans la qualité de la modélisation ? L'adéquation des i-vectors à un modèle génératif, hautement discriminant vis à vis du locuteur, tient-elle à l'ensemble des méthodes, ou principalement à certaines d'entre elles ?
- peut-on quantifier ces impacts ? Les effets des différentes procédures, en terme de performance, peuvent-ils être isolés ?
- pour mieux cerner ces systèmes en terme de fiabilité et pour orienter les recherches futures, une telle évaluation peut-elle dégager les facteurs-clés de la réussite du concept i-vectors ? Les étapes du système ont été élaborées sur une vingtaine d'années, au fur et à mesure de l'avancement du domaine. La pertinence de certaines démarches dans le nouveau contexte i-vectors est-elle toujours effective ?

Les trois méthodes les plus communément adoptées aux différentes étapes du système sont les suivantes :

- représentation par supervecteurs-MAP,
- réduction de dimensionnalité par FA-total variability,
- PLDA gaussienne

Pour mesurer la part jouée par ces méthodes dans la qualité de la modélisation et la comparer à celle imputable aux procédures de normalisation, nous avons proposé une alternative à chacune d'entre elles, basée sur une approche déterministe ou non-paramétrique. Les trois méthodes alternatives mises à contribution sont les suivantes :

- la représentation par clés binaires qui, comme celle par supervecteurs, s'appuie initialement sur la structuration de l'espace acoustique par le GMM-UBM, avant transformation non-paramétrique vers un espace binaire.
- la PCA, méthode de réduction de dimensionnalité déterministe. Nous l'appliquons à des supervecteurs adaptés (aux quantités d'informations associées à chaque composante gaussienne), afin de rendre la procédure comparable à la Factor analysis.
- les modèles LDA-Mahalanobis ou LDA-two-covariance, qui s'exemptent des contraintes de décomposition par maximum de vraisemblance de la PLDA.

L'évaluation a confirmé la part prépondérante des normalisations dans la qualité de la modélisation. Les méthodes initiales n'améliorent les performances que d'un facteur réduit, voire négligeable, en comparaison de celui induit par les transformations telles que SphN. La synthèse des résultats a permis de mettre en lumière les étapes fondamentales du paradigme i-vector, indépendamment des méthodes employées. Une fois

produites les trames acoustiques d'un segment de voix et estimé un GMM-UBM, ces étapes sont les suivantes :

1. Une représentation vectorielle est formée par concaténation d'informations des composantes gaussiennes. Le GMM-UBM est capable, quelles que soient les voies de représentation et de modélisations choisies, de structurer l'espace d'une manière suffisamment cohérente pour servir de socle à un système performant. Cette évaluation permet donc une nouvelle fois de lui associer le qualificatif de "paradigme".
2. L'extraction d'un i-vector à partir de cette représentation est une simple compression des données (réduction de dimensionnalité suivant la variabilité totale). C'est bien l'énoncé de voix complet qui est représenté par le vecteur en faible dimension et non des facteurs associés à certaines de ses variabilités. Cette compression est drastique et aveugle à toute variable latente, en particulier à la variable locuteur. Elle s'explique par l'opportunité de résumer les comportements corrélés des informations par gaussiennes, produisant ainsi des composantes principales du signal de voix. Ces composantes principales constituent les variables statistiques de nouvelles informations acoustiques de haut niveau. L'opportunité de la compression est donc essentielle, et non la méthode
3. La représentation compressée est alors transformée (post-normalisation) par des procédures qui ne sont ni linéaires ni issues de fonctions probabilistes (la standardisation suivant une variabilité-cible est suivie d'une uniformisation des longueurs). Ces procédures s'avèrent en effet les moyens les plus efficaces pour mettre les données à conformité de modèles et scorings à hypothèses gaussiennes. Aucun des raffinements probabilistes mis en oeuvre durant la chaîne de traitement ne contribue autant que ces normalisations à la qualité finale du système.

Ces procédures sont indissociables. On ne peut parler d'efficacité de la solution i-vector si l'un ou l'autre de ces principes n'est pas respecté : (1) vecteur de représentation par concaténation d'informations par composante gaussienne, (2) compression suivant la variabilité totale, (3) normalisation non-linéaire.

Une étude complémentaire a confirmé le caractère essentiel de ces étapes. Cette fois, les représentations sont constituées de vecteurs d'accumulation issus du GMM. Les probabilités d'occupation de chaque composante suivant la distribution du GMM, ou bien les comptes du modèle de clés binaire, sont considérés comme l'unique représentation de l'énoncé dont on dispose. Le résultat est bien, comme le préconise le principe (1), une concaténation d'informations par composante gaussienne (ici, une valeur réelle par composante). Puis, ces représentations ont été soumises à une modélisation PLDA, éventuellement précédée par une réduction de dimensionnalité (2) et par une normalisation de type Spherical Nuisance (3). Les résultats sont très corrects en terme de performance au regard du volume d'information des représentations, ce qui confirme le principe (1). De plus, ils montrent l'opportunité des principes (2) et (3) : l'amélioration relative en taux d'erreur avoisine les 40%, dont 25% sont induits par la réduction et 15% par la normalisation.

Le paradigme i-vector, défini par ces étapes fondamentales, est une représentation compacte et normalisée des énoncés complets (et non de facteurs ciblés de variabilité), issue d'une concaténation vectorielle d'informations locales du GMM-UBM. Le résultat est un vecteur de variables acoustiques de très haut niveau informatif. Ces supervariables quantifient les causes indépendantes de variabilité du signal de voix, dont celles propres au locuteur, qui nous intéressent. D'autre part, elles obéissent suffisamment à des hypothèses probabilistes pour leur appliquer des métriques robustes. Les représentations vectorielles des énoncés de voix à partir de ces supervariables répondent ainsi efficacement aux problématiques de la reconnaissance automatique du locuteur.

Lors de l'élaboration d'une technique de transformation des i-vectors (la normalisation des i-vectors *Spherical Nuisance* avant PLDA, paragraphe *4.6.1*), nous avons signalé la difficulté à estimer un métaparamètre linéaire universel de variabilité intra-locuteur lorsque les données avaient été transportées sur la surface d'une hypersphère pour gaussianisation. Cette surface est non-linéaire, ôtant toute validité a priori à une hypothèse d'homoscédasticité. Nous avons donc effectué une analyse supplémentaire qui vise à éprouver les limites du concept.

Les résultats de cette analyse montrent que l'utilisation de métaparamètres linéaires après sphérisation entraîne une anisotropie de l'espace de représentation. Une corrélation forte est mise en avant entre la position d'un centre de classe-locuteur et des caractéristiques statistiques de sa variabilité interne. Les estimations de proximité entre énoncés de voix vis à vis de la variable locuteur en sont affectées et leur précision limitée.

La démarche fondamentale présente donc des limites pratiques d'efficacité. Elles tiennent à la présence conjointe de caractères linéaire et non-linéaire. Plus généralement, ce fait rappelle que l'emploi de modélisations basées sur des estimations de paramètre globaux linéaires (matrices de covariance par exemple) s'avère contraignant, voire un obstacle à la qualité des systèmes. Des approches non-linéaires, ou linéaires localement, apparaissent comme indiquées, à l'horizon des recherches futures. Le caractère non aléatoire de l'anisotropie permet toutefois, dans le cadre de la RAL par i-vectors, d'envisager son redressement, une première piste de recherche étant présentée à l'annexe H.

Autour du modèle de clés binaires du locuteur

L'évaluation des étapes d'un système i-vector nous a amené à utiliser la représentation des énoncés de voix basée sur le modèle de clés binaires du locuteur. Parallèlement à ces investigations, nous avons contribué à ce modèle, qui propose une voie originale, semi-paramétrique et basée sur une nouvelle approche de la problématique. Nos contributions ont porté sur les points suivants :

– l'élaboration d'un recouvrement de l'espace acoustique, base sous-jacente au GMM-UBM d'un générateur de clés binaires.

– une technique d'égalisation des vecteurs avant binarisation, employée dans un esprit proche des transformations appliquées sur les i-vectors. Le but, dans un contexte non-paramétrique, n'est pas la mise à conformité à un modèle gaussien mais la meilleure prise en compte et détection d'exceptions, informations non décelables par une démarche statistique conventionnelle et qui participent pourtant efficacement à la caractérisation des énoncés vocaux.
– une dernière partie couvre la gestion des typicalités, qui sont des informations inter-dimensionnelles liées à une variabilité ciblée. Par une logique comparable à celle de la génomique, nous avons proposé des procédures recherchant des facteurs de variabilité sous forme d'héritages dans les clés binaires. Notre étude a porté sur les typicalités de type session, intrinsèques au locuteur. Elle a permis de mettre en place une technique de détection de ces variabilités nuisibles, dans un cadre formel binaire (matrice binaire de vecteurs propres, avec courbe d'énergie comparable au spectre de valeurs propres). Ce cadre est la traduction d'une démarche mathématique orientée *dimensions*.

D'une manière plus générale, l'ensemble des investigations menées autour du modèle de clés binaires démontrent la capacité à produire des systèmes de reconnaissance du locuteur, basés sur le GMM-UBM, dont les vecteurs de représentation et les métaparamètres sont beaucoup moins volumineux que ceux usuellement employés, en terme de quantité d'information. Le fait important est que cette réduction de complexité est rendue possible en intégrant dans la démarche des logiques d'*exception* et des transformations non- ou semi-paramétriques.

Des contextes de reconnaissance particuliers

Enfin, deux études ont été effectuées, visant à mettre en place et adapter des systèmes de vérification basés sur les i-vectors dans des contextes particuliers : le cas où la décision doit être prise avec tout ou partie d'énoncés de courte durée (moins de 20 secondes) et le cas où les énoncés de voix à comparer (enrôlement vs test) diffèrent d'un facteur binaire connu a priori (durée longue vs durée courte, microphone vs téléphone, langage A vs langage B).

Pour le premier cas, des hypothèses de complémentarité des bases d'acquisition de connaissances composées d'énoncés de courte et/ou longue durée, ont été validées expérimentalement. Elles montrent l'intérêt, pour des applications de reconnaissance du locuteur travaillant sur des énoncés de courte durée, d'inclure aux bases d'apprentissage des énoncés de plus longue durée. Une analyse de variance prouve que ces derniers contiennent une part plus forte d'information propre au locuteur. Les modèles de décomposition en facteurs locuteur et nuisance tels que la PLDA tendent soit à estimer par défaut la part de variabilité nuisible, soit à produire un facteur locuteur trop peu varié pour une discrimination efficace. La combinaison de systèmes basés sur des apprentissages de durée variable permet de mieux gérer le problème de la discrimination d'énoncés pauvres en information locuteur.

La procédure de normalisation dans l'espace des i-vectors s'avère utile à condition-

ner des données de durée hétérogène. Les représentations i-vectors des énoncés courts ou longs sont transportés sur la surface de la sphère en sorte que les sous-classes locuteur d'énoncés courts et longs ne présentent plus de divergence nette en terme de variance.

Pour le deuxième cas, l'adaptation bayesienne du modèle two-covariance prend en compte le facteur de non-correspondance entre les énoncés. Cette sophistication probabiliste affine la précision d'un modèle déterministe et ouvre des perspectives pour une adaptation de la PLDA à ce contexte particulier.

Perspectives

Plusieurs perspectives de recherche se présentent, suite à l'ensemble de ces travaux :

Optimisation du modèle de décomposition en facteurs

L'initialisation de la PLDA après normalisation Spherical Nuisance réduit considérablement l'utilité de la démarche probabiliste par maximum de vraisemblance. Les études effectuées dans cette thèse ont confirmé le rôle minime -voire négligeable- d'une procédure itérative par maximum de vraisemblance EM-ML dans la qualité de la modélisation PLDA. Ces travaux ont été étendus sur une large variété de configurations. Une fois les données normalisées, les métaparamètres du modèle ont été initialisés suivant la méthode que nous avons présentée au 4.6.7, en incluant cette fois les valeurs singulières de $\mathbf{B}$ dans la matrice $\mathbf{\Phi}$ de l'équation 4.24, en lieu et place de l'identité $\mathbf{I}$. Sur la quasi-totalité des configurations et conditions expérimentales testées, les métaparamètres ne nécessitent plus alors d'ajustement aux hypothèses par maximum de vraisemblance : sans leur modification par EM, le scoring PLDA atteint sa performance maximale. Ce constat ouvre plusieurs voies de recherche :

- l'algorithme EM par maximum de vraisemblance s'avère inopérant sur la surface sphérique de représentation. Son adaptation à la distribution sphérisée des données est à étudier.
- l'analyse d'anisotropie montre la non-optimalité d'une modélisation linéaire isotropique. Cette faiblesse explique sans doute l'inefficacité de la démarche EM-ML. La prise en compte d'une variabilité du métaparamètre de classe-locuteur suivant la position de cette classe dans l'espace doit permettre d'améliorer la qualité du système. Une modélisation non-linéaire est également envisageable.

Analyse de données : expliquer les variables

La réduction de dimensionnalité projette des vecteurs de l'espace acoustique (concaténés) dans l'espace des i-vectors. Elle projette donc les gaussiennes du GMM-initiales

dans un espace mono-gaussien en résumant, comme nous l'avons indiqué, les corrélations entre les gaussiennes d'origine. L'analyse spatiale de ces projections permet donc d'associer les composantes principales du signal de voix (supervariables de l'espace i-vectors) à des relations inter-composantes dans l'espace acoustique initial. Une telle analyse reste à effectuer.

Les composantes principales dans l'espace des i-vectors sont des supervariables acoustiques, d'effectif réduit (400 à 600). Leur "explication" mérite toute l'attention des chercheurs. En effet, si les systèmes de reconnaissance peuvent s'avérer performants, en particulier celui basé sur les i-vectors, l'acquisition de connaissances par leurs modèles reste opaque. Dans le domaine de l'analyse de données, les analyses en composantes principales n'ont pas pour seul but d'éliminer la part résiduelle, non ou peu informative. Elles permettent également de mettre en lumière des indicateurs, supervariables, qui synthétisent les variables initiales. Expliquer ces composantes revient à identifier les causes essentielles de variabilité, c'est à dire les ressorts de la population étudiée (pour nous les énoncés vocaux), vis à vis de ces variables. Dans le cadre de l'étude du signal de voix, associer ces composantes à des informations acoustiques de haut niveau (caractéristiques acoustiques, phonétiques, effort vocal, ...) constitue une voie de recherche prioritaire.

De la performance comme moyen de validation de la solution i-vector, il serait utile de passer à la notion de "fiabilité" (Boë and Bonastre, 2012), (van Leeuwen and Brummer, 2013), (Doddington et al., 1998), (Campbell et al., 2009), (Kahn, 2011), (Kahn et al., 2010) : seule une compréhension claire des contenus et détails de la représentation compacte normalisée peut permettre de mieux cerner son matériau de base, la voix humaine. Il sera alors possible de circonscrire et maîtriser les systèmes de reconnaissance basés sur cette dernière, de la manière la plus indépendante possible des contextes d'expérimentation.

Alternatives à l'état de l'art

La mise en lumière des étapes fondamentales de la solution i-vector montre sa relative indépendance aux méthodes employées à chacune de ses étapes. D'autres représentations en haute dimension, issues du GMM-UBM, peuvent s'avérer plus pertinentes, ces représentations n'étant plus décisionnelles mais seulement intermédiaires avant une réduction de dimensionnalité. Le modèle de clés binaires s'est ainsi montré capable de représenter les énoncés avec pertinence. Si l'on garde à l'esprit la faible quantité d'information d'un vecteur de clés, ce modèle montre l'opportunité d'élargir le cadre de représentation au delà des voies statistiques et probabilistes usuelles.

Les travaux réalisés sur les représentations par accumulation ont montré que le pouvoir discriminant des vecteurs de comptes issus du générateur de clés binaires surpasse celui des vecteurs de probabilités gaussiennes d'occupation des composantes. De même, la réduction de dimensionnalité probabiliste (Factor analysis) peut inclure de nouvelles contraintes et de nouvelles formulations, pour améliorer la modélisation des i-vectors qu'elle produit. De nombreux travaux tentent notamment d'appliquer cer-

taines techniques de normalisation durant, et non après, la phase d'extraction (Kenny, 2012). Les modèles et scorings dans l'espace des i-vectors peuvent donc être re-façonnés (prise en compte de l'anisotropie) ou, simplement, remplacés par des modèles non- ou semi-paramétriques, moins dépendants d'exigences probabilistes, en particulier gaussiennes.

Approche analytique

L'extraction de i-vectors s'effectue dans tous les cas par un calcul matriciel, qui peut être décomposé. Par exemple, lorsque les i-vectors sont calculés par PCA sur des vecteurs de clés binaires, ils sont égaux à une somme de vecteurs (ces points d'ancrage sont les images des spécificités de l'espace initial), pondérée par 1 ou 0 (un i-vector d'un énoncé est la somme d'une sélection de ces points). La nature additive du i-vector peut faciliter son analyse, notamment dans le cadre précédent de l'explication des composantes principales. Elle peut aussi servir de base à la mise en place de méthodes plus complexes. Ces méthodes extrairaient du i-vector des typicalités, élimineraient certaines anomalies et plus généralement remplaceraient leur décomposition additive en facteurs par une batterie de sous-vecteurs, associé chacun à une variabilité ciblée.

Troisième partie

Contributions complémentaires

Chapitre 8

Contributions aux modèles de clés binaires du locuteur

Nous présentons ici l'ensemble des travaux que nous avons effectués dans le champ d'études du modèle de clés binaires du locuteur. Ceux-ci portent en premier lieu sur l'élaboration d'un recouvrement de l'espace acoustique sous-jacent au GMM-UBM, base du générateur de clés binaires. Nous présentons et justifions également les techniques dites d'égalisation par gaussiennes, qui visent à mieux détecter les *exceptions* d'un énoncé de voix, c'est à dire des particularités de l'énoncé non nécessairement paramétriques. Nous détaillons enfin un travail réalisé autour du thème des *typicalités*. Son but est d'extraire d'une représentation par clés binaires des caractéristiques informatives, vis à vis d'une variabilité ciblée, dans un contexte formel de type binaire orienté *dimensions*.

8.1 Algorithmes de recouvrement

Déterminer les spécificités par gaussienne (définies en 2.8) revient à former un recouvrement discret de chaque région gaussienne de l'espace acoustique structuré par le GMM-UBM. Les points-spécificités choisis par composante sont censés résumer la variabilité possible des trames acoustiques sur celle-ci.

8.1.1 Recouvrement par base de modèles imposteurs

Une première famille d'algorithmes de recouvrement s'appuie sur une base de modèles imposteurs. Les spécificités sont alors sélectionnées parmi les vecteurs de moyennes par gaussienne (composant le supervecteur adapté par MAP).

Il est possible de sélectionner la quantité de spécificités voulue sur chaque gaussienne par tirage aléatoire ou bien par un algorithme de type *Maximum Relevance Minimum Redundancy MRMR* (Bonastre et al., 2011b). Cet algorithme est hiérarchique : il

forme le recouvrement en q points d'une gaussienne par sélection consécutive de modèles (i.e. de moyennes) et non simultanée. Le double critère de sélection du $(k+1)^{\text{ème}}$ modèle est la maximisation de la pertinence de ce choix et la minimisation de sa redondance. Pour respecter le premier objectif, le modèle doit être le plus proche possible de la moyenne du monde, pic de densité avéré de la composante. Et pour respecter le second objectif, il doit être le plus éloigné possible des k modèles déjà sélectionnés. Le critère de sélection d'une spécificité de la gaussienne g est :

$$\arg\max_{s\in\overline{\mathcal{S}}} \frac{\sum_{e\in\mathcal{S}} d\left(s,e,\mathbf{\Sigma}_g\right)}{d\left(s,\mu_g,\mathbf{\Sigma}_g\right)} \tag{8.1}$$

où $\mathcal{S}$ est le sous-ensemble des modèles de la gaussienne g déjà sélectionnées, $\overline{\mathcal{S}}$ son complémentaire et $d\left(x,y,\mathbf{\Sigma}_g\right)$ une distance entre composantes gaussiennes, par exemple la densité gaussienne $\mathcal{N}\left(x|y,\mathbf{\Sigma}_g\right)$.

Il peut être aussi effectué une classification en q classes sur chaque gaussienne suivant un algorithme d'agrégation autour des centres mobiles (*kmeans*), celui-ci utilisant la métrique Σ_g^{-1} de la matrice de précision du monde. Les moyennes imposteurs sélectionnés sont alors les q-plus proches voisins des centres constitués. On note $\mathbf{\Sigma}^{-1}$-*KNN-kmeans* cet algorithme.

8.1.2 Recouvrement par extension locale du GMM-UBM

La seconde stratégie consiste à prolonger la mixture du GMM-UBM sur chacune de ses composantes pour former un sous-GMM le plus vraisemblable possible. Mais l'obtention de ce GMM par algorithme EM-ML pour chaque composante est rendu peu efficient par la faible quantité d'informations d'apprentissage. Pour contourner cette carence en informations, la tactique consiste à remplacer la densité gaussienne de paramètres $\left(\mu_g,\Sigma_g\right)$ par la mixture de q densités la plus vraisemblable en l'absence de nouvelles informations. La matrice de covariance Σ_g étant diagonale, donc les dimensions acoustiques indépendantes, ce modèle est constitué de la famille de densités de moyennes et covariance :

$$\left\{\mu_g \pm \delta_f \sqrt{\left(\mathbf{\Sigma}_g\right)_{f,f}}, \mathbf{\Sigma}_g\right\}_{f=1}^{F} \tag{8.2}$$

où $\sqrt{\left(\mathbf{\Sigma}_g\right)_{f,f}}$ est le vecteur des écart-type de la composante et δ_f est le vecteur de Kronecker de dimension F égal à 1 sur sa $f^{\text{ème}}$ coordonnée, à 0 sinon.

Ici, F est la dimension de l'espace acoustique. Une spécificité est obtenue par translation de la moyenne monde μ_g, suivant une dimension donnée, d'un pas d'un écart-type de cette dimension $\sqrt{\left(\mathbf{\Sigma}_g\right)_{f,f}}$ ou $-\sqrt{\left(\mathbf{\Sigma}_g\right)_{f,f}}$. Le nombre total de spécificités est donc $2GF$. Les covariances des sous-gaussiennes ainsi formées sont identiques à l'initiale $\mathbf{\Sigma}_g$ du monde.

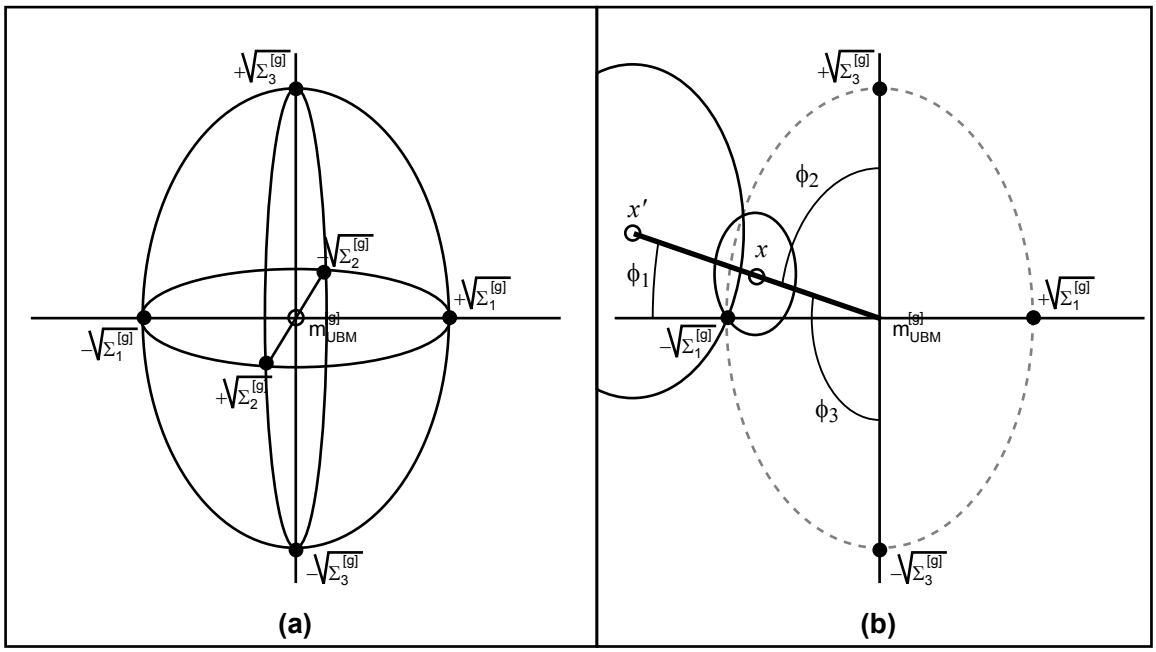

FIGURE 8.1 – *Illustration en 3D du recouvrement par coquille gaussienne à covariance diagonale, autour de la moyenne monde.*

La figure 8.1 partie (a) représente ce recouvrement en dimension 3. Les dimensions étant indépendantes, ce recouvrement tapisse l'espace suivant la coquille elliptique gaussienne de la covariance $\mathbf{\Sigma}_g$.

Ce recouvrement possède plusieurs propriétés :

- le principe de sélection de spécificités par trame suivant les ellipses de distribution de Σ_g entraîne que les vecteurs binaires obtenus seront identiques si l'on remplace le coefficient $\sqrt{(\mathbf{\Sigma}_g)_{f,f}}$ par $\alpha\sqrt{(\mathbf{\Sigma}_g)_{f,f}}$, α réel positif. En effet, le calcul de distance trame-spécificités ne tient compte que de proximité directionnelle, comme l'indique la figure 8.1 partie (b) où les deux trames x et x' "allument" la même spécificité.
- les spécificités sont équidistantes de la moyenne monde au sens de la métrique de $\mathbf{\Sigma}_g^{-1}$, donc de la loi gaussienne $N\left(|\mu_g, \mathbf{\Sigma}_g\right)$, ce qui signifie qu'elles sont équiprobables vis à vis du monde. De même, étant donnée une spécificité $s^+_{g,f} = \mu_g + \delta_f\sqrt{(\mathbf{\Sigma}_g)_{f,f}}$, celle-ci se trouve à égale distance de toutes les autres spécificités de la composante au sens de $\mathbf{\Sigma}_g^{-1}$ hormis de son opposée sur l'axe $s^-_{g,f} = \mu_g - \delta_f\sqrt{(\mathbf{\Sigma}_g)_{f,f}}$. Chaque spécificité a donc un voisinage constitué de $(N-2)$ plus-proches-voisines équidistantes sur $(N-1)$ possibles. Cette structuration présente l'avantage de contourner la question de la "cartographie" du recouvrement : à savoir la nécessité de disposer des voisinages de chaque spécificité pour déterminer

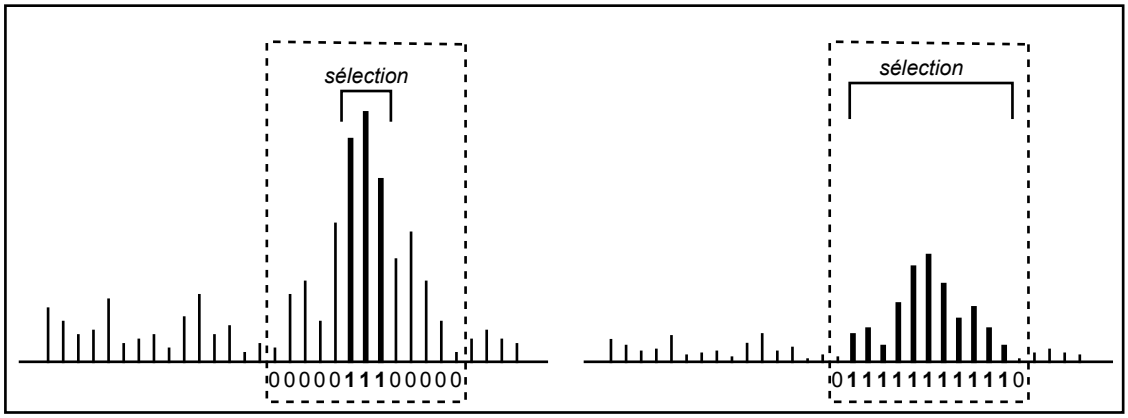

FIGURE 8.2 – *Sélection de zones de haute densité avec intervalle de confiance.*

les zones connexes de haute densité à résumer. Il permet également de lever une difficulté d'analyse du modèle binaire : en effet, nous montrons plus bas que les clés binaires des imposteurs produites à partir de ce recouvrement tendent à l'indépendance statistique

Ce recouvrement peut être considéré comme une solution à la recherche du sous-GMM équilibré le plus vraisemblable à posteriori d'une composante du GMM-UBM. Ne dépendant ni de nouvelles informations ni de nouvelles hypothèses, il constitue une base fonctionnelle pour étudier la capacité discriminante du modèle de clés binaires.

8.2 Egalisation des appels par gaussienne

Les vecteurs de clés binaires de la collection de trames sont synthétisés en une représentation de taille fixe par addition terme à terme. Le *vecteur de comptes* obtenu est un vecteur d'accumulation, qui est ensuite seuillé pour produire le vecteur de clés binaires du segment.

Lorsqu'un certain nombre de points d'un même voisinage présentent des pics de densité élevés (c'est à dire lorsque leurs valeurs dans le vecteur de comptes sont élevées), la question se pose de les résumer ou pas et en quelle proportion : une logique intuitive tendrait à conserver les points présentant les densités les plus élevées dans l'absolu. Mais suivant une logique statistique, plus la quantité d'information qui a permis de "détecter" ces points est élevée, plus l'intervalle de confiance sur la localisation de leur centre de masse est restreint (son amplitude 1/précision tend vers 0). Et donc, il est préférable de ne conserver qu'un nombre restreint de points, voire un seul, sa localisation s'appuyant sur une estimation fiable. Par contre, si la quantité d'information est faible, l'intervalle de confiance sera grand (la précision tend vers 0) et donc le nombre de points à retenir sera élevé pour localiser cette zone à haute densité.

La figure 8.2 illustre ce propos. Sur un espace unidimensionnel discrétisé représenté

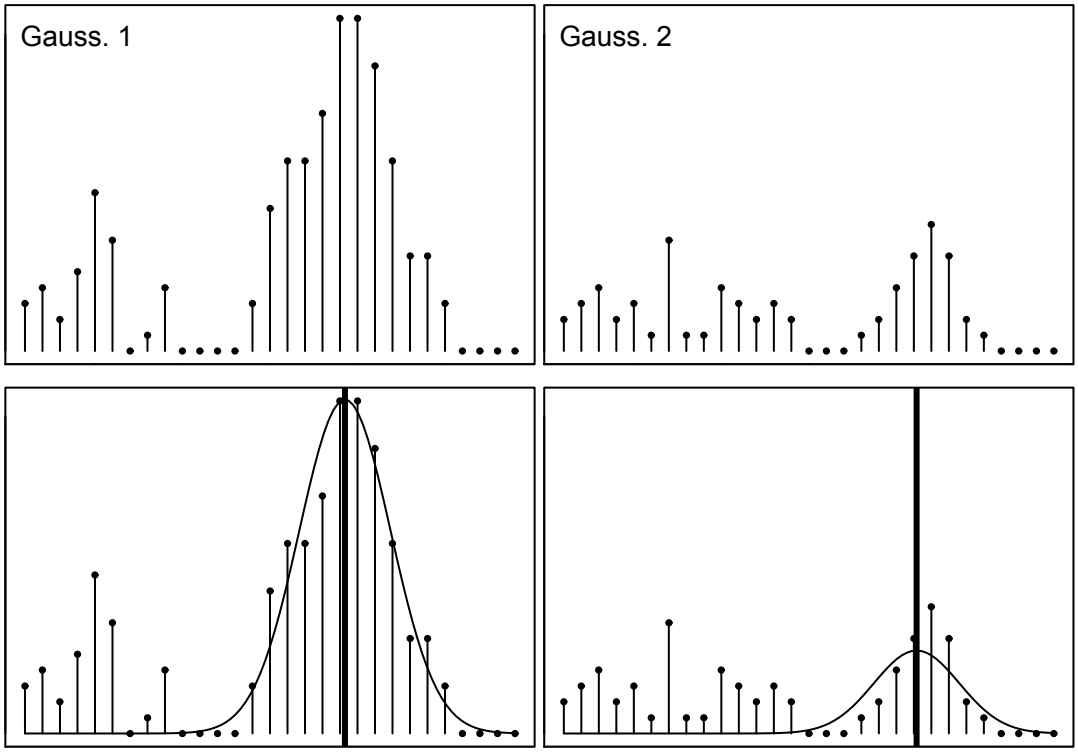

FIGURE 8.3 – *Lois gaussiennes déduites des zones principales de densité de deux gaussiennes (gauche et droite).*

par l'axe horizontal, est affiché en ordonnée l'histogramme des densités. La zone de haute densité à gauche a été obtenue avec une quantité d'informations importante (effectifs élevés) et seul un nombre réduit de pics a été retenu. A droite, la zone de haute densité étant obtenue avec peu d'informations, sa confiance est limitée et un nombre importants de pics doit être choisi pour la résumer fiablement.

La logique d'*exceptions* décrite précédemment dans la description du modèle peut être ainsi prolongée par cette notion d'intervalles de confiance.

La figure 8.3 affiche les comptes obtenus sur les spécificités de deux gaussiennes (première ligne). La deuxième ligne affiche les lois gaussiennes déduites des zones principales de densité des gaussiennes 1 et 2. Rappelons que la faible quantité d'informations sur la gaussienne 2 attribuera à sa loi un faible poids.

La figure 8.4 indique, sur sa première ligne, les spécificités qui seront retenues pour constituer le vecteur binaire du segment originel (8 plus forts comptes par exemple). On voit que sont sélectionnés des spécificités de la gaussienne 1, qui regroupe en effet la majorité des appels. Sept spécificités résument la loi gaussienne précédente, la huitième est isolée : le modèle binaire permet de relever plusieurs pics de densité par composante gaussienne. Mais, d'une part, l'intervalle de confiance sur la loi principale de la gaussienne 1 doit être réduit, du fait de sa grande quantité d'appels. D'autre part,

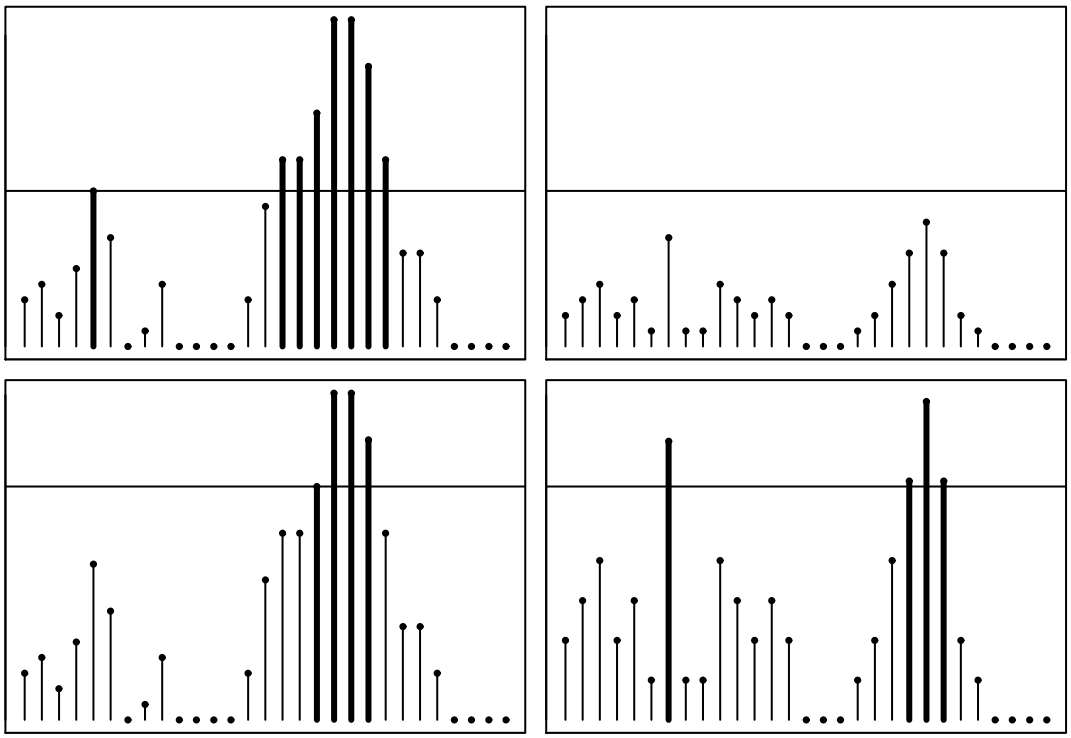

FIGURE 8.4 – *Spécificités retenues pour les clés binaires d'une trame, sans égalisation préalable (en haut) et avec égalisation (en bas).*

la gaussienne 2 est ignorée, jugée insuffisante en quantité totale d'appels.

Nous avons mis en place des techniques destinées à prendre en compte ces remarques, par exemple en limitant le nombre d'appels total par gaussienne. La méthode finalement adoptée normalise simplement les appels par gaussienne. Nous la nommerons *égalisation des appels*. Il s'agit, étant donné un vecteur de comptes c, de produire sa version égalisée $\widehat{c}$ en divisant chaque compte par le total d'appels de sa gaussienne. C'est à dire que, pour toute gaussienne g :

$$\widehat{c}_{g,k} = \frac{c_{g,k}}{\sum_{k=1}^{q} c_{g,k}} \tag{8.3}$$

si bien qu'on a $\sum_{k=1}^{q} \widehat{c}_{g,k} = 1$.

Après égalisation, la deuxième ligne de la figure 8.4 indique les spécificités retenues pour constituer le vecteur binaire du segment. La loi principale de la gaussienne 1 a engendré la sélection de quatre spécificités seulement. Elles sont trois pour la loi de la gaussienne 2, où une exception a été retenue (le pic unique à gauche de cette gaussienne). Par contre, l'exception de la gaussienne 1 n'en est plus une et n'est pas sélectionnée, son énergie n'étant pas suffisante sur la région gaussienne.

Nous disposons donc d'un moyen d'exalter les exceptions tout en continuant de relever des zones de forte densité, ces dernières étant régulées par la notion d'intervalle de confiance.

8.3 Performance sans apprentissage de la variable locuteur

Nous présentons ici le résultat d'expériences menées sur le modèle de clés binaires. Aucune technique basée sur l'apprentissage de la variable locuteur (*intersession compensation*) n'a été encore incluse dans le système et nous donnons pour comparaison la performance obtenue par le système noté MAP dans la partie 2.7. Nous rappelons que ce système utilise un GMM-UBM, la représentation par statistique d'ordre 1 du GMM (supervecteurs) adaptée par MAP et un scoring LLR-by-frame dans lequel le modèle imposteur est approximé par le GMM-UBM. L'expérience est menée sur la même condition *det 7* de NIST-SRE 2008 qu'au paragraphe 2.7. Pour le modèle de clés binaires, la configuration est la suivante :

- le GMM-UBM est celui codé GMM-UBM-LIA détaillé en annexe A.
- Le recouvrement est constitué par coquille gaussienne, tel que décrit précédemment 8.1. La dimension de l'espace acoustique étant $F = 50$, il contient $N = 512 \times 50 \times 2 = 51200$ spécificités. Pour chaque trame, le nombre de gaussiennes retenues vaut 3 et le nombre de spécificités par gaussienne est fixé à 40. Le nombre

	EER %	DCF min	dim	octets	taille
MAP	**7.74**	0.0354	25600	8	204 800
BK	**8.88**	0.0469	51200	1/8	6 400

TABLE 8.1 – *Performances comparées, en termes d'EER et DCF min, des systèmes supervecteurs-MAP et clés binaires sans compensation de l'effet session. Les tailles des représentations haute-dimension sont également indiquées.*

de valeurs mises à 1 dans le vecteur binaire final d'un segment est de 19200. L'égalisation par gaussienne décrite précédemment a été effectuée avant cette binarisation finale.

8.3.1 Score modèle vs modèle

La table 8.1 affiche la comparaison du baseline **MAP** et du modèle de clés binaire noté **BK**. Le modèle binaire atteint une performance de 1.14% inférieure en écart relatif à celle du baseline MAP en terme d'EER. Le vecteur de clés binaires s'avère moins performant que le supervecteur. Mais ce résultat doit être re-considéré en tenant compte que le système binaire compare deux modèles de clés binaires, celui du locuteur-cible et celui constitué à partir du segment de test (*modèle vs modèle*), alors que le modèle MAP compare le modèle du locuteur-cible à la totalité des trames du test, passées en en revue lors du score LLR-by-frame (*modèle vs trames*). Une comparaison plus juste est effectuée dans le paragraphe suivant. Mais ces résultats méritent d'abord une discussion : la même table 8.1 indique la dimension de la représentation vectorielle utilisée ($512 \times 50 = 25600$ pour le supervecteur et $512 \times 50 \times 2 = 51200$ pour le modèle binaire), le codage de ses coefficients en octets (réels doubles pour les supervecteurs, bits pour le binaire) et donc la quantité d'informations ayant permis d'atteindre ces performances. Comme cette table l'illustre, un vecteur de clé binaire est bien plus compact qu'un supervecteur, d'un facteur 32. Cette expérience montre qu'à partir d'une structuration de l'espace acoustique par GMM-UBM, il est possible d'atteindre une performance proche de l'optimal à partir de représentations de taille beaucoup plus réduite que celle des supervecteurs usuels.

8.3.2 Score modèle vs trames

Afin de comparer les capacités discriminantes des modèles MAP et de clés binaires, nous introduisons un équivalent du scoring LLR-by-frame au modèle binaire.

Son principe est le suivant : étant donné un locuteur-cible et un segment de test, on note $\widehat{c}$ le vecteur de comptes égalisé de ce locuteur. Pour toutes les trames x du segment de test, sont calculées leurs log-vraisemblances par rapport à toutes les spécificités $s_{g,k}$ (c'est à dire les logarithmes des densités $\log \mathcal{N}\left(x|s_{g,k}, \mathbf{\Sigma}_g\right)$). Ces vraisemblances sont cumulées dans un vecteur qu'on notera $llk_{[\text{test}]}$.

	EER %	DCF min
MAP	**7.74**	0.0354
BK llr-by-frame	**7.74**	0.0502

TABLE 8.2 – *Performances comparées, en termes d'EER et DCF min, des mêmes systèmes sur scorings LLR-by-frame.*

Le vecteur $\widehat{c}$ est alors considéré comme un vecteur de poids d'un sur-GMM de taille N. A partir de celui-ci, est constitué le vecteur de clés binaires du locuteur, noté b. Un vecteur d'accumulation $\widehat{\widehat{c}}$ est alors calculé, en multipliant simplement les coefficients de $\widehat{c}$ et b, puis en réexécutant la procédure d'égalisation par gaussienne. Le résultat est un vecteur de poids d'un sur-GMM de taille N dont les composantes non retenues pour cette cible sont nulles.

Nous considérons le vecteur $\widehat{\widehat{c}}$ comme vecteur de poids du sur-GMM du locuteur-cible et le vecteur $\widehat{c}$ comme celui de ses imposteurs. Un score LLR-by-frame s'en déduit, en pondérant les vraisemblances $llk_{[\text{test}]}$ par les poids $\widehat{c}$ ou $\widehat{\widehat{c}}$:

$$score\,(cible, test) = llk_{[\text{test}]}.\widehat{\widehat{c}} - llk_{[\text{test}]}.\widehat{c} \tag{8.4}$$

où "." représente le produit scalaire.

Quelques précautions doivent être prises : si une gaussienne n'est jamais appelée par le test, ses coefficient dans $llk_{[\text{test}]}$ sont nuls, valeur maximale pour la log-vraisemblance. Cette gaussienne doit être ignorée dans les calculs. Il en va de même si aucune spécificité d'une gaussienne n'a été retenue dans b.

La table 8.1 affiche les résultats obtenus avec la même configuration du modèle binaire que précédemment, mais utilisant ce nouveau scoring. Ce système est noté BK llr-by-frame. Une coïncidence exacte (et fortuite) se produit en terme d'EER. La DCF minimale reste supérieure à celle du baseline MAP.

Ces résultats montrent que le modèle par clés binaires peut concurrencer celui par supervecteurs et MAP dans les systèmes n'incluant pas d'apprentissage de la variable locuteur.

8.3.3 Indépendance des clés binaires

Dans une logique tirée de la génomique, à laquelle le modèle de clés binaires du locuteur peut être apparenté, on attend du vecteur binaire d'un énoncé qu'il hérite des caractéristiques vocales de son locuteur. La quantité astronomique de clés binaires qu'il est possible de constituer, en sélectionnant n valeurs binaires parmi N, doit être relativisé par le degré de corrélation entre ces clés. L'indépendance des représentations de deux locuteurs distincts est donc un objectif à se fixer.

A partir des scores imposteurs obtenus par la première des expériences précédentes, il est possible d'estimer le degré d'indépendance des clés binaires de segment de locuteurs distincts. Nous mesurons si les clés produites par le recouvrement en coquille gaussienne de l'UBM ne présentent pas de corrélations significatives non-informatives, c'est à dire non imputables à la différenciation d'identité. Etant donné un test-imposteur, ses deux vecteurs binaires à comparer contiennent chacun n valeurs à 1 parmi N. L'espérance de leur fréquence de 1 communs, sous l'hypothèse d'indépendance de leurs dimensions, peut être approximée par la valeur $\left(\frac{n}{N}\right)^2$ tirée d'une loi binomiale [1]. Avec la configuration de l'expérience précédente, cette valeur théorique est de 0.1406. La fréquence des scores imposteurs sur l'expérience est de 0.1496. L'indépendance des clés binaires est quasiment vérifiée sur cette expérience.

Ce fait est important : il montre la capacité du modèle de clés binaires à produire des représentations propres au locuteur, c'est à dire à variables aléatoires binaires quasi-indépendantes. Sa capacité à reconnaître des énoncés d'un locuteur-cible va donc dépendre du degré de dépendance que cette seule variable latente induit dans la modélisation.

8.3.4 Conclusion

Ces expériences montrent la capacité du modèle binaire à discriminer le locuteur, à partir d'une faible quantité d'informations, avec des performances d'ordre proche de l'état-de-l'art. Mais des techniques basées sur l'apprentissage de la variabilité locuteur doivent être mises en place dans l'espace des clés binaires pour comparer pleinement cette nouvelle représentation aux autres approches. Ce sont les objectifs de la partie suivante.

8.4 Extraction de typicalités

Les clés binaires indiquent, pour chacune des N spécificités du modèle, si elle doit être retenue pour représenter le segment de voix initial. Les N dimensions de l'espace binaire sont des variables catégorielles ("facteurs") avec seulement deux niveaux : *oui* ou *non*. Nous décrivons ci-dessous la technique d'apprentissage et prise en compte de la variable latente locuteur (*intersession compensation*) que nous avons mise en place pour améliorer le taux de discrimination des locuteurs dans cet espace binaire (Bousquet and Bonastre, 2012).

Le concept de "typicalité" est introduit, constituant des liens entre spécificités porteurs d'information discriminante. Puis nous montrons que des techniques linéaires usuelles de compensation de l'effet session peuvent être employées dans cet espace de clés binaires. Enfin, nous présentons un exemple d'algorithme d'extraction de ces typicalités et validons son efficacité sur une expérience de vérification du locuteur.

1. Approximée seulement car n fixé ôte un degré de liberté au problème.

La proximité entre deux deux clés binaires pour une spécificité donnée peut être obtenue simplement par l'opérateur logique ET. Les deux catégories étant codées 0 et 1, une similarité entre deux clés binaires v_1 et v_2 peut être alors calculée par :

$$S(v_1, v_2) = v_1.v_2 = \sum_{i=1}^{N} (v_1)_i (v_2)_i \tag{8.5}$$

où le produit scalaire "." est équivalent dans l'espace binaire au ET logique pour deux variables catégorielles. La similarité est éventuellement normalisée par division par N (ou par le nombre maximal de spécificités sélectionnées) pour fournir une valeur dans $[0, 1]$.

Notons que les spécificités constituent bien des variables aléatoires catégorielles (i.e. "des "facteurs"). L'utilisation du produit scalaire entre valeurs binaires ne se justifie, ici comme par la suite, que dans le sens où il coïncide avec la mesure de proximité de l'opérateur ET entre facteurs.

La vérification du locuteur peut être améliorée par l'emploi de nouvelles similarités. Notre but ici est de révéler des familles de spécificités liées par une relation à caractère discriminant. Nous supposons qu'un nouveau calcul de similarité entre deux vecteurs-clés binaires prenant en compte ces relations améliorera le processus de décision.

Une **typicalité** est un sous-ensemble $\mathcal{L}$ des spécificités $\{1, ..., N\}$ à partir duquel il est possible d'affiner la discrimination par ajout d'informations complémentaires sur les proximités entre vecteurs. Une typicalité peut contenir des informations sur les variabilités totale, inter-locuteur, intra-locuteur ou résiduelle mais aussi sur la distribution des imposteurs. Cette similarité considère que chaque combinaison des spécificités sélectionnées pour cette typicalité indique une "trace" d'un effet utile à la discrimination (par analogie à la génomique) et est ainsi plus efficace que le produit scalaire qui procède seulement dimension par dimension.

L'ensemble des aspects mathématiques, ainsi qu'un exemple d'extraction de typicalités, sont détaillés en annexe E. Nous présentons seulement ici (figure 8.5) le schéma général de cet algorithme d'extraction de typicalité et scoring dans le domaine binaire. La particularité de la démarche est de produire un score uniquement basé sur des vecteurs et matrices de vecteurs propres binaires :

1. Une décomposition en valeurs singulières de la matrice de covariance intra-locuteur des clés binaires produit un jeu de r vecteurs principaux $\lambda_r v_r$, où λ_r et v_r sont respectivement les valeurs et vecteurs propres issus de la décomposition.
2. Cette matrice est binarisée : deux matrices de vecteurs propres binaires "antagonistes" $\mathbf{B}^{(1)}$, $\mathbf{B}^{(2)}$ sont générées, qui relèvent les covariances significatives entre dimensions exprimées par $\{\lambda_r v_r\}$. La méthode limite le nombre de 1 par dimension (colonnes) et non par vecteur, pour préserver la quasi-orthogonalité des vecteurs propres résultants.
3. Une mesure de similarité est alors produite à partir de ces matrices $\mathbf{B}^{(1)}$ et $\mathbf{B}^{(2)}$, suivant une métrique de type NAP.

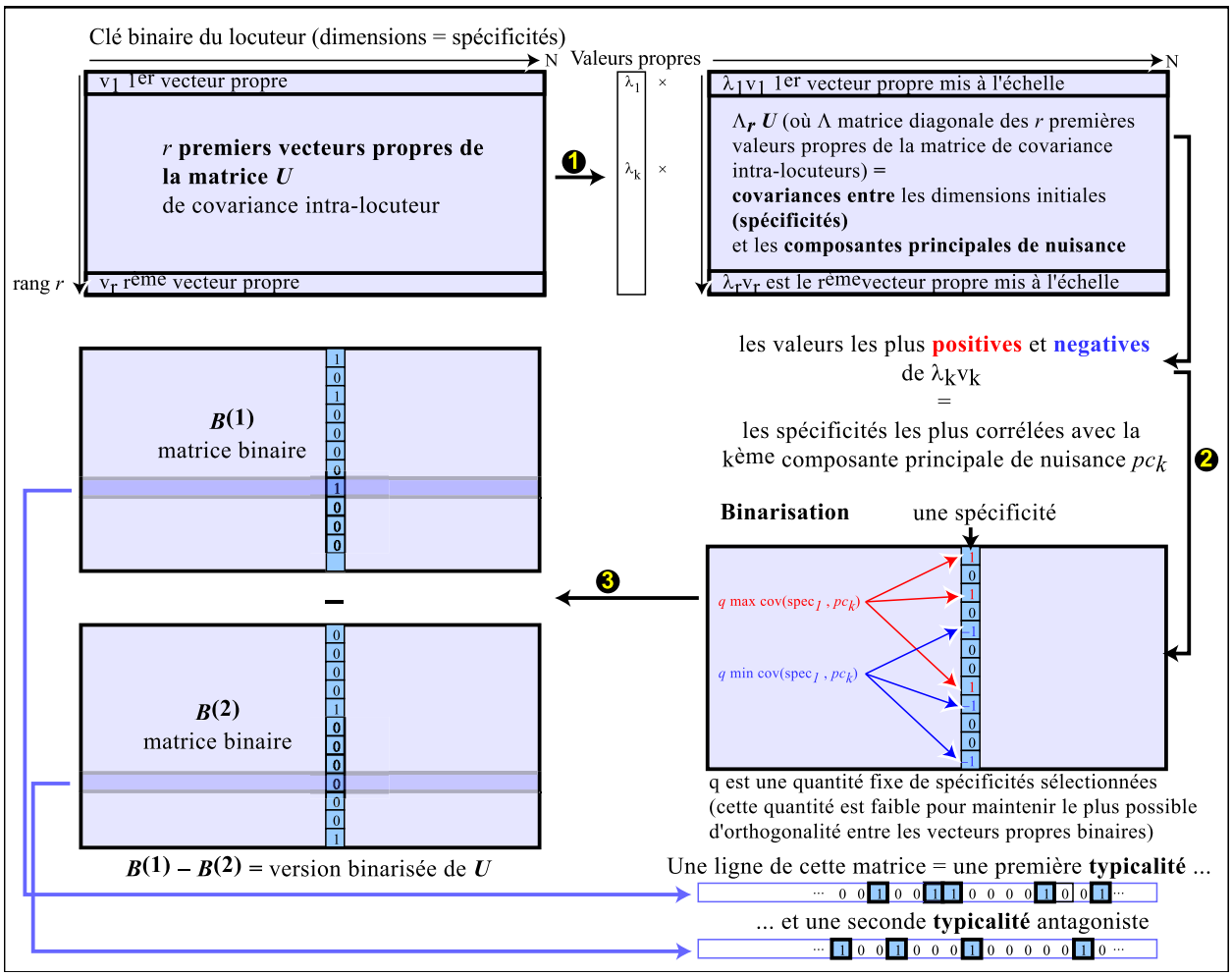

FIGURE 8.5 – *Les étapes de génération de matrices de vecteurs propres binaires basées sur les typicalités.*

L'application de cette technique présente l'avantage de mieux éclairer la question de la précision des systèmes de reconnaissance, au regard des quantités d'information de leurs modélisations. La comparaison de systèmes vis à vis de ce paramètre fait l'objet du paragraphe suivant.

8.4.1 Quantité d'informations

Le système mis en place de reconnaissance du locuteur par clés binaires et typicalités (dont le détail est présenté en annexe E) utilise la matrice **A** de l'équation E.21 pour calculer une similarité entre deux clés binaires. Nous nous intéressons ici à la capacité de ce système à discriminer le locuteur en terme de performance mais aussi de quantité d'informations matricielles nécessaire à la détermination du score. Les systèmes de vérification du locuteur emploient en général une ou plusieurs matrices censées synthétiser les divers types d'informations (locuteur, session, résidu, ...). Il s'agit donc de comparer les systèmes état-de-l'art JFA sur les supervecteurs de moyenne et vecteurs d'occupation et PLDA sur les i-vectors à notre système NAP sur clés binaires.

La table 8.3 compare les performances de ces systèmes en terme d'EER et de quantité d'information matricielle nécessaire à produire cette performance. Le jeu d'évaluation commun à ces expériences est toujours la condition *det 7* de NIST-SRE 2008. Pour la PLDA, elle a été appliquée aux i-vectors LIA-hommes et BUT-hommes. La table indique les rangs de matrices locuteur et sessions ayant produit les meilleures performances. La dernière ligne de la table indique la taille totale en kilo-octets des matrices utilisées. Par exemple, pour le système "i-vectors LIA" de la deuxième colonne, les matrices locuteur et session, de dimension 400×300 et 400×400 codées en 64 bits, occupent un volume en kilo-octets de $(400 \times (300 + 400) \times 64/8)\ /1000 = 2240$ Ko.

Comparée aux systèmes JFA et i-vectors (1) (ce dernier correspondant à la performance optimale), l'EER du NAP binaire, de 6.64 %, est loin des EER de ces deux précédents. Mais la quantité matricielle d'information nécessitée par le NAP binaire est considérablement réduite, de 20480, 2240 et 3264 Ko à 82 Ko.

La dernière partie à droite de la table affiche les valeurs correspondant à des systèmes i-vectors LIA ou BUT de même quantité d'informations que le NAP binaire (de l'ordre de 80 Ko). Nous avons réduit les rangs locuteur et session, pour obtenir la meilleure performance de ces systèmes sous la contrainte de quantité d'informations.

Les EER obtenus (8.90 et 36.62 %) sont moins bons que ceux du NAP binaire. La valeur très élevée de la performance BUT est surprenante. Elle rappelle que ces i-vectors, plus discriminants que ceux de la plupart des systèmes dans le monde, dépendent considérablement de l'accroissement de dimension mis en place.

Système					
JFA	i-vectors (1)		NAP	i-vectors (2)	
	LIA	BUT	binaire	LIA	BUT
Dimension de l'espace					
512×50	400	600	128×256	400	600
Représentation					
supervecteurs	i-vector		clé binaire	i-vector	
Type de données (bits)					
(double) 64	(double) 64		(binaire) 1	(double) 64	
Taille représentation					
204.8 Ko	3.2 Ko	4.8 Ko	4.1 Ko	3.2 Ko	4.8 Ko
Rangs des matrices					
$40 + 60$	locuteur 300 session 400	locuteur 80 session 600	$10 + 10$	locuteur 10 session 16	locuteur 8 session 9
Quantité d'informations des métaparamètres matriciels					
20480 Ko	2240 Ko	3264 Ko	82 Ko	83 Ko	82 Ko
EER %					
3.89 %	1.58 %	0.90 %	6.64 %	8.90 %	36.62 %

TABLE 8.3 – *Comparaison de performances et quantités d'informations matricielles requises, pour différents systèmes : Joint Factor Analysis (JFA), i-vectors et NAP-binaire. Aucune normalisation de scores n'a été appliquée.*

8.5 Perspectives

Faire apparaître des liens entre spécificités permet d'améliorer le pouvoir discriminant du recouvrement. Un algorithme itératif de recouvrement est concevable, qui utiliserait les informations extraites des typicalités pour sélectionner les spécificités. L'emploi de matrices de vecteurs propres binaires, qui approche les matrices continues en terme de performance de reconnaissance, montre la qualité des liens inter-spécificités et leur faculté à résumer un type de variabilité.

De plus, le problème de la discrimination des locuteurs peut être considérablement réduit en terme de dimensionnalité, comme le montrent les i-vectors, mais aussi en terme de quantité d'informations, comme le montrent les clés binaires.

Chapitre 9

Adaptations des systèmes i-vectors à des contextes particuliers

9.1 Introduction

Les techniques de normalisation que nous avons présentées dans ce document ont été évaluées et validées sur des applications de vérification indépendante du texte et sur des corpus de segments vocaux de longue durée (d'une trentaine de secondes à plusieurs minutes). Nous avons participé à une étude (Larcher et al., 2012a) sur un système de vérification dépendante du texte dont les énoncés sont de courte durée. Elle a notamment montré le bon comportement expérimental de la normalisation EFR suivie du scoring de Mahalanobis. Cette métrique s'est avérée efficace lorsqu'elle a modélisé dans la matrice **W** la variabilité interne des sous-classes croisées (locuteur + énoncé phonétique). Le prolongement de cette étude (Larcher et al., 2013) a montré l'efficacité de la normalisation Spherical Nuisance et du modèle PLDA dans le contexte de la vérification dépendante du texte. L'algorithme EFR et le scoring de Mahalanobis dans le contexte de la segmentation en locuteurs (Rouvier and Meignier, 2012a) (Rouvier and Meignier, 2012b) ont produit des performances intéressantes. La capacité du scoring de Mahalanobis dans un contexte de classifieur a, là encore, été confirmée.

Nous présentons ici deux études réalisées pour mettre en place et adapter des systèmes de vérification basés sur les i-vectors dans des contextes particuliers : le cas où la décision doit être prise avec tout ou partie d'énoncés de courte durée (moins de 20 secondes) et le cas où les énoncés de voix à comparer (enrôlement vs test) diffèrent d'un facteur binaire connu a priori (durée longue vs durée courte, microphone vs téléphone, langage A vs langage B, ...).

Nous validons pour le premier cas des hypothèses de complémentarité des bases d'acquisition de connaissances composées d'énoncés de courte et/ou longue durée, ce dans le cadre de la modélisation par pré-normalisation et PLDA.

Pour le deuxième cas, nous présentons une adaptation bayesienne du modèle two-

covariance présenté en 2.6.4 qui prend en compte le facteur de non-correspondance entre les énoncés. Nous montrons que cette sophistication probabiliste affine la précision d'un modèle déterministe et ouvre des perspectives pour une adaptation de la PLDA à ce contexte particulier.

9.2 Application des i-vectors sur les énoncés de courte durée

De nombreuses applications de reconnaissance automatique de la voix imposent des contraintes sur la quantité d'informations utilisables pour l'apprentissage et le test. Ainsi, la vérification du locuteur doit être alors effectuée par comparaison d'énoncés de courtes durées, ou d'énoncés de durées significativement différentes (des énoncés de longue durée sont disponibles pour mettre en place des modèles-locuteur mais les énoncés à tester sont de courte durée).

Nous classerons dans la suite de cette partie comme de courte durée des énoncés dont la durée n'excède pas 20 secondes. Deux cas de reconnaissance du locuteur sont envisagés :

i) Des énoncés de longue durée des locuteurs-cible sont disponibles (enrôlements de longue durée), mais les énoncés à évaluer sont de courte durée (tests de courte durée).

ii) L'ensemble des énoncés disponibles et de ceux à évaluer sont de courte durée.

Dans le premier cas, les durées des observations d'enrôlement et de test différent significativement. Dans le second, elles correspondent en durée, mais souffrent de la contrainte de leur brièveté. Nous présentons ici le détaillé des analyses et résultats synthétisés dans l'article (Sarkar et al., 2012) auquel nous avons contribué.

Les systèmes de reconnaissance du locuteur basés sur les modèles de mixture de gaussiennes ont été conçus pour des énoncés de longue durée. La constitution de supervecteurs à partir des trames acoustiques et d'une couverture de l'espace acoustique par un nombre important de composantes gaussiennes (512 à 4096) n'est réellement adaptée qu'aux cas d'énoncés constitués, après VAD, d'un nombre conséquent de trames. Lorsque ce nombre est réduit (quelques centaines à moins de 2000 pour des énoncés courts), l'estimation d'une moyenne par gaussienne à partir d'un faible effectif de trames est peu fiable : pour 1000 trames et 512 gaussiennes par exemple, l'effectif moyen n'est que de 2. Comme tout modèle paramétrique, le GMM-UBM nécessite des quantités suffisantes d'informations pour ajuster une distribution empirique à un modèle théorique a priori.

Nous avons cherché à mettre en place, depuis la modélisation par GMM-UBM, un ensemble de systèmes capables de circonscrire la variabilité nuisible des vecteurs et d'exalter la variabilité seulement dépendante du locuteur.

Nous avons émis plusieurs hypothèses qui doivent être confirmées expérimentalement :

– Intuitivement, un énoncé de courte durée est pauvre en information propre au locuteur. Nous présentons dans la suite des mesures de variabilité effectuées sur des énoncés courts, puis longs, qui valident ce constat.
– La seconde intuition qui nous a guidés est la suivante : ayant à effectuer une tâche de reconnaissance du locuteur à partir d'énoncés de courte durée, l'élaboration de modèles d'extraction et de scoring des i-vectors, tels que ceux de la FA-total-var et de la PLDA, peut s'opérer à partir d'énoncés similaires. Une base d'énoncés de courte durée sera constituée à cet effet. Mais la constitution d'un modèle à partir d'énoncés de plus longue durée, que nous supposons plus riches en information propre au locuteur, doit alors nécessairement complémenter cette base d'énoncés courts. Nous entendons par là que cette deuxième base est capable de modéliser des spécificités-locuteur qui échappent à la première. La base initiale reste utile, seuls les modèles constitués à partir d'énoncés courts permettant de prendre en compte la forte variabilité nuisible de ces énoncés. La combinaison des acquis de ces différentes modélisations doit donc améliorer la qualité de l'application de reconnaissance.
– Enfin, les modèles appris sur une réunion d'énoncés courts et longs méritent d'être analysés. En particulier, la procédure de normalisation SphN peut combattre un décalage entre les classes-locuteur des deux sous-ensembles de leurs i-vectors.

Afin de vérifier expérimentalement ces hypothèses, nous avons mis en place un ensemble de systèmes dont les apprentissages, de la matrice **T** de la FA-total-var et des paramètres de la PLDA, sont basés sur des énoncés de durée distincte. D'autre part, nous avons effectué des mesures de variabilité destinées à confirmer la meilleure aptitude des énoncés longs à modéliser l'information propre au locuteur et celle des énoncés courts à modéliser la variabilité nuisible.

Pour réaliser cette étude, il faut d'abord circonscrire les variantes possibles dans l'ajout d'informations issues d'énoncés longs à un système incluant des courts. Les différents systèmes possibles sont les suivants :

L'apprentissage du modèle PLDA peut être effectué à partir d'énoncés de durée :

– courte seulement,
– longue seulement,
– courte et longue, réunion des deux jeux de données précédents.

D'autre part, plusieurs types d'évaluation sont possibles. Elles peuvent porter sur des énoncés de durées :

– enrôlement : courte - test : courte
– enrôlement : courte - test : longue
– enrôlement : longue - test : courte

Pour nos expériences, nous avons utilisé le jeu de i-vectors d'apprentissage LIA-hommes décrit en annexe. Les énoncés de courte durée, d'apprentissage et d'évaluation, ont été formés par sélection d'un nombre fixé de trames énergisées après VAD. Les vecteurs de trames sont normalisés à une moyenne de 0 et une variance unité. Les

Initial (avant normalisation)		
Données	Variance totale	R^2
court 5sec	15.508	11.7%
court 10 sec	4.161	14.5%
court 20 sec	2.026	18.5%
long	0.199	34.0%

TABLE 9.1 – *Analyse des diverses variances selon les durées des segments d'apprentissage (données homogènes).*

nombres fixés de trames que nous avons considérés sont 500, 1000 et 2000, simulant des durées de 5, 10 et 20 secondes. L'extraction de i-vectors a été réalisée chaque fois pour les apprentissages "court", "long" et "court + long".

Les hypothèses sur les meilleures aptitudes des différents systèmes à modéliser certaines variabilités sont évaluées dans le paragraphe suivant.

9.2.1 Analyse de variance

Pour chacun des systèmes i-vectors basés sur les jeux d'apprentissage précédents (court, long ou court+long), nous avons calculé les variances inter et intra locuteur. Ces valeurs sont des indicateurs de la dispersion moyenne entre les classes-locuteur d'apprentissage et à l'intérieur de leurs classes. Nous rappelons que la variance totale est égale à leur somme. Le rapport de la variance inter-locuteur à la variance totale (noté R^2) permet d'apprécier la capacité a priori du modèle PLDA à discriminer les énoncés de locuteurs distincts.

Dans un premier temps, ces valeurs sont obtenues sur le jeu de données initial, immédiatement après extraction et donc avant toute forme de normalisation.

La table 9.1 affiche les valeurs obtenues pour des systèmes utilisant des données d'apprentissage de différentes durées. La variance totale décroît avec la durée de ses énoncés d'apprentissage. Ainsi, les i-vectors extraits à partir d'énoncés basés sur des durées courtes présentent une très grande variabilité. Mais cette variabilité est essentiellement non-informative. Le taux d'information locuteur dans chaque système croît avec la durée des énoncés, de 11.71 % pour les 5 secondes à 34.09 % pour les longues durées. Ces résultats confirment la première de nos hypothèses : les énoncés de courte durée s'avèrent pauvres en information propre au locuteur.

Pour le cas de systèmes basés sur un apprentissage mixte (court + long), les valeurs suivantes ont été calculés : variance totale et R^2 de l'apprentissage complet, du sous-ensemble des données courtes et du sous-ensemble des données longues. La table 9.2 affiche ces valeurs. Elle permet d'apprécier les distributions des énoncés courts et longs à travers leur représentation dans un espace i-vector commun.

Les énoncés courts sont beaucoup plus variés que les énoncés longs (par exemple, pour court 5 sec + long, leur variance totale est de 4.704 contre 0.803). Mais leur R^2 est

Initial (avant normalisation)		
Données	Variance totale	R^2
court 5sec + long		
court et long	2.757	12.1%
court seulement	4.704	12.5%
long seulement	0.803	31.2%
court 10sec + long		
court et long	1.523	16.9%
court seulement	2.369	15.6%
long seulement	0.674	31.4%
court 20sec + long		
court et long	0.664	23.1%
court seulement	0.905	20.8%
long seulement	0.421	32.3%

TABLE 9.2 – *Analyse des diverses variances selon les durées des segments d'apprentissage (données mixtes).*

Après normalisation Sph.Nuisance		
Données	Variance totale	R^2
court 5sec	1	18.32%
court 10 sec	1	24.31%
court 20 sec	1	32.19%
long	1	49.97%

TABLE 9.3 – *Analyse, après normalisation SphN, des diverses variances selon les durées des segments d'apprentissage (données homogènes).*

par contre beaucoup plus faible (toujours pour court 5 sec + long, 12.5% contre 31.2%). Les mêmes constats se dégagent des valeurs pour 10 et 20 secondes, le phénomène tendant à s'atténuer. Ils confirment le constat précédent du manque d'information locuteur dans les énoncés courts. L'apprentissage de paramètres à partir d'énoncés courts, s'il utilise des données de type concordant, modélise essentiellement la variabilité nuisible.

Avant d'aborder la question du ou des systèmes à adopter pour gérer la reconnaissance du locuteur à partir d'énoncés de courte durée, nous présentons la même analyse de variance effectuée sur des vecteurs préalablement normalisés.

Après normalisation

La table 9.3 présente les mêmes calculs de variance que ceux de la table 9.1, après deux itérations de l'algorithme de normalisation Spherical- Nuisance sur les fichiers d'apprentissage. La normalisation uniformisant les variances totales, celles-ci sont toutes égales à 1. Le R^2 est plus élevé après normalisation, sur toutes les durées testées. La capacité discriminante des systèmes s'est donc accrue. Mais la part de variabilité locuteur

Après normalisation Sph.Nuisance		
Données	Variance totale	R^2
court 5sec + long		
court et long	1	28.99%
court seulement	0.999	19.24%
long seulement	0.999	48.22%
court 10sec + long		
court et long	1	34.43%
court seulement	0.999	25.8%
long seulement	0.999	48.77%
court 20sec + long		
court et long	1	40.13%
court seulement	1	33.79%
long seulement	1	49.24%

TABLE 9.4 – *Analyse, après normalisation SphN, des diverses variances selon les durées des segments d'apprentissage (données mixtes).*

des apprentissages constituées de court seulement reste faible (R^2 égal à 18.32% pour 5 secondes).

La table 9.4 présente les mêmes calculs de variance que ceux de la table 9.2, après deux itérations de l'algorithme de normalisation. Les variances totales des sous-ensembles de données "courts seulement" ou "longs seulement" sont approximativement égales à 1. La normalisation a égalisé leurs variances à la variance totale, ce qui signifie que ces sous-ensembles de i-vectors ne présentent plus de divergence [1].

9.2.2 Combinaison des systèmes

Les analyses de variance précédentes montrent que, ayant à effectuer une tâche de reconnaissance du locuteur à partir d'énoncés de test de courte durée :

- l'inclusion de données de longue durée dans l'apprentissage des modèles de la FA-total-var et de la PLDA peut enrichir le système,
- la normalisation de type SphN contribue à homogénéïser les observations de différentes durées,
- la forte variabilité nuisible des énoncés courts est correctement modélisée par un système à apprentissage basé sur du court. Par contre, la variabilité inter-locuteur est alors mal apprise, de par cette forte instabilité des énoncés courts.A l'opposé, les systèmes à apprentissage basé sur du long parviennent à modéliser précisé-

1. En l'occurrence des localisations globalement distinctes sur la sphère de représentation. Pour mieux appréhender ce fait, rappelons que l'écart entre les centres des classes "court seulement" et "long seulement" est ici égal à $\sqrt{1-(0.999+0.999)/2} \approx 3 \times 10^{-2}$, soit une valeur négligeable au regard du rayon 1 de la sphère.

ment la variabilité inter-locuteur, mais minimisent alors la part de variabilité nuisible dans l'énoncé court de test.

Aucun des systèmes inventoriés ne peut donc modéliser précisément ces énoncés courts : les estimations de leur variabilité résiduelle pêchent forcément par excès ou par défaut. A cela s'ajoute, dans le cas des évaluations mixtes (longue vs courte durée), la difficulté supplémentaire d'une modélisation commune, nécessairement sous-optimale.

Nous avons émis l'hypothèse que ces différents systèmes étaient nécessairement complémentaires. Nous entendons par ce terme qu'accordant chacun des parts trop fortes ou trop faibles aux variabilités locuteur et nuisible, leurs décompositions par la PLDA présenteront nécessairement une part d'indépendance. Ainsi la synthèse de ces systèmes améliorera la qualité du système discriminant. La validité de ce postulat doit être confirmée par l'expérimentation. Nous décrivons dans le paragraphe suivant les expériences mises en place pour estimer le degré de validité de cette hypothèse.

9.2.3 Expérimentation

Nous avons effectué un certain nombre d'expériences, utilisant toujours le même jeu de données.

- **Eval : 5s-5s** : l'évaluation porte sur des énoncés d'enrôlement et tests de courte durée, de 500 trames,
- **Eval : long-5s** : l'évaluation porte sur des énoncés d'enrôlement de longue durée et de tests de 500 trames,

et de même pour 10 secondes :

- **Eval : 10s-10s** : l'évaluation porte sur des énoncés d'enrôlement et tests de courte durée, de 1000 trames,
- **Eval : long-10s** : l'évaluation porte sur des énoncés d'enrôlement de longue durée et de tests de 500 trames,

A chaque fois, trois jeux de i-vectors sont engendrés :

- **Apprentissage 5s** : la matrice T de variabilité totale utilisée par l'extraction EM-FA n'a été apprise qu'à partir d'énoncés de 500 trames. Les i-vectors issus de ces énoncés sont utilisés comme apprentissage pour la technique de normalisation et la PLDA.
- **Apprentissage long** : même chose, mais seulement à partir d'énoncés longs.
- **Apprentissage (5s et long)**: la matrice T a été apprise à partir de la réunion des énoncés courts (500 trames) et longs précédentes. Même chose ensuite, pour les apprentissages dans l'espace des i-vectors.

et de même pour 10 secondes (1000 trames) au lieu de 5.

EER %	Apprentissage			*Fusion (addition)*
	5s	*long*	*(5s et long)*	
Eval : 5s-5s	18.79	19.81	17.81	**15.26**
	10s	*long*	*(10s + long)*	
Eval : 10s-10s	10.70	10.70	9.40	**8.14**

TABLE 9.5 – *Performances, en terme d'EER, des évaluations de type évaluations "court vs court" (enrôlement vs test).*

EER %	Apprentissage			*Fusion (addition)*
	5s	*long*	*(5s et long)*	
Eval : long-5s	15.72	**9.34**	10.24	9.55
	10s	*long*	*(10s et long)*	
Eval : long-10s	9.11	5.32	5.62	**5.01**

TABLE 9.6 – *Performances, en terme d'EER, des évaluations de type évaluations mixtes "long vs court" (enrôlement vs test).*

Les tables 9.5 et 9.6 montrent les résultats, en terme d'EER de ces différentes expérimentations. Les tests réalisés sont ceux de la condition *det 7* de NIST-SRE 2008. Le rang de la matrice session $\mathbf{\Gamma}$ de la PLDA est fixé à 400 (rang maximal). A chaque fois, est indiqué le meilleur EER obtenu en variant les rangs de la matrice locuteur $\mathbf{\Phi}$ de 50 à 400 par pas de 50.

La dernière colonne des deux tables mesure la validité de l'hypothèse précédente, sur les scores obtenus par les différents systèmes. Leur complémentarité est testée par fusion. Pour procéder, les scores des trois systèmes individuels, pour l'ensemble des rangs de la matrice locuteur $\mathbf{\Phi}$ de 50 à 400 par pas de 50 ont été fusionnés par leur simple moyenne (à pondérations égales) [2].

La première table (9.5) affiche les résultats des évaluations dans lesquels les énoncés d'enrôlement et de test sont de courte durée (5 ou 10 secondes). Si l'apprentissage obtenu par réunion de court et long procure un gain de performance (17.81% et 9.40%) par rapport à des systèmes à durée unique, la combinaison des trois systèmes par fusion de leurs scores s'avère significativement plus performante que chacun d'entre eux (15.26% et 8.14%). Ces résultats confirment notre hypothèse, sur la complémentarité de ces systèmes.

La seconde table (9.5) affiche les résultats d'évaluations mixtes : les énoncés d'enrôlement sont de longue durée mais ceux de test de courte durée (5 ou 10 secondes). Cette fois, le second système, basé sur un apprentissage par énoncés de longue durée, égale la performance de la fusion des scores des trois systèmes (9.34% vs 9.55% pour 5 secondes, 5.32% vs 5.01% pour 10 secondes). Ce résultat ne remet pas en question

2. Nous avons inclus dans la somme l'ensemble des scores de chaque système individuel et non ceux de leur seul meilleur rang, d'une part pour ne pas réduire la robustesse des résultats par l'optimisation de paramètres, d'autre part pour inclure dans le score les détaillés des décompositions en facteurs de la PLDA.

l'hypothèse de complémentarité : les systèmes basés sur des apprentissages incluant des courtes durées (1ère et 3ème colonnes de la table) dégradent la performance (pour 5 secondes : 15.72% et 10.24% vs 9.34% pour le système longue durée seulement, pour 10 secondes : 9.11% et 5.62% vs 5.323%). La complémentarité permet pourtant à la fusion de ces trois systèmes de rejoindre le meilleur système individuel.

En gardant à l'esprit que les énoncés d'enrôlement sont de longue durée, ces dernières expériences montrent également que l'ajout en modélisation de données plus pauvres en information cible n'améliore pas nécessairement la qualité d'un système.

9.2.4 Complément : échelonnement des durées d'apprentissage

Les expériences précédentes ont montré la complémentarité des systèmes basés sur une hérérogénéïté des durées. Pour traiter les évaluations sur des données courtes seulement ("5s-5s" et "10s-10s" ci-dessus), l'apport d'un apprentissage long à l'apprentissage court ("5s") a été déterminant.

Il est possible d'analyser plus finement l'apport de données plus riches en information explicative à un système basé sur les durées courtes. Nous avons réalisé l'expérience suivante :

Pour l'évaluation 5s-5s, c'est à dire cibles et tests de 500 trames, nous avons élaboré plusieurs apprentissages :

- **5s** : la matrice **T** de variabilité totale utilisée par l'extraction EM-FA a été apprise à partir d'énoncés de 500 trames. Les i-vectors issus de ces énoncés sont utilisés comme apprentissage pour la technique de normalisation et la PLDA.
- **10s** : même procédure avec des énoncés d'apprentissage de 1000 trames.
- **20s** : même procédure avec des énoncés d'apprentissage de 2000 trames.
- **long** : même procédure avec les énoncés d'apprentissage de longue duré (segments complets).

Sur chacun de ces modèles, les i-vectors d'évaluation de 500 trames ont été extraits à partir des métaparamètres de la FA obtenus. Les scores de l'évaluation ont été alors calculés, pour les quatre systèmes.

Enfin, nous avons fusionné ces scores par simple moyenne (fusion équipondérée, donc sans aucune optimisation des coefficients par une technique quelconque comme la régression logistique), de manière échelonnée :

- **5s** : scores issus du système 5s seul,
- **5s + 10s** : addition des scores issus des apprentissages 5s et 10s,
- **5s + 10s +20s** : addition des scores issus des apprentissages 5s, 10s et 20s,
- **5s + 10s +20s + long** : addition des scores issus des quatre systèmes,

ce afin d'analyser les apports progressifs de données enrichies en information locuteur par leur durée. La table 9.7 indique les EER obtenus par cette expérience.

EER %	**Apprentissages**			
Eval : 5s-5s	5s	5s + 10s	5s+10s+20s	5s+10s+20s+long
SphN + GPLDA	19.25	16.17	15.04	14.81 (*13.66*)

TABLE 9.7 – *Performances, en terme d'EER, d'évaluations basées sur des réunions de fichiers d'apprentissages échelonnés, dans la durée de leurs segments.*

La performance progresse strictement au fil des étapes de l'échelonnement et aboutit à un EER (14.81 %) meilleur que celui obtenant précédemment (table 9.5). Les apprentissages de taille intermédiaires (10s et 20s) semblent former un compromis dans la captation de l'information locuteur, utile à un système de discrimination.

9.2.5 Conclusion

L'étude que nous avons conduite montre l'intérêt, pour des applications de reconnaissance du locuteur travaillant sur des énoncés de courte durée, d'inclure aux bases d'apprentissage des modèles d'énoncés de plus longue durée. L'analyse de variance a montré que ces derniers contenaient une part plus forte d'information propre au locuteur. D'autre part, les expériences précédentes ont confirmé notre hypothèse de complémentarité de systèmes basés sur des apprentissages de durée variable : leur combinaison, dans notre étude par la simple moyenne des scores, permet de mieux gérer le problème de la discrimination d'énoncés pauvres en information locuteur. Les modèles de décomposition en facteurs locuteur et nuisance tels que la PLDA tendent soit à estimer par défaut la part de variabilité nuisible, soit à produire un facteur locuteur trop peu varié pour une discrimination efficace.

La procédure de normalisation dans l'espace des i-vectors s'avère utile à conditionner des données de durée hétérogéne. Les représentations i-vectors des énoncés courts ou longs sont transportés sur la surface de la sphère en sorte que les sous-classes locuteur d'énoncés courts et longs ne présentent plus de divergence nette en terme de variance.

Les expériences précédentes ont employé des bases d'apprentissage constituées chaque fois à partir de deux jeux de données : l'un formé d'énoncés courts dont l'effectif de trames, après VAD, est faible et fixé (par exemple 500), l'autre d'énoncés longs (effectif supérieur à 5000). Il est possible de renseigner plus finement un système, en poussant plus loin la logique de complémentarité des durées variables. Nous avons mené une expérience qui combine cette fois-ci les scores de systèmes à apprentissage de durées échelonnées (5 secondes, 10 secondes, 20 secondes et longue durée), toujours par leur moyenne équipondérée. Sur l'évaluation 5sec-5sec de la table 9.5, l'EER s'améliore, de 15.26% à 14.81%. Ce résultat confirme la pertinence de nos hypothèses et l'intérêt de la démarche.

Celle-ci présente toutefois ses limites : l'élaboration d'une représentation par supervecteurs de collections de quelques centaines de trames, à partir d'un GMM-UBM à 512 composantes tel que celui utilisé dans nos expériences, souffre d'un manque de

précision. Les moyennes par gaussienne sont estimées à partir d'effectifs par trame très réduits. La solution par mixtures de gaussiennes se trouve alors aux limites de son champ de validité et des alternatives non ou semi paramétriques sont envisageables.

9.3 Extension bayésienne du modèle two-covariance aux applications mixtes

La section précédente a proposé des stratégies de prise en compte d'énoncés de voix de durée variable. Les applications comparant des énoncés d'enrôlement-cible de longue durée à des énoncés de test courts seront qualifiées d'applications *mixtes*.

Les modèles i-vectors tels que la PLDA estiment des métaparamètres de variabilité inter et intra locuteurs. Lorsque les énoncés de voix à comparer (enrôlement vs test) diffèrent d'un facteur binaire connu **a priori** (durée longue vs durée courte, microphone vs téléphone, langage A vs langage B), les métaparamètres de chacun de ces deux types ne sont pas nécessairement égaux. La partie précédente a montré que tel était le cas pour un facteur durée à deux types court et long. S'il est aisé d'estimer ces métaparamètres indépendamment pour chacun des deux types d'énoncés, la difficulté à surmonter alors est d'adapter les formules de scoring à cette modélisation à quatre métaparamètres de covariance. Cette adaptation doit notamment permettre de relier par un pont probabiliste les deux distributions. Nous présentons ici l'extension du modèle et scoring two-covariance à des applications mixtes, que nous avons présenté dans (Bousquet et al., 2012b).

9.3.1 Le modèle 4-covariance

Le modèle two-covariance est basé sur l'hypothèse que les énoncés d'enrôlement et de test pour un locuteur donné ont la même densité de distribution de l'équation 2.61 et que tous les facteurs locuteurs ont la même densité de distribution de l'équation 2.62. Dans le cas d'expériences mixtes, nous postulons que le type 1 des i-vectors d'enrôlement et le type 2 des i-vectors de test ont des distributions intra et inter-locuteur distinctes. Etant donné un locuteur s et pour toute observation w de s, le modèle two-covariance devient :

$$\left|\begin{array}{l} \text{- si } w \text{ appartient à la classe 1 : } P(y_1) = \mathcal{N}(\mu_1, \mathbf{B}_1) \text{ et } P(w|s) = \mathcal{N}(y_1, \mathbf{W}_1) \\ \text{- si } w \text{ appartient à la classe 2 : } P(y_2) = \mathcal{N}(\mu_2, \mathbf{B}_2) \text{ et } P(w|s) = \mathcal{N}(y_2, \mathbf{W}_2) \end{array}\right.$$

où pour $i = 1$ ou 2, y_i est le facteur locuteur du type i, μ_i la moyenne des observations du type i et $\mathbf{B}_i$, $\mathbf{W}_i$ les matrices de covariance inter- et intra-locuteur du type i. Les paramètres $\{\mu_1, \mathbf{B}_1, \mu_2, \mathbf{B}_2\}$ sont calculées à partir de données d'apprentissage de l'un et l'autre type. Nous appellerons "4-covariance" ce modèle étendu du modèle two-covariance.

9.3.2 Adaptation du scoring au modèle 4-covariance

Il est nécessaire de reformuler le score two-covariance par log-vraisemblance d'hypothèses complémentaires. Nous présentons ici les calculs abrégés, leur détail pouvant être consulté en annexe F.

Non-cible

La mesure du score non-cible de l'équation 2.36 devient dans le cas du modèle 4-covariance :

$$\log P\left(w_1, w_2|\theta_{\text{non}}\right) = \int P\left(w_i, w_2, y_1, y_2\right) dy_1 dy_2 \tag{9.1}$$

L'indépendance supposée des observations w_1 et w_2 entraîne que :

$$\begin{aligned}\log P\left(w_1, w_2|\theta_{\text{non}}\right) &= \log P\left(w_1\right) P\left(w_2\right) \\ &= \log \int P\left(w_1|y, \mathbf{W}_1\right) P\left(y|\mu_1, \mathbf{B}_1\right) dy - \log \int P\left(w_2|y, \mathbf{W}_2\right) P\left(y|\mu_2, \mathbf{B}_2\right) dy\end{aligned} \tag{9.2}$$

La solution explicite est :

$$\begin{aligned}\log P\left(w_1, w_2|\theta_{\text{non}}\right) = K &+ w_1^t \mathbf{N}_{11}^{\text{non}} w_1 + w_2^t \mathbf{N}_{22}^{\text{non}} w_2 \\ &+ n_1^{\text{non}} w_1 + n_2^{\text{non}} w_2\end{aligned} \tag{9.3}$$

où K est une constante et, pour $i = 1, 2$,

$$\begin{cases} \mathbf{N}_{ii}^{\text{non}} = \mathbf{W}_i^{-1} \left(\mathbf{B}_i^{-1} + \mathbf{W}_i^{-1}\right)^{-1} \mathbf{W}_i^{-1} - \mathbf{W}_i^{-1} \\ n_i^{\text{non}} = 2 \left(\mathbf{B}_i^{-1} \mu_i\right)^t \left(\mathbf{W}_i^{-1} + \mathbf{W}_i^{-1}\right)^{-1} \mathbf{W}_i^{-1} \end{cases} \tag{9.4}$$

Cible

Formulation : La mesure du score cible de l'équation 2.35 devient dans le cas du modèle 4-covariance :

$$P\left(w_1, w_2|\theta_{\text{tar}}\right) = \iint P\left(w_1, w_2, y_1, y_2\right) dy_1 dy_2 \tag{9.5}$$

Cette log-vraisemblance peut être réécrite :

$$P(w_1, w_2|\theta_{\text{tar}}) = \iint P(w_1, w_2|y_1, y_2) P(y_1, y_2) dy_1 dy_2$$
$$= \iint P(w_1|y_1, y_2) P(w_2|y_1, y_2) P(y_1, y_2) dy_1 dy_2 \quad (9.6)$$

en utilisant l'indépendance de w_1 et w_2 sachant y_1 et y_2. Or

$$P(w_1|y_1, y_2) = \frac{P(w_1, y_1, y_2)}{P(y_1, y_2)} = \frac{P(w_1, y_1|y_2) P(y_2)}{P(y_1, y_2)}$$
$$= \frac{P(w_1|y_2) P(y_1|y_2) P(y_2)}{P(y_1, y_2)} = P(w_1|y_2) \quad (9.7)$$

qui confirme l'indépendance de w_1 et y_1 sachant y_2, d'où

$$P(w_1, w_2|\theta_{\text{tar}}) = \iint P(w_1|y_1) P(w_2|y_2) P(y_2|y_1) P(y_1) dy_1 dy_2 \quad (9.8)$$

Relation entre les facteurs locuteurs Le problème restant à surmonter est de former une relation probabiliste entre les facteurs locuteurs y_1, y_2 des deux classes. Un cadre probabiliste pour le facteur $P(y_2|y_1)$ de l'équation précédente doit être mis en place. Nous avons exploré deux voies. Elles ont en commun de nécessiter la disponibilité d'un jeu de données multi-locuteur et multi-types contenant pour chacun de ses locuteurs des énoncés des deux types. L'effectif en nombre de locuteurs de ce jeu mixte doit être supérieur ou égal à la dimension de l'espace des i-vectors, pour obtenir des matrices inversibles. Pour chaque locuteur pour lequel on dispose d'énoncés des deux types, nous calculons la moyenne de ses énoncés de type 1, notée y_1 puis de ses énoncés de type 2, notée y_2. L'ensemble des facteurs y_1 et y_2 des locuteurs d'apprentissage forment deux ensembles appariés notés Y_1 et Y_2.

i) Régression linéaire :

Une première voie consiste à appliquer une régression linéaire entre les deux jeux Y_1 et Y_2 de facteurs locuteur des deux types. L'hypothèse de relation linéaire entre les deux facteurs d'un même locuteur $(y_1, y_2) \in Y_1 \times Y_2$ s'écrit :

$$y_2 = \mu_2 + \mathbf{Å}(\mathbf{y}_1 - \mu_1) + \varepsilon \quad (9.9)$$

où μ_1 et μ_2 sont les moyennes globales des classes 1 et 2 et $\mathbf{Å}$ est la matrice $p \times p$ (p dimension de l'espace) calculée en régression linéaire par optimisation suivant les moindres carrés. Nous rappelons que le résidu ε n'est pas nécessairement gaussien dans un modèle de régression.

Soient $\overline{Y_1}$ et $\overline{Y_2}$ les matrices $S \times p$ centrées des facteurs locuteurs, où S est l'effectif de locuteurs multi-classes dans l'apprentissage. L'estimation par les moindres carrés $\widehat{\mathbf{A}}$ de $\mathbf{\mathring{A}}$ est la projection orthogonale de $\overline{Y_2}$ sur le sous-espace engendré par $\overline{Y_1}$, égale à :

$$\widehat{\mathbf{A}} = \left(\overline{Y_1}^t \overline{Y_1}\right)^{-1} \overline{Y_1}^t \overline{Y_2} \tag{9.10}$$

L'orthogonalité implique que $\overline{Y_1}\widehat{A}$ et le résidu ε sont indépendants et donc, supposant la gaussianité du résidu, la relation probabiliste entre les locuteurs facteurs est :

$$P\left(y_2|y_1\right) = \mathcal{N}\left(y_2 \mid \mu_2 + \widehat{\mathbf{A}}\left(y_1 - \mu_1\right), cov\left(\overline{Y_2} - \overline{Y_1}\widehat{\mathbf{A}}\right)\right) \tag{9.11}$$

La qualité de l'estimation par régression linéaire peut être mesurée par le coefficient de détermination R^2, égal au rapport de la variance expliquée par la variance totale :

$$R^2 = \frac{var\left(\overline{Y_2}\right)}{var\left(\overline{Y_1}\widehat{\mathbf{A}}\right)} \tag{9.12}$$

où la variance peut être calculée par la trace de la matrice de covariance. L'estimation étant orthogonale, le R^2 est nécessairement compris entre 0 et 1.

Sur les jeux d'apprentissage des expériences présentées au paragraphe suivant d'expérimentation, les R^2 obtenus ont tous dépassé 0.75. Le degré de linéarité entre Y_1 et Y_2 est suffisant pour appliquer l'équation 9.9.

ii) Fonction gaussienne de transfert :

La relation précédente présente l'avantage de relier les facteurs locuteurs avec une perte minimale en terme de distance euclidienne mais la gaussianité du résidu n'est pas garantie. Pour pallier cet inconvénient, une relation inférentielle basée sur les distributions empiriques déduites de l'apprentissage (et non sur les relations ponctuelles entre chacun de ses couples) peut sembler plus robuste. Nous utilisons les distributions gaussiennes estimées $\mathcal{N}\left(\mu_1, \mathbf{B}_1\right)$ et $\mathcal{N}\left(\mu_2, \mathbf{B}_2\right)$ issues des jeux d'apprentissage Y_1 et Y_2.

La figure 9.1 montre une image en 2 dimensions de ces distributions. Pour passer de l'une à l'autre, la fonction de transfert standardise d'abord les facteurs de la 1$^{\text{ère}}$ classe pour les normaliser, puis les re-déploie suivant la distribution de la classe 2 par l'opération de standardisation réciproque. Cette relation :

$$\mathbf{B}_2^{\frac{1}{2}}\left(y_2 - \mu_2\right) \leftrightarrow \mathbf{B}_1^{-\frac{1}{2}}\left(y_1 - \mu_1\right) \tag{9.13}$$

permet de proposer un modèle de transfert :

$$y_2 = \mu_2 + \mathbf{B}_2^{\frac{1}{2}}\mathbf{B}_1^{-\frac{1}{2}}\left(y_1 - \mu_1\right) + \varepsilon \tag{9.14}$$

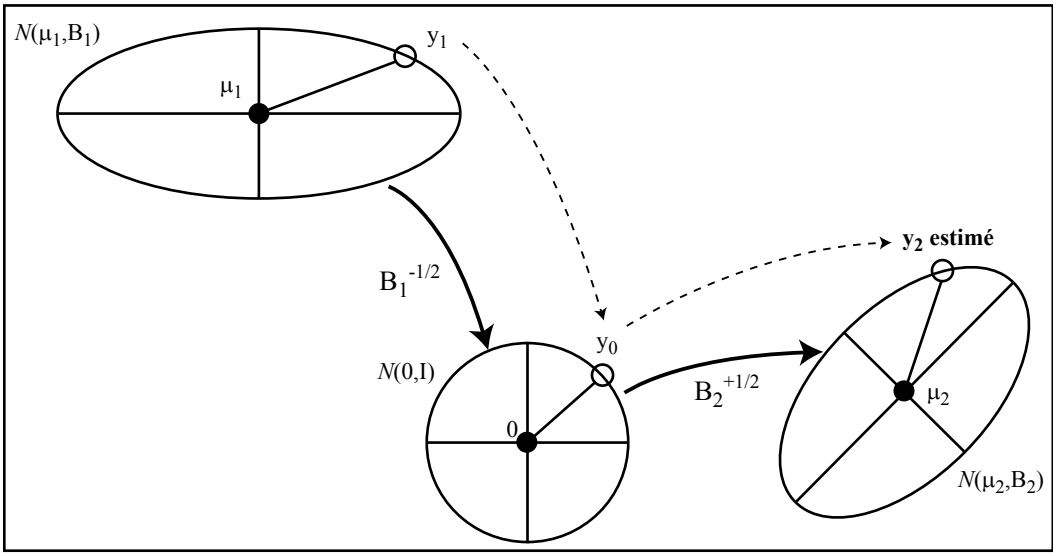

FIGURE 9.1 – *Fonction de transfert gaussienne entre les deux distributions des facteurs locuteur.*

L'erreur ε peut être alors estimée en calculant sa covariance sur les données d'apprentissage appariées et centrées $\overline{\mathbf{Y}_1}$ et $\overline{\mathbf{Y}_2}$:

$$cov\left(\varepsilon\right) = cov\left(\overline{Y_2} - \overline{Y_1}\mathbf{B}_2^{\frac{1}{2}}\mathbf{B}_1^{-\frac{1}{2}}\right) \tag{9.15}$$

Sous l'hypothèse de gaussianité de ce résidu, la relation probabiliste entre les locuteurs facteurs devient :

$$P\left(y_2|y_1\right) = \mathcal{N}\left(y_2 \mid \mu_2 + \mathbf{B}_2^{\frac{1}{2}}\mathbf{B}_1^{-\frac{1}{2}}\left(y_1 - \mu_1\right), cov\left(\overline{Y_2} - \overline{Y_1}\mathbf{B}_2^{\frac{1}{2}}\mathbf{B}_1^{-\frac{1}{2}}\right)\right) \tag{9.16}$$

Notons qu'ici, l'absence d'orthogonalité entre les variabilités expliquée et résiduelle rend inopérante la mesure de qualité précédente par le R^2.

Solution explicite Dans les deux cas d'estimation précédents, la relation entre y_2 et y_1 s'écrit sous une forme commune, à la matrice de transfert près. Etant donnée cette matrice que nous noterons par commodité $\widehat{A}$ (qu'elle ait été obtenue par régression linéaire ou par fonction gaussienne de transfert), la solution explicite est alors la même dans les deux cas. Les calculs détaillés sont présentés en annexe F. La log-vraisemblance sous l'hypothèse θ_{tar} s'écrit :

$$\begin{aligned}\log P\left(w_1, w_2|\theta_{\text{tar}}\right) = K &+ w_1^t\mathbf{N}_{11}^{\text{tar}}w_1 + w_2^t\mathbf{N}_{22}^{\text{tar}}w_2 \\ &+ 2w_1^t\mathbf{N}_{12}^{\text{tar}}w_2 + n_1^{\text{tar}}w_1 + n_2^{\text{tar}}w_2\end{aligned} \tag{9.17}$$

où le scalaire K est une constante et

$$\begin{cases} \mathbf{N}_{11}^{\text{tar}} = \mathbf{W}_1^{-1}\mathcal{C}^t\mathcal{B}^{-1}\mathcal{C}\mathbf{W}_1^{-1} - \mathbf{W}_1^{-1} + \mathbf{W}_1^{-1}\mathcal{A}\mathbf{W}_1^{-1} \\ \mathbf{N}_{22}^{\text{tar}} = \mathbf{W}_2^{-1}\mathcal{B}^{-1}\mathbf{W}_2^{-1} - \mathbf{W}_2^{-1} \\ \mathbf{N}_{12}^{\text{tar}} = \mathbf{W}_1^{-1}\mathcal{C}^t\mathcal{B}^{-1}\mathbf{W}_2^{-1} \\ n_1^{\text{tar}} = -2\left(a - \mathcal{C}b\right)^t\mathcal{B}^{-1}\mathcal{C}\mathbf{W}_1^{-1} + 2b^t\mathcal{A}^{-1}\mathbf{W}_1^{-1} \\ n_2^{\text{tar}} = -2\left(a - \mathcal{C}b\right)^t\mathcal{B}^{-1}\mathbf{W}_2^{-1} \end{cases} \tag{9.18}$$

En posant $\mathbf{M} = cov\left(\overline{Y_2} - \overline{Y_1}\widehat{\mathring{\mathbf{A}}}\right)$, on obtient :

$$\begin{cases} \mathcal{A} = \mathbf{W}_1^{-1} + \mathbf{B}_1^{-1} + \widehat{\mathbf{A}}^t\mathbf{M}^{-1}\widehat{\mathbf{A}} \\ \mathcal{B} = \mathbf{M}^{-1} - \mathbf{M}^{-1}\widehat{\mathbf{A}}\mathcal{A}^{-1}\widehat{\mathbf{A}}^t\mathbf{M}^{-1} + \mathbf{W}_2^{-1} \\ \mathcal{C} = \mathbf{M}^{-1}\widehat{\mathbf{A}}\mathcal{A}^{-1} \\ a = \mathbf{M}^{-1}(\widehat{A}\mu_1 - \mu_2) \\ b = \mathbf{B}_1^{-1}\mu_1 + \widehat{\mathbf{A}}^t\mathbf{M}^{-1}(\widehat{\mathbf{A}}\mu_1 - \mu_2) \\ c = \mu_1^t\mathbf{B}_1^{-1}\mu_1 + (\widehat{\mathbf{A}}\mu_1 - \mu_2)^t\mathbf{M}^{-1}(\widehat{\mathbf{A}}\mu_1 - \mu_2) \end{cases} \tag{9.19}$$

9.3.3 Expérimentation

Nous proposons une étude de performances du modèle 4-covariance sur un exemple de données mixtes pour lequel on dispose de quantités suffisantes d'énoncés multi-classes du même locuteur. Il s'agit du cas de données d'enrôlement de longue durée et de test de courte durée. Les premières nous sont fournies par le jeu de données d'apprentissage décrit au 2.7, auquel nous nous référerons par l'abréviation "long". Les secondes, de courte durée, ont été obtenues en en sélectionnant aléatoirement un nombre désiré de trames, après VAD, des énoncés de longue durée précédents. Les trames sont ensuite normalisées (moyenne 0 et variance unité). Trois types d'énoncés de courte durée ont été ainsi constitués : "5 secondes" qui contiennent 500 trames et simulent des énoncés de 5 secondes ; "10 secondes" (1000 trames) et "20 secondes" (2000 trames). L'évaluation est effectuée avec les conditions téléphone *det 6* et *det 7* de NIST-SRE 2008, dans lesquels les énoncés d'enrôlement ont été maintenus à leur longueur complète mais ceux de test limités au nombre de trames voulu, par le même procédé que précédemment.

Les tables 9.8, 9.9 et 9.10 détaillent chaque fois les performances comparées de six systèmes, en terme d'EER :

- long & 2-cov. : le fichier d'apprentissage utilisé pour l'extraction des i-vectors et les traitements des i-vectors (normalisation et modélisation) est celui constitué des segments longs seulement. Les modèles et scorings two-covariance sont réalisés pour évaluer le jeu d'enrôlement long contre les tests courts (respectivement "5sec", "10 sec" et "20 sec" dans les tables 9.8, 9.9 et 9.10).
- (5s et long) ou (10s et long) ou (20s et long) et 2-cov. : Le même modélisation des i-vectors est appliquée, mais ici le fichier d'apprentissage commun à l'extraction et au traitement des i-vectors est formé de la réunion des énoncés longs et courts.

(a) *enrôlement : longue durée - test : 5 sec*			
Apprentissage	Scoring	det 6 (tél-tél tous) EER%	det 7 (tél-tél anglais) EER%
long	2-cov.	13.17	9.87
(5s et long)	2-cov.	12.68	9.87
long	PLDA	12.92	9.34
(5s et long)	PLDA	12.46	10.24
(5s et long)	4-cov. reg. lin. $\mathbf{Y}_{\text{long}}$~$\mathbf{Y}_{\text{court}}$	11.80	9.66
(5s et long)	4-cov. transf. gauss. $\mathbf{Y}_{\text{long}} \leftrightarrow \mathbf{Y}_{\text{court}}$	**11.78**	**8.96**

TABLE 9.8 – *Pour le cas de tests 5 secondes, comparaison de performances, en terme d'EER, de différents scorings basés sur les modèles LDA-two covariance, PLDA gaussienne et LDA-4covariance, ce dernier avec pont bayesien par régression linéaire ou par fonction de transfert gaussienne.*

(b) *enrôlement : longue durée : 10 sec*			
Apprentissage	Scoring	det 6 (tél-tél tous) EER%	det 7 (tél-tél anglais) EER%
long	2-cov.	9.30	**5.32**
(10s et long)	2-cov.	8.82	5.71
long	PLDA	9.60	**5.32**
(10s et long)	PLDA	**8.69**	5.69
(10s et long)	4-cov. reg. lin. $\mathbf{Y}_{\text{long}}$~$\mathbf{Y}_{\text{court}}$	8.98	5.69
(10s et long)	4-cov. transf. gauss. $\mathbf{Y}_{\text{long}} \leftrightarrow \mathbf{Y}_{\text{court}}$	**8.69**	5.45

TABLE 9.9 – *Pour le cas de tests 10 secondes, comparaison de performances, en terme d'EER, de différents scorings basés sur les modèles LDA-two covariance, PLDA gaussienne et LDA-4covariance, ce dernier avec pont bayesien par régression linéaire ou par fonction de transfert gaussienne.*

- long & 2-cov. : le fichier d'apprentissage des segments longs est utilisé, mais cette fois les modélisation et scoring sont ceux de la PLDA Gaussienne.
- (5s et long) ou (10s et long) ou (20s et long) et PLDA : même modèle mais sur le fichier d'apprentissage du deuxième système.
- (5s et long) ou (10s et long) ou (20s et long) et 4-cov : cette fois, le fichier d'apprentissage obtenu par réunion des segments courts et longs est utilisé pour le modèle proposé 4-covariance. A l'intérieur de ce fichier d'apprentissage, l'ensemble de segments des deux types (court, long) d'un même locuteur alimente le pont bayésien nécessaire au modèle. L'estimation de ces paramètres est effectuée par la régression linéaire de l'équation 9.10.
- Avec les mêmes fichier d'apprentissage et modèle, est expérimenté le pont bayésien basé sur la fonction de transfert de l'équation 9.14.

La modélisation 4-covariance fournit les meilleures performances pour la condition 6 (tél-tél tous) quelle que soit la durée de segments considérée. Il en va de même pour

(c) *enrôlement : longue durée : 20 sec*			
Apprentissage	Scoring	det 6 (tél-tél tous) EER%	det 7 (tél-tél anglais) EER%
long	2-cov.	7.09	3.86
(20s et long)	2-cov.	6.85	3.96
long	PLDA	7.21	3.99
(20s et long)	PLDA	7.06	4.14
(20s et long)	4-cov. reg. lin. $\mathbf{Y}_{\text{long}}$~$\mathbf{Y}_{\text{court}}$	6.85	3.83
(20s et long)	4-cov. transf. gauss. $\mathbf{Y}_{\text{long}} \leftrightarrow \mathbf{Y}_{\text{court}}$	**6.75**	**3.72**

TABLE 9.10 – *Pour le cas de tests 20 secondes, comparaison de performances, en terme d'EER, de différents scorings basés sur les modèles LDA-two covariance, PLDA gaussienne et LDA-4covariance, ce dernier avec pont bayesien par régression linéaire ou par fonction de transfert gaussienne.*

la condition 7 (tél-tél anglais natifs) sur les segments courts "5sec" et "20sec", mais pas sur "10sec".

Dans cinq cas sur six, ce modèle est le plus performant. Ce résultat est obtenu à chaque fois par le pont bayésien basé sur la fonction de transfert de l'équation 9.14. Son caractère probabiliste (la régression linéaire n'étant qu'une optimisation géométrique) contribue clairement à son efficacité.

9.3.4 Conclusion

Sans être exceptionnel, le gain de performance du modèle 4-covariance pour les applications mixtes par rapport au modèle PLDA s'avère significatif dans le cas le plus délicat de courtes durées "5sec" de 500 trames seulement. Le pont bayesien basé sur la fonction de transfert gaussien est le plus pertinent. Cette étude montre que l'efficience de la modélisation probabiliste PLDA sur des données mixtes est remise en question : une technique PLDA adaptée à ce type d'applications reste à définir. Des pistes de recherche (*tied-PLDA*) peuvent être trouvées dans (Prince et al., 2008) (Prince and Elder, 2007).

L'estimation par régression linéaire du pont probabiliste entre les facteurs locuteurs y_1 et y_2 nécessite un fichier de locuteurs multi-types (c.à.d. contenant à la fois des énoncés de classe 1 et de classe 2 de ce locuteur) pour l'estimation de la matrice $\widehat{\mathbf{A}}$ de transfert entre les classes. Celle-ci n'est calculable que si la matrice $\left(\overline{Y_1}^t \overline{Y_1}\right)$ est inversible, donc si l'on dispose d'un effectif de données appariées (Y_1, Y_2) supérieur à la dimension de l'espace des i-vectors.

Par la fonction gaussienne de transfert de l'équation 9.14, l'estimation s'appuie sur les matrices $\mathbf{B}_1$ et $\mathbf{B}_2$ de covariance inter-locuteur, qui sont calculées sur un jeu de données de grande taille non nécessairement appariées (contenant notamment des énoncés

de type 1 seulement pour certains locuteurs , de type 2 seulement pour d'autres). Mais le calcul de la covariance du résidu (équations 9.11 et 9.15) pour l'estimation de $y_2|y_1$ nécessite, lui, un fichier apparié de plein rang.

De nombreux cas se présentent dans le domaine de la reconnaissance vocale où de telles données appariées manquent pour constituer un tel fichier. Actuellement, les dimensions efficaces des espaces de i-vectors sont supérieures ou égales à 400 et l'insuffisance de données d'apprentissage à disposition (multi-langages, multi-canal téléphone / microphone) empêche la plupart des organismes d'appliquer un modèle tel que le 4-covariance. Nous avons mis en place des adaptations mathématiques au cas de matrices de variabilité inter-type singulières mais les résultats s'avèrent peu probants. Seule la collecte de très grandes quantités de données d'apprentissage, des deux types et appariées par locuteur, permettra dans l'avenir de lever cet obstacle à la qualité des applications de reconnaissance basées sur des énoncés mixtes.

Quatrième partie

Annexes

Annexe A

Description des corpus d'apprentissage et d'évaluation

Les jeux de données d'apprentissage et d'évaluation ainsi que les conditions d'évaluation utilisés pour nos analyses et expérimentations sont détaillés dans cette section. Chaque fois, le titre du paragraphe est le code utilisé pour les nommer dans le document.

A.1 GMM-UBM-LIA

Ce GMM-UBM a été utilisé pour la première fois pour la soumission LIA de la campagne NIST-SRE 2006, hommes seulement. Son apprentissage est réalisé à partir de la base Fisher (Fisher English Training Speech Part 1, LDC :LDC2004S13) et consiste en quelques 10 millions de trames de parole. Ces trames sont composées de 19 paramètres LFCC, leurs dérivées premières et 11 dérivées secondes (la fenêtre de fréquence est restreinte à 300-3400 Hz). Une normalisation est appliquée pour standardiser chaque coefficient cepstral (moyenne 0, variance 1) d'un segment donné. Le GMM a 512 composantes dont les paramètres de variance sont seuillés à un plancher de 50% de la variance globale (0.5).

A.2 T15660-LIA

Cette matrice de variabilité totale T du système i-vector LIA est apprise à partir de 15660 segments de 1147 locuteurs (NIST 2004, 2005, 2006 et Switchboard II part 1, 2 & 3, Switchboard cellular part 1 & 2, à peu près 14 sessions par locuteur). Les modèles locuteurs sont dérivées d'une adaptation bayesienne des moyennes des composantes gaussiennes par MAP avec un facteur de confiance (*relevant factor* τ) universel de 14. La dimension de l'espace des i-vectors résultant est de 400. La base d'apprentissage de T

		Apprentissage	Evaluation (NIST-SRE 2010)	dimension i-vectors
BUT	hommes	21475	2294 (téléphone seulement)	600
BUT	femmes	27155	2740 (téléphone seulement)	600

TABLE *A.1* – *Détail des diverses configurations i-vectors du laboratoire BUT utilisées : tailles des fichiers d'apprentissage (extraction et modélisation), de test et dimension des espaces.*

		Apprentissage	Evaluation (NIST-SRE 2010)	dimension i-vectors
LIA	hommes	19349	5200 (tous)	400

TABLE *A.2* – *Détail des diverses configurations i-vectors du laboratoire LIA utilisées : tailles des fichiers d'apprentissage (extraction et modélisation), de test et dimension des espaces.*

est réutilisée pour estimer les matrices de covariances LDA et WCCN dans l'espace des i-vectors.

A.3 BUT-hommes, BUT-femmes

Fournis par le Brno University of Technology BUT [1], ces jeux de i-vectors ont été produits suivant une configuration d'extraction décrite dans (Bousquet et al., 2012a) et (Matejka et al., 2011). La table A.1 indique les tailles des fichiers utilisés, ainsi que les dimensions des i-vectors. La première ligne indique, par exemple, que 21475 i-vectors sont utilisés comme apprentissage pour la normalisation et la modélisation et 2294 durant les tests de l'évaluation.

A.4 LIA-hommes

Ces i-vectors sont extraits par le LIA suivant les configurations GMM-UBM-LIA et T15660-LIA (dont les détails se trouvent dans cette section). La table A.2 indique les tailles des fichiers utilisés, ainsi que la dimensions des i-vectors.

A.5 NIST-SRE 2008

Il s'agit de l'évaluation NIST-SRE 2008 codée short2-short3, hommes seulement. Les données consistent en fichiers de conversation de 2 à 3 minutes. Ce protocole utilise 3798 locuteurs. Les énoncés des locuteurs-cibles et tests contiennent en moyenne 2.5 minutes de parole, dont à peu près 30% des trames ont été retenues. Les 6615 tests de la condition "téléphone- téléphone anglais natifs det 7" considérée se séparent en 439 tests-cible et 6176 tests-imposteur. Les 12511 tests de la condition "téléphone- téléphone tous det 6" considérée se séparent en 874 tests-cible et 11637 tests-imposteur.

1. Merci à Pavel Matejka.

A.6 NIST-SRE-2010-det5Extended

Ce jeu d'essais de NIST-SRE 2010 est constitué de segments téléphone seulement (enrôlement et test), effort vocal normal et utilisé dans sa version étendue (condition *det 5 Extended*). Le fichier de test hommes compte 179338 tests, dont 3465 tests-cible, celui des femmes compte 236781 tests, dont 3704 tests-cible.

Annexe B

Formulation du score two-covariance

B.1 Sous l'hypothèse θ_{tar}

On note $\mathcal{M}$ le modèle two-covariance des équations 2.61 et 2.62.

$$P\left(\{w_1, w_2\}\,|\theta_{\text{tar}}\right) = \int P\left(w_1|s,\mathcal{M}\right) P\left(w_2|s,\mathcal{M}\right) P\left(s\right) ds \tag{B.1}$$

$$= \frac{1}{\left(2\pi\right)^{-\frac{3p}{2}}} \left|\mathbf{W}\right|^{-1} \left|\mathbf{B}\right|^{-\frac{1}{2}} \times \int \exp\left(-\frac{1}{2}\mathcal{A}\right) ds \tag{B.2}$$

où $\mathcal{A} = \left(w_1 - s\right)^t \mathbf{W}^{-1} \left(w_1 - s\right) + \left(w_2 - s\right)^t \mathbf{W}^{-1} \left(w_2 - s\right) + \left(s - \mu\right)^t B^{-1} \left(s - \mu\right)$.

On a

$$\mathcal{A} = w_1\mathbf{W}^{-1}w_1 + w_2\mathbf{W}^{-1}w_2 + \left(s - a_0\right)^t \left(\mathbf{B}^{-1} + 2\mathbf{W}^{-1}\right) \left(s - a_0\right) - a_0^t \left(\mathbf{B}^{-1} + 2\mathbf{W}^{-1}\right) a_0 \tag{B.3}$$

avec $a_0 = \left(\mathbf{B}^{-1} + 2\mathbf{W}^{-1}\right)^{-1} \left(\mathbf{B}^{-1}\mu + \mathbf{W}^{-1}\left(w_1 + w_2\right)\right)$

d'où

$$\log P\left(\{w_1, w_2\}\,|\theta_{\text{tar}}\right) = C - \frac{1}{2}\left\{w_1\mathbf{W}^{-1}w_1 + w_2\mathbf{W}^{-1}w_2 - a_0^t \left(\mathbf{B}^{-1} + 2\mathbf{W}^{-1}\right) a_0\right\} \tag{B.4}$$

où C constante.

B.2 Sous l'hypothèse $\theta_{\mathbf{non}}$

$$P\left(\{w_1, w_2\}|\theta_{\text{non}}\right) = P\left(\{w_1\}|\mathcal{M}\right) P\left(\{w_2\}|\mathcal{M}\right) \tag{B.5}$$

$$= \int P\left(w_1|w_2\right) P\left(w_2\right) dw_2 \int P\left(w_2|w_2\right) P\left(w_2\right) dw_2 \tag{B.6}$$

$$= \frac{1}{(2\pi)^{-\frac{3p}{2}}} \left|\mathbf{W}\right|^{-1} \left|\mathbf{B}\right|^{-1} \tag{B.7}$$

$$\times \int \exp\left(-\frac{1}{2}\left\{\left(w_1 - s\right)^t \mathbf{W}^{-1}\left(w_1 - s\right) + \left(s - \mu\right)^t \mathbf{B}^{-1}\left(s - \mu\right)\right\}\right) ds \tag{B.8}$$

$$\times \int \exp\left(-\frac{1}{2}\left\{\left(w_2 - s\right)^t \mathbf{W}^{-1}\left(w_2 - s\right) + \left(s - \mu\right)^t \mathbf{B}^{-1}\left(s - \mu\right)\right\}\right) ds \tag{B.9}$$

et, pour $i = 1$ ou 2

$$\left(w_i - s\right)^t \mathbf{W}^{-1}\left(w_i - s\right) + \left(s - \mu\right)^t \mathbf{B}^{-1}\left(s - \mu\right) = w_i \mathbf{W}^{-1} w_i + \left(s - a_i\right)^t \left(\mathbf{B}^{-1} + \mathbf{W}^{-1}\right)\left(s - a_i\right) \tag{B.10}$$

$$- a_i^t \left(\mathbf{B}^{-1} + \mathbf{W}^{-1}\right) a_i \tag{B.11}$$

avec $a_i = \left(\mathbf{B}^{-1} + \mathbf{W}^{-1}\right)^{-1} \left(\mathbf{B}^{-1}\mu + \mathbf{W}^{-1} w_i\right)$ d'où

$$\log P\left(\{w_1, w_2\}|\theta_{\text{non}}\right) = C' - \tag{B.12}$$

$$\frac{1}{2}\left\{w_1 \mathbf{W}^{-1} w_1 + w_2 \mathbf{W}^{-1} w_2 - a_1^t \left(\mathbf{B}^{-1} + \mathbf{W}^{-1}\right) a_1 - a_2^t \left(\mathbf{B}^{-1} + \mathbf{W}^{-1}\right) a_2\right\} \tag{B.13}$$

où C' constante.

On déduit des deux formules précédentes le score :

$$score\left(w_1, w_2\right) = K - \frac{1}{2}\left\{-a_0^t \left(\mathbf{B}^{-1} + 2\mathbf{W}^{-1}\right) a_0 + a_1^t \left(\mathbf{B}^{-1} + \mathbf{W}^{-1}\right) a_1 + a_2^t \left(\mathbf{B}^{-1} + \mathbf{W}^{-1}\right) a_2\right\} \tag{B.14}$$

B.3 Score log-ratio de vraisemblance

Après normalisation, la moyenne globaleμ est quasiment égale à 0, d'où :

$$a_0 = \left(\mathbf{B}^{-1} + 2\mathbf{W}^{-1}\right)^{-1} \mathbf{W}^{-1} (w_1 + w_2)$$

et

$$a_i = \left(\mathbf{B}^{-1} + \mathbf{W}^{-1}\right)^{-1} \mathbf{W}^{-1} w_i$$

d'où

$$\log P(\{w_1, w_2\} | \theta_{\text{tar}}) = C - \frac{1}{2} w_1 \mathbf{W}^{-1} w_1 - \frac{1}{2} w_2 \mathbf{W}^{-1} w_2 \tag{B.15}$$

$$+ \frac{1}{2} \left(\left(\mathbf{B}^{-1} + 2\mathbf{W}^{-1}\right)^{-1} \mathbf{W}^{-1} (w_1 + w_2) \right)^t \tag{B.16}$$

$$\left(\mathbf{B}^{-1} + 2\mathbf{W}^{-1}\right) \left(\mathbf{B}^{-1} + 2\mathbf{W}^{-1}\right)^{-1} \mathbf{W}^{-1} (w_1 + w_2) \tag{B.17}$$

$$\log P(\{w_1, w_2\} | \theta_{\text{tar}}) = C - \frac{1}{2} w_1 \mathbf{W}^{-1} w_1 - \frac{1}{2} w_2 \mathbf{W}^{-1} w_2 \tag{B.18}$$

$$+ \frac{1}{2} (w_1 + w_2)^t \mathbf{W}^{-1} \left(\mathbf{B}^{-1} + 2\mathbf{W}^{-1}\right)^{-1} \mathbf{W}^{-1} (w_1 + w_2) \tag{B.19}$$

et

$$\log P(\{w_1, w_2\} | \theta_{\text{non}}) = C' - \frac{1}{2} w_1 \mathbf{W}^{-1} w_1 - \frac{1}{2} w_2 \mathbf{W}^{-1} w_2 \tag{B.20}$$

$$+ \frac{1}{2} \left(\left(\mathbf{B}^{-1} + \mathbf{W}^{-1}\right)^{-1} \mathbf{W}^{-1} w_1 \right)^t \left(\mathbf{B}^{-1} + \mathbf{W}^{-1}\right) \left(\mathbf{B}^{-1} + \mathbf{W}^{-1}\right)^{-1} \mathbf{W}^{-1} w_1 \tag{B.21}$$

$$+ \frac{1}{2} \left(\left(\mathbf{B}^{-1} + \mathbf{W}^{-1}\right)^{-1} \mathbf{W}^{-1} w_i \right)^t \left(\mathbf{B}^{-1} + \mathbf{W}^{-1}\right) \left(\mathbf{B}^{-1} + \mathbf{W}^{-1}\right)^{-1} \mathbf{W}^{-1} w_i \tag{B.22}$$

$$\log P(\{w_1, w_2\} | \theta_{\text{non}}) = C' - \frac{1}{2} w_1 \mathbf{W}^{-1} w_1 - \frac{1}{2} w_2 \mathbf{W}^{-1} w_2 \tag{B.23}$$

$$+ \frac{1}{2} w_1 \mathbf{W}^{-1} \left(\mathbf{B}^{-1} + \mathbf{W}^{-1}\right)^{-1} \mathbf{W}^{-1} w_1 \tag{B.24}$$

$$+ \frac{1}{2} w_2 \mathbf{W}^{-1} \left(\mathbf{B}^{-1} + \mathbf{W}^{-1}\right)^{-1} \mathbf{W}^{-1} w_2 \tag{B.25}$$

On déduit des deux formules précédentes le score final (la constante étant retirée) :

$$score\,(w_1, w_2) = \frac{1}{2}\,(w_1 + w_2)^t\,\mathbf{W}^{-1}\left(\mathbf{B}^{-1} + 2\mathbf{W}^{-1}\right)^{-1}\mathbf{W}^{-1}\,(w_1 + w_2) \tag{B.26}$$

$$- \frac{1}{2} w_1^t \mathbf{W}^{-1}\left(\mathbf{B}^{-1} + \mathbf{W}^{-1}\right)^{-1}\mathbf{W}^{-1} w_1 - \frac{1}{2} w_2^t \mathbf{W}^{-1}\left(\mathbf{B}^{-1} + \mathbf{W}^{-1}\right)^{-1}\mathbf{W}^{-1} w_2 \tag{B.27}$$

Ce score est aussi décomposable en fonction de w_1 et w_2 de la manière suivante :

$$score\,(w_1, w_2) = \frac{1}{2}\left\{w_1^t \mathcal{P} w_1 + w_2^t \mathcal{P} w_2 + 2 w_1^t \mathcal{Q} w_2\right\} \tag{B.28}$$

où

$$\mathcal{P} = \mathbf{W}^{-1}\left\{\left(\mathbf{B}^{-1} + 2\mathbf{W}^{-1}\right)^{-1} - \left(\mathbf{B}^{-1} + \mathbf{W}^{-1}\right)^{-1}\right\}\mathbf{W}^{-1} \tag{B.29}$$

$$\mathcal{Q} = \mathbf{W}^{-1}\left(\mathbf{B}^{-1} + 2\mathbf{W}^{-1}\right)^{-1}\mathbf{W}^{-1} \tag{B.30}$$

Annexe C

Gaussianité de divers jeux d'apprentissage et de test

En complément de la figure 4.2 du paragraphe 4.2.2 mesurant la gaussianité des i-vectors , nous présentons les évolutions des normes carrées de différents jeux de i-vectors. Ces valeurs indiquent le degré de gaussianité (ou plutôt de non-gaussianité) des données, au fur et à mesure des itérations de l'algorithme de normalisation EFR.

La figure C.1 présente les histogrammes successifs des normes carrées des jeux de données standardisés d'apprentissage et d'évaluation BUT-femmes, ainsi que la densité du χ^2 à p degrés de liberté (ici $p = 600$). Ils permettent notamment de constater que les conclusions du paragraphe 4.2.2 sont indépendants du genre.

La figure C.2 présente les histogrammes successifs des normes carrées des jeux de données standardisés d'apprentissage et d'évaluation LIA-hommes, ainsi que la densité du χ^2 à p degrés de liberté (ici $p = 400$). Les conclusions tirées de l'étude effectuée au paragraphe 4.2.2 restent valables pour les données de notre laboratoire (les différences apparentes ne tiennent qu'aux échelles distinctes de l'axe horizontal).

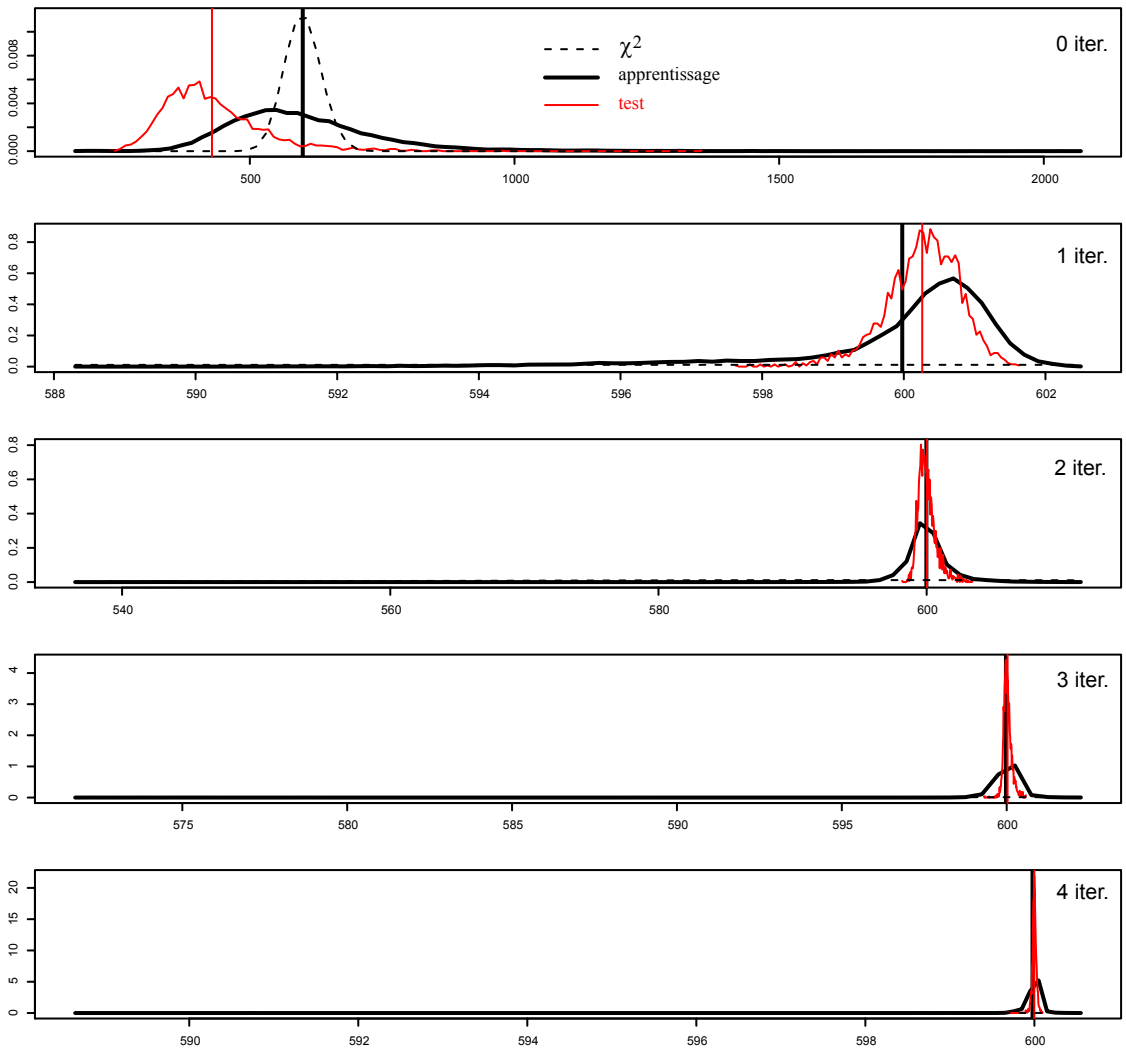

FIGURE C.1 – *Histogrammes successifs des normes carrées des jeux de données standardisés d'apprentissage et d'évaluation BUT-femmes, et densité du χ^2 à p degrés de liberté ($p = 600$).*

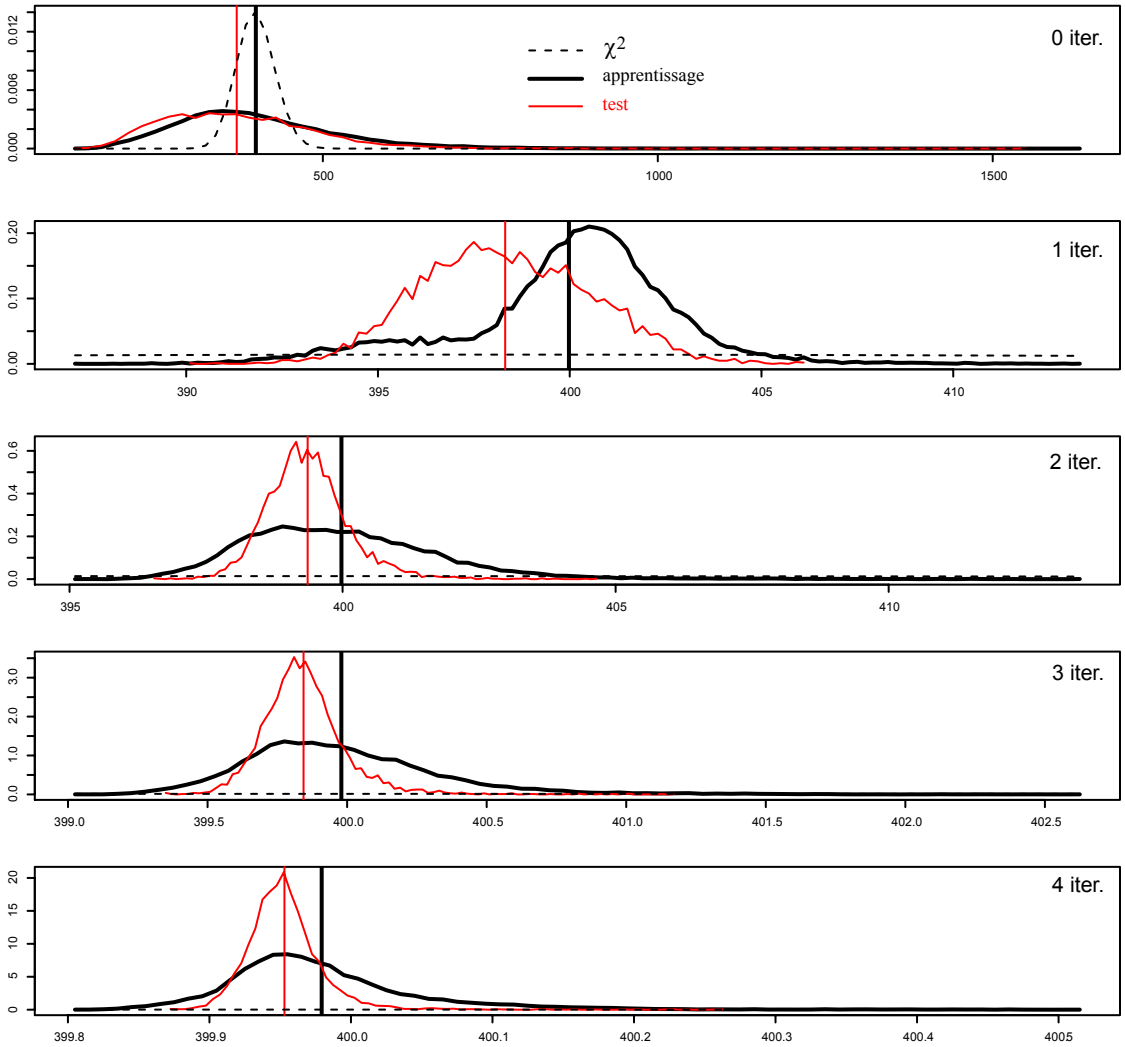

FIGURE C.2 – *Histogrammes successifs des normes carrées des jeux de données standardisés d'apprentissage et d'évaluation LIA-hommes, et densité du χ^2 à p degrés de liberté ($p = 400$).*

Annexe D

Résultats de la comparaison de systèmes du chapitre 5

Les tables D.1, D.2 et D.3 affichent les résultats de la comparaison de systèmes du chapitre 5, respectivement pour :

- les évaluations hommes, en terme de DCF minimale,
- les évaluations femmes, en terme d'EER,
- les évaluations femmes, en terme de DCF minimale.

DCF min					2008		2010	
	repr	reduc dim	norm	scoring	det 7	det 6	det 5 Ext	**moy.**
1	BKbinary	PCA	non	PLDA	0.16	0.33	0.65	**0.38**
2	BKcounts	PCA	non	PLDA	0.16	0.33	0.63	**0.37**
3	sv	PCA	no	PLDA	0.17	0.34	0.62	**0.38**
4	sv	FA	no	LDAMaha	0.33	0.42	0.86	**0.54**
5	sv	FA	no	LDATwoCov	0.19	0.33	0.60	**0.37**
6	sv	FA	no	PLDA	0.17	0.32	0.59	**0.36**
7	BKbinary	PCA	SphN	PLDA	0.13	0.29	0.54	**0.32**
8	BKcounts	PCA	SphN	PLDA	0.12	0.29	0.57	**0.33**
9	sv	PCA	SphN	PLDA	0.12	0.28	0.52	**0.31**
10	sv	FA	cosine	LDA-WCCN-Cosine	0.17	0.33	0.55	**0.35**
11	sv	FA	EFR	LDAMaha	0.12	0.29	0.52	**0.31**
12	sv	FA	EFR	LDATwoCov	0.12	0.28	0.49	**0.30**
13	sv	FA	LNorm	PLDA	0.12	0.28	0.49	**0.30**
14	sv	FA	EFR	PLDA	0.12	0.28	0.48	**0.29**
15	sv	FA	SphN	PLDA	0.12	0.28	0.47	**0.29**

TABLE *D.1* – *Synthèse des performances obtenues par les différents systèmes envisagés pour l'évaluation **hommes** et en terme de **DCF minimale**.*

EER (%)					2008		2010	
	repr	reduc dim	norm	scoring	det 7	det 6	det 5 Ext	**moy**
1	BKbinary	PCA	non	PLDA	3.04	6.75	4.75	4.85
2	BKcounts	PCA	non	PLDA	3.70	7.34	6.06	5.70
3	sv	PCA	non	PLDA	3.68	7.32	5.27	5.42
4	sv	FA	non	LDAMaha	5.46	9.37	8.42	7.75
5	sv	FA	non	LDATwoCov	3.30	7.21	5.48	5.33
6	sv	FA	non	PLDA	3.15	6.86	5.35	5.12
7	BKbinary	PCA	SphN	PLDA	2.82	6.76	3.86	4.48
8	BKcounts	PCA	SphN	PLDA	2.92	6.65	4.10	4.56
9	sv	PCA	SphN	PLDA	3.04	6.76	3.85	4.55
10	sv	FA	cosine	LDA-WCCN-Cosine	3.55	8.09	4.18	5.27
11	sv	FA	EFR	LDAMaha	2.72	6.79	3.32	4.28
12	sv	FA	EFR	LDATwoCov	2.66	6.63	3.21	4.17
13	sv	FA	LNorm	PLDA	2.66	6.27	3.34	4.09
14	sv	FA	EFR	PLDA	2.66	6.43	3.29	4.13
15	sv	FA	SphN	PLDA	2.53	6.31	3.02	3.95

TABLE *D.2* – *Synthèse des performances obtenues par les différents systèmes envisagés pour l'évaluation **femmes**, en terme d'**EER**.*

DCF min					2008		2010	
	repr	reduc dim	norm	scoring	det 7	det 6	det 5 Ext	**moy**
1	BKbinary	PCA	non	PLDA	0.15	0.35	0.67	0.39
2	BKcounts	PCA	non	PLDA	0.16	0.37	0.69	0.41
3	sv	PCA	non	PLDA	0.16	0.36	0.66	0.39
4	sv	FA	non	LDAMaha	0.26	0.48	0.94	0.56
5	sv	FA	non	LDATwoCov	0.15	0.35	0.69	0.40
6	sv	FA	non	PLDA	0.13	0.36	0.62	0.37
7	BKbinary	PCA	SphN	PLDA	0.13	0.35	0.65	0.38
8	BKcounts	PCA	SphN	PLDA	0.13	0.35	0.67	0.38
9	sv	PCA	SphN	PLDA	0.12	0.34	0.61	0.36
10	sv	FA	cosine	LDA-WCCN-Cosine	0.17	0.39	0.63	0.40
11	sv	FA	EFR	LDAMaha	0.13	0.35	0.61	0.36
12	sv	FA	EFR	LDATwoCov	0.11	0.34	0.59	0.35
13	sv	FA	LNorm	PLDA	0.11	0.33	0.57	0.34
14	sv	FA	EFR	PLDA	0.11	0.33	0.58	0.34
15	sv	FA	SphN	PLDA	0.11	0.33	0.58	0.34

TABLE *D.3* – *Synthèse des performances obtenues par les différents systèmes envisagés pour l'évaluation **femmes** , en terme de **DCF minimale**.*

Annexe E

Extraction de typicalités pour le modèle de clés binaires du locuteur

Mathématiquement, nous définissons une typicalité comme un sous-ensemble de spécificités liées par une relation autorisant à calculer entre elles un *produit total* et pas seulement scalaire. Le produit total est calculé de la manière suivante : pour les deux vecteurs, les valeurs 0 ou 1 pour le seul sous-ensemble des spécificités correspondant à une typicalité $\mathcal{L}$ sont sélectionnées. Puis une valeur de similarité entre les deux sous-vecteurs suivant $\mathcal{L}$ est calculée par multiplications croisées entre toutes les valeurs binaires retenues (somme des produits entre toutes les paires de valeurs). La similarité entre deux vecteurs v_1 et v_2 suivant la typicalité $\mathcal{L}$ est donc le produit total suivant $\mathcal{L}$, défini par :

$$S_{\mathcal{L}}(v_1, v_2) = \sum_{i \in \mathcal{L}} \sum_{j \in \mathcal{L}} (v_1)_i (v_2)_j \tag{E.1}$$

La figure E.1 montre une illustration du produit "total". Pour deux vecteurs binaires, leurs seules valeurs binaires (0 ou 1) pour la typicalité $\mathcal{L}$ considérée sont sélectionnées : $(0,1,0,0,1,1)$ et $(0,1,1,0,0,0)$. Une valeur de similarité entre les deux sous-vecteurs est alors calculée par multiplication croisée entre toutes leurs valeurs. Le résultat, égal à 6, est le produit total entre les deux vecteurs. On considère donc le sous-vecteur des seules spécificités de $\mathcal{L}$ comme la trace d'un lien, hérité d'une origine commune dans une classification.

E.1 Calcul rapide du produit total

La similarité précédente peut être calculée plus simplement : si les sous-vecteurs des indices de $\mathcal{L}$ pour v_1 (resp. v_2) contiennent n_1 (resp. n_2) valeurs à 1, le produit total entre v_1 et v_2 est égal à $n_1 \times n_2$.

Soit $\delta_{\mathcal{L}}$ le vecteur de Kronecker de IR^n défini par :

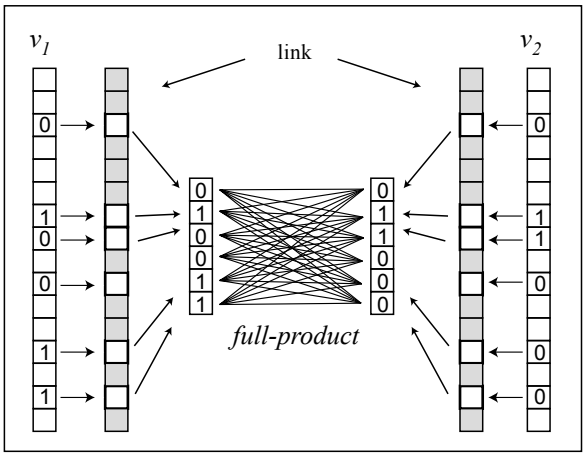

FIGURE *E.1* – *Illustration du produit total entre deux clés binaires.*

$(\delta_{\mathcal{L}})_k = 1$ si la $k^{\text{ème}}$ spécificité appartient à $\mathcal{L}$
$(\delta_{\mathcal{L}})_k = 0$ sinon

Le produit total peut être réécrit :

$$S_{\mathcal{L}}(v_1, v_2) = (\delta_{\mathcal{L}}.v_1) \times (\delta_{\mathcal{L}}.v_2) \tag{E.2}$$

où "." est le produit scalaire.

Pour calculer une similarité entre deux vecteurs suivant un jeu $\{\mathcal{L}_1, ..., \mathcal{L}_l\}$ de l typicalités, nous introduisons la matrice L de dimensions $l \times N$ définie par : $k^{\text{ème}}$ ligne de $L = \delta_{\mathcal{L}_k}$, vecteur de Kronecker du $k^{\text{ème}}$ lien. La similarité totale suivant le jeu de typicalités $\{\mathcal{L}_1, ..., \mathcal{L}_l\}$ s'écrit :

$$S_{\{\mathcal{L}\}}(v_1, v_2) = Lv_1.Lv_2 \tag{E.3}$$

Cette similarité peut être pondérée. Par exemple, les cardinaux des typicalités ne sont pas identiques. Pour standardiser les effets des diverses typicalités, on peut introduire la matrice diagonale $\mathbf{D}$ remplie avec les cardinaux des $\mathcal{L}_k$: pour tout k de $\{1, .., l\}$:

$$\mathbf{D}_{k,k} = card(\mathcal{L}_k) \tag{E.4}$$

c'est à dire le nombre de 1 dans la $k^{\text{ème}}$ typicalité.

Un vecteur de lien devient :

$$v \to \mathbf{D}^{-\frac{1}{2}} Lv \tag{E.5}$$

et la similarité pondérée :

$$S_{\{\mathcal{L}\}}(v_1, v_2) = v_1.\mathbf{L}^t\mathbf{D}\mathbf{L}v_2 \tag{E.6}$$

E.2 Décomposition en valeurs singulières et variables catégorielles

Nous rappelons ici le principe de la procédure NAP (*Nuisance Attribute projection*) de compensation de l'effet session utilisée notamment en reconnaissance du locuteur. Puis l'utilisation d'une décomposition en valeurs singulières, telle que nécessitée par NAP, sera justifiée lorsqu'elle s'applique aux clés binaires, qui sont pourtant des vecteurs de variables aléatoires catégorielles et non numériques.

E.2.1 Technique NAP

La procédure NAP (Campbell et al., 2006) estime la variabilité session comme un sous-espace linéaire de rang $r < p$. Il est obtenu par décomposition en valeurs singulières, la diagonalisation fournissant ses axes principaux (il s'agit des vecteurs propres de la matrice de covariance intra-classes $\mathbf{W}$ associés aux plus grandes valeurs propres de son spectre). Les vecteurs initiaux sont alors projetés sur le sous-espace complémentaire orthogonal de ce sous-espace principal, supposé être une estimation du sous-espace locuteur principal.

Soit $\mathbf{U}$ la matrice $r \times N$ des r vecteurs propres principaux de $\mathbf{W}$ en colonnes. La projection vectorielle sur l'orthogonal utilise la matrice $\mathbf{I} - \mathbf{U}\mathbf{U}^t$ et la similarité entre deux vecteurs dont l'effet-session a été compensé par NAP s'écrit :

$$S\left(v_1, v_2\right) = \left(\mathbf{I} - \mathbf{U}^t\mathbf{U}\right) v_1 . \left(\mathbf{I} - \mathbf{U}^t\mathbf{U}\right) v_2 \tag{E.7}$$

Puisque une matrice de projection est symétrique et idempotente (égale à son carré), la similarité précédente s'exprime plus simplement - et se calcule plus rapidement- par :

$$S\left(v_1, v_2\right) = v_1 . v_2 - \mathbf{U}v_1 . \mathbf{U}v_2 \tag{E.8}$$

La technique de compensation NAP apparaît comme un redressement de la similarité initiale par le produit scalaire entre leurs projetés de dimension réduite.

E.2.2 Application aux variables catégorielles

Une technique comme NAP emploie une décomposition en valeurs singulières. La diagonalisation peut sembler inadéquate étant donnée la nature catégorielle (qualitative) des variables aléatoires spécificités. Il existe une technique équivalente à la PCA adaptée à de telles variables : l'ACM, analyse des correspondances multiples (*Multiple Correspondance analysis MCA*). Celle-ci détermine des axes principaux de variabilité et un spectre de valeurs propres associé à partir de données non-numériques, à valeurs dans un ensemble fini de modalités. L'ACM bâtit une hypertable de contingence et un tableau de Burt pour charger les fréquences de chaque paires de modalités (Lebart et al., 2000). Puis une norme induite par la métrique du χ^2 permet de comparer les modalités

des variables et de calculer des "axes factoriels" de variabilité principale, sur lesquels sont projetés les données.

Mais dans le cas spécial de variables binaires (i.e. dont toutes les variables aléatoires, c'est à dire les dimension des vecteurs, n'ont que deux niveaux/modalités), il est démontré que l'ACM est équivalent à une PCA effectuée sur des vecteurs codant en binaire (0 ou 1) l'occurrence des facteurs. Il est ainsi possible d'employer les techniques usuelles de réduction de dimension par projection sur des axes principaux dans le cas de notre espace de spécificités.

E.2.3 Variabilité totale

Cette remarque étant faite, il est aussi envisageable de diagonaliser la matrice de covariance totale (et non intra-classe) pour obtenir une matrice rectangulaire **T** de rang réduit r'. En projetant les vecteurs initiaux sur le sous-espace engendré par **T**, leurs résidus en haute dimension initiale sont éliminés. Nous appellerons *T-similarité* la similarité obtenue après cette projection. Elle est définie par :

$$\mathbf{T}v_1.\mathbf{T}v_2 \tag{E.9}$$

Mais si, comme l'a remarqué (Matrouf and Bonastre, 2009) une majorité de la variabilité entre les vecteurs est due à la variabilité session, il est intéressant de combiner similarité initiale et *T*-similarité dans une sorte de technique de type NAP de variabilité totale. Nous présentons plus loin les résultats d'une telle étude.

E.3 Un exemple d'extracteur de typicalités par binarisation de vecteurs propres

Nous présentons ci-dessous un exemple d'algorithme d'extraction de typicalités : des familles de spécificités sont extraites par cet algorithme, dotées de caractéristiques discriminantes.

E.3.1 Algorithme de binarisation

Relation entre spécificités et axes principaux Il est possible d'extraire des familles de liens inter-spécificités à partir des vecteurs propres des matrices de variabilité **U** et **T** précédentes. Souvent, dans le domaine de l'analyse de données, il ne s'agit pas seulement de produire des axes principaux résumant la variabilité : le lien entre ces nouveaux axes et les axes initiaux, en l'occurrence des dimensions associées à des variables portant un sens (indicateurs sociologiques, économiques, scientifiques, ...), doit être déterminé le plus clairement possible. Ainsi, chaque axe principal, c'est à dire chaque vecteur propre d'une matrice de covariances dans le cas d'une décomposition en valeurs

singulières, est "expliqué" par les variables initiales : on dira que ce vecteur propre est principalement associé à certaines variables, plus associé à l'une qu'à l'autre, en mesurant un indice de corrélation entre ce vecteur propre et les variables initiales. La conclusion dans de telles analyses de données est d'expliquer chaque axe principal par la famille de variables qui lui sont le plus corrélées (celles-ci ont souvent un sens, commun, qui permet de résumer l'axe par un titre : l'axe des "revenus", du "niveau de développement", de "l'éducation", ...) Pour procéder et déterminer ainsi la proportion de chaque variable, pour nous les spécificités, dans la génération d'un axe principal, le jeu de données d'apprentissage d'effectif n est d'abord projeté sur cet axe. Le vecteur aléatoire de $\mathbb{R}^n$ obtenu est la *composante principale* de cet axe, bien connue en analyse de données. Puis sont calculés les N coefficients de covariance (ou de corrélation si l'on souhaite annuler les différences d'échelle) entre la composante principale et les vecteurs aléatoires de $\mathbb{R}^n$ de chaque variable dimensionnelle.

Ceci s'écrit mathématiquement ainsi :

Soit X la matrice du jeu de données d'apprentissage, à n lignes (taille de l'échantillon d'apprentissage) et N colonnes (dimension de l'espace de représentation = nombre de spécificités). Chaque ligne de X est un vecteur d'apprentissage. Cette matrice est centrée, suivant la moyenne globale, pour la variabilité totale, ou bien du locuteur de chaque session, pour la variabilité session. Soit $\overline{X}$ la version centrée de X. La matrice de covariance V à diagonaliser est $\mathbf{\Sigma} = \frac{1}{n}\overline{X}^t\overline{X}$.

Soit v_j le $j^{\text{ème}}$ vecteur propre, unitaire, de la matrice de diagonalisation utilisée et soit λ_j sa valeur propre associée. La $j^{\text{ème}}$ composante principale, noté p_j est égale à :

$$p_j = \overline{X}v_j \tag{E.10}$$

Notons $cov\left(s_k, p_j\right)$ la covariance entre la $k^{\text{ème}}$ spécificité et la $j^{\text{ème}}$ composante principale p_j. Nous entendons ici par $k^{\text{ème}}$ spécificité les valeurs de la $k^{\text{ème}}$ dimension des vecteurs binaires d'apprentissage. Les composantes principales étant centrées, on a :

$$cov\left(s_k, p_j\right) = \frac{1}{n}\left(\overline{X}_{*,k}\right).\left(\overline{X}v_j\right) \tag{E.11}$$

où $\overline{X}_{*,k}$ est la $k^{\text{ème}}$ colonne de $\overline{X}$.

Une valeur forte (positivement ou négativement) de cette covariance indique une forte relation entre l'axe principal j et la spécificité k. D'autre part, pour j fixé, les plus négatives de ces covariances et les plus positives forment deux familles antagonistes : les éléments de la première sont de même comportement vis à vis de l'axe principal j, donc identifiables comme corrélées. Les élément de la seconde famille le sont de même. Mais les deux familles entre elles sont corrélées négativement : de fortes valeurs dans l'une induisent statistiquement de faibles valeurs dans l'autre.

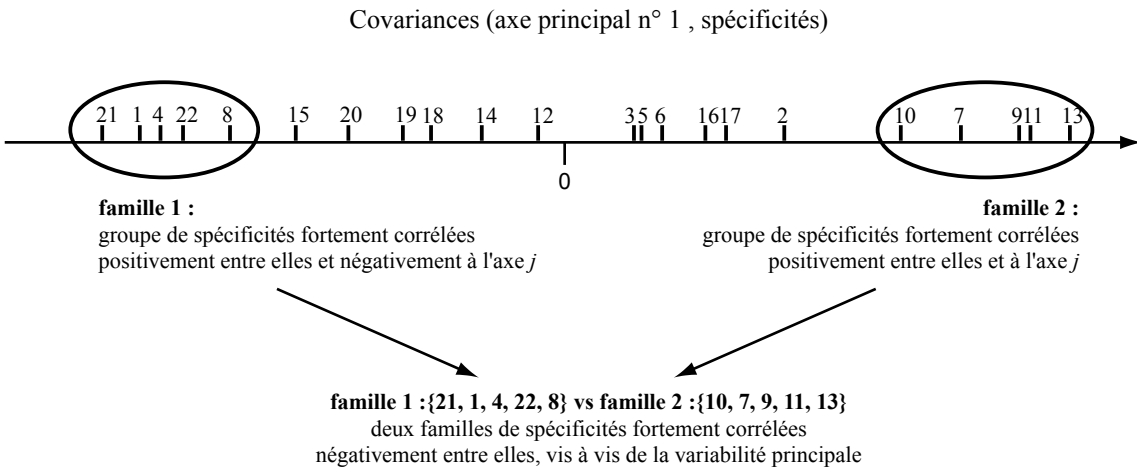

FIGURE *E.2* – *Exemple, pour un axe principal donné, de covariances avec les variables-dimensions initiales. La principale cause de variabilité est une opposition entre deux familles de spécificités.*

La figure E.2 représente un exemple, pour un axe principal donné, de ses covariances avec les variables-dimensions initiales. Deux familles de spécificités ont été sélectionnées comme les plus significatives de l'axe principal. Si par exemple l'axe principal est le premier, donc de plus grande variabilité, on peut en conclure que la principale cause de variabilité dans ces données est une opposition entre deux familles de spécificités : la famille 1 :$\{21, 1, 4, 22, 8\}$ dont les éléments ont un comportement proche et la famille 2 :$\{10, 7, 9, 11, 13\}$ dont, de même, les comportements sont proches.

Ceci peut se résumer par :

pour tout axe principal j :
- déterminer une famille 1 des spécificités s_k qui maximisent $cov\left(s_k, p_j\right)$
- déterminer une famille 2 des spécificités s_k qui minimisent $cov\left(s_k, p_j\right)$

Ces deux familles sont considérées comme les plus importantes pour expliquer le $j^{\text{ème}}$ axe principal. L'ensemble des $N \times N$ covariances entre composantes principales et spécificités doit donc être calculé. Notons $\mathbf{C}$ la matrice $N \times N$ de ces covariances. Son calcul peut être effectué matriciellement (et un sens plus précis donné à ses valeurs) par l'équivalence suivante : si l'on note

$$\mathbf{C}_{j,*} = \left\{cov\left(s_k, p_j\right)\right\}_{k=1,\ldots,N} \tag{E.12}$$

la $j^{\text{ème}}$ ligne de C, correspondant aux covariances par rapport à la $j^{\text{ème}}$ composante principale, on a :

$$\mathbf{C}_{j,*} = \left\{\frac{1}{n}\left(\overline{X}_{*,k}\right).\left(\overline{X}v_j\right)\right\}_{k=1,\ldots,N} = \frac{1}{n}\overline{X}^t\overline{X}v_j = \Sigma v_j = \lambda_j v_j \tag{E.13}$$

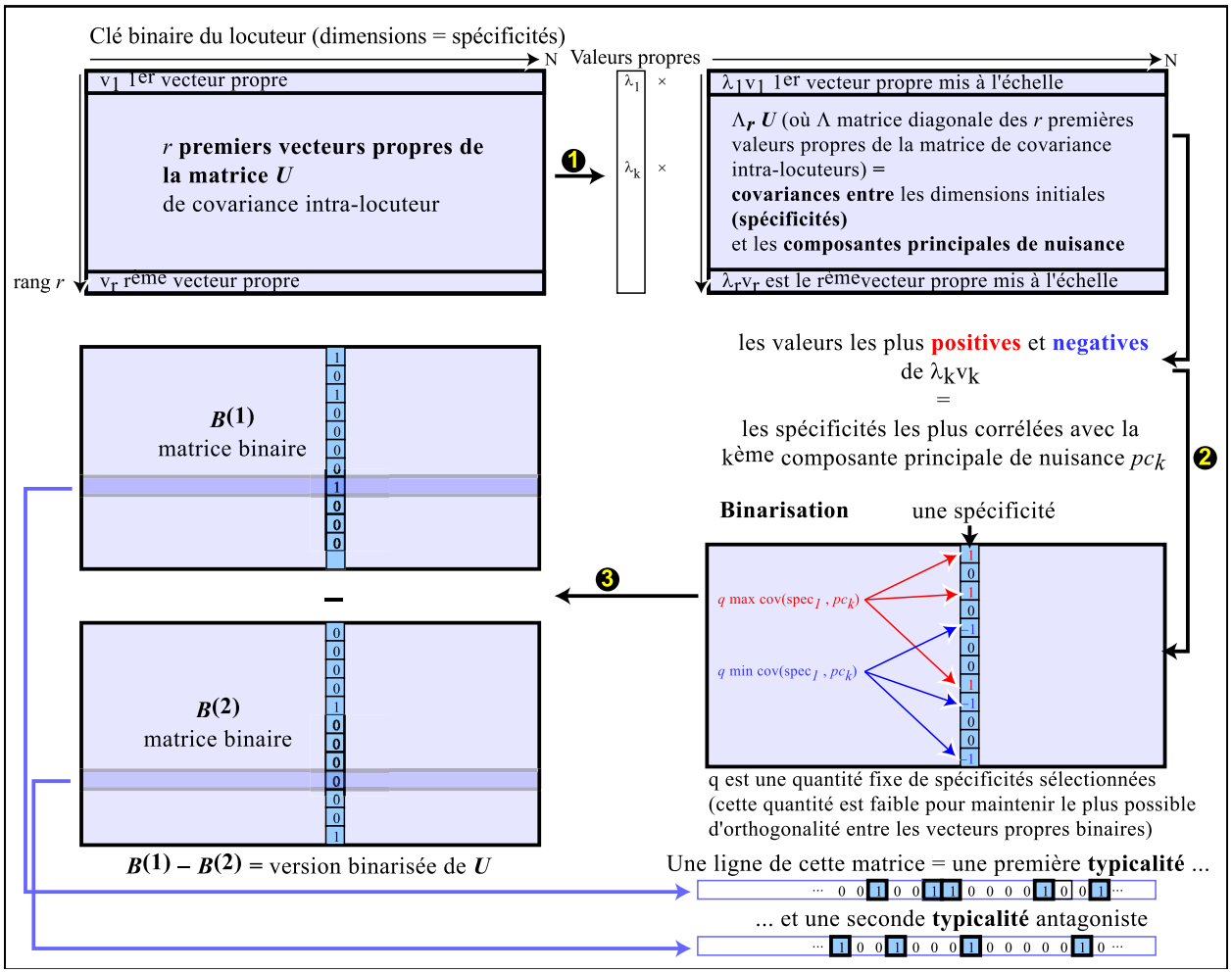

FIGURE E.3 – *Les étapes de génération de matrices de vecteurs propres binaires basées sur les typicalités.*

car v_j est vecteur propre de $\mathbf{\Sigma}$ pour la valeur propre λ_j.

Considérons maintenant le cas d'une réduction de dimension au rang r. La matrice $\mathbf{C}_r$, de dimensions $r \times N$, des covariances à calculer, peut donc s'écrire

$$\mathbf{C}_r = \mathbf{\Lambda}_r \mathbf{V}_r \tag{E.14}$$

où Λ_r est la matrice diagonale des plus grandes valeurs propres de $\mathbf{\Sigma}$ et $\mathbf{V}_r$ est la matrice $r \times N$ des r premiers vecteurs propres de $\mathbf{V}$ (en lignes). Comme indiqué précédemment, les valeurs les plus négatives et les plus positives de chaque ligne de C_r indiquent les spécificités les plus impliquées dans la génération de l'axe principal de variabilité correspondant.

La figure E.3 décrit l'étape de génération des covariances décrite précédemment.

Binarisation Considérant la matrice $\mathbf{C}_r$ précédente de covariances entre les r premières composantes principales et les spécificités initiales, il apparaît clairement que les deux familles de spécificités antagonistes dégagées pour chaque axe constituent des typicalités, résumant des informations essentielles de variabilité, utiles à un processus de discrimination.

Les difficultés à surmonter pour livrer ces typicalités sont de deux ordres :

- individuellement : pour un axe principal donné, comment sélectionner les spécificités des deux familles antagonistes ?
- globalement : comment obtenir un jeu de typicalités le plus indépendantes possibles, pour éviter des redondances perturbantes pour la mesure de similarité ?

La voie que nous avons choisie pour lever ces questions passe par la notion même de vecteurs propres : il est possible de considérer les deux typicalités par axe comme un vecteur "propre" de cet axe, au sens littéral comme mathématique du terme. Et donc, l'ensemble des typicalités comme une matrice de vecteurs propres : l'indépendance entre ses vecteurs devant être la plus forte possible.

L'extraction de typicalités se ramène alors à une binarisation de matrice de vecteurs propres. Considérant la matrice C_r, qui n'est autre que la matrice V_r des principaux vecteurs propres pondérés par leurs valeurs propres, il s'agit de binariser cette matrice pour former des typicalités. Individuellement, l'opposition entre les deux familles d'une ligne amène à proposer une décomposition binarisée de C_r en deux matrices binaires :

$$\mathbf{C}_r = \mathbf{B}_r^{(1)} - \mathbf{B}_r^{(2)} \tag{E.15}$$

où $\mathbf{B}_r^{(1)}$ et $\mathbf{B}_r^{(2)}$ sont à valeurs dans $\{0, 1\}$. Les $i^{\text{ème}}$ lignes de $\mathbf{B}_r^{(1)}$ et $\mathbf{B}_r^{(2)}$ contiennent les deux typicalités antagonistes du $i^{\text{ème}}$ axe principal. Notons que l'intersection de ces deux familles de spécificités est nécessairement vide.

Globalement, l'indépendance entre les typicalités par axe se traduit par l'absence (ou le minimum) de spécificités appartenant à deux typicalités distinctes. Cela se traduit matriciellement par le fait que, notant $(\mathbf{C}_r)_{i,*}$ et $(\mathbf{C}_r)_{i',*}$ les deux vecteurs-lignes i et i' de $\mathbf{C}_r$, leur produit scalaire $(\mathbf{C}_r)_{i,*} . (\mathbf{C}_r)_{i',*}$ doit être le plus proche de 0. il s'agit donc d'éviter la sélection d'une même spécificité par trop de lignes de $\mathbf{C}_r$.

Enfin, la première ligne de C_r correspond à la plus grande cause de variabilité : c'est donc celle à prendre le plus en compte. On s'attend donc à ce que le nombre de spécificités sélectionnées par ligne soit plus ou moins décroissant. L'application de cette contrainte est facilitée par le fait qu'une ligne j est constituée des coordonnées d'un vecteur propre multipliées par la valeur propre λ_j, les valeurs propres étant décroissantes. Les ordres de grandeur des covariances sont donc approximativement décroissantes de la ligne 1 à la ligne r.

Nous tentons de prendre en compte toutes ces remarques et spécifications et proposons l'algorithme suivant de binarisation et donc d'extraction de typicalités :

Contrairement à l'intuition première qui voudrait considérer les lignes de $\mathbf{C}_r$ de la première -la plus significative- à la dernière puis les binariser une à une avec le souci du nombre de 1 à affecter, nous procédons colonne par colonne : pour chaque colonne (donc chaque spécificité), les q valeurs les plus positives de $\mathbf{C}_r$ sont passées à 1, les q plus valeurs les plus négatives à -1 et les valeurs restantes à 0. La matrice obtenue est tronquée en deux parties binaires : les matrices $\mathbf{B}_r^{(1)}$ et $\mathbf{B}_r^{(2)}$.

Ce processus est illustré dans la suite de la figure 8.5 que nous commenterons plus bas. L'algorithme d'extraction de typicalité par binarisation de matrices de vecteurs propres s'écrit en deux étapes :

Etape 1 :

Calculer $\mathbf{C}_r = \mathbf{\Lambda}_r \mathbf{V}_r = \{\lambda_k v_k\}_k$
Pour chaque spécificité s_k
  mettre à 1 les q valeurs les plus positives de la $k^{\text{ème}}$ colonne de $\mathbf{C}_r$
  mettre à -1 les q valeurs les plus positives de la $k^{\text{ème}}$ colonne de $\mathbf{C}_r$
  mettre à 0 les autres valeurs de la $k^{\text{ème}}$ ligne.

où q est un nombre fixé de valeurs par spécificité. La matrice discrète $\mathbf{C}_r$ obtenue contient des $-1, 0$ et 1. Soient $\mathbf{B}_r^{(1)}$ et $\mathbf{B}_r^{(2)}$ les matrices $r \times N$ définies par :

$$\left(\mathbf{B}_r^{(1)}\right)_{ij} = 1 \quad \text{si } (\mathbf{C}_r)_{ij} = 1 \quad \text{et 0 sinon}$$
$$\left(\mathbf{B}_r^{(2)}\right) ij = 1 \quad \text{si } (\mathbf{C}_r)_{ij} = -1 \quad \text{et 0 sinon}$$

$\mathbf{C}_r$ devient donc égale à $\mathbf{B}_r^{(1)} - \mathbf{B}_r^{(2)}$, où $\mathbf{B}_r^{(1)}$ et $\mathbf{B}_r^{(2)}$ sont de matrices binaires.

La figure 8.5 décrit les étapes de cette binarisation.

Cet algorithme se justifie par les remarques suivantes :

- les variabilités, donc les covariances, tendant à se réduire à chaque axe, du premier au $r^{\text{ème}}$, les valeurs les plus volontiers sélectionnées pour une spécificité donnée seront celles des premiers axes, les plus significatifs. Le nombre de spécificités sélectionnées va donc décroître par axe.

- le nombre q d'axes sélectionnés par spécificité est fixe et réduit ($q << r$). De cette manière, l'indépendance entre deux lignes est proche d'être réalisée. Le produit scalaire de deux lignes i et i' de la matrice discrète C_r obtenue est égal à :

$$\sum_{k=1}^{N} (\mathbf{C}_r)_{i,k} (\mathbf{C}_r)_{i',k} \tag{E.16}$$

et la somme totale des produits scalaires des vecteurs distincts deux à deux ne contient au plus que $Nq\dfrac{q-1}{2}$ valeurs non nulles sur $Nr\dfrac{r-1}{2}$ possibles. Qui plus est,

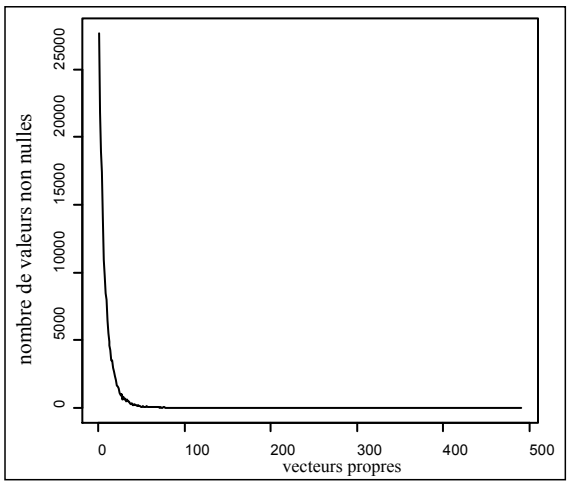

FIGURE *E.4* – *Spectre d'énergies issu de vecteurs propres binaires.*

ces valeurs non nulles (-1 ou 1) peuvent être opposées et donc tendre à s'annuler entre elles par sommation. Cela assure une semi-orthogonalité des vecteurs propres binarisés, donc leur semi-indépendance.

Le nombre de valeurs non nulles tendant à décroître de l'axe 1 à l'axe r, l'algorithme va se poursuivre par une seconde étape. Mais celle-ci s'appuie sur la notion de spectre de vecteurs propres binaires, que nous introduisons au paragraphe suivant.

Spectre de vecteurs propres binaires Un grand nombre de vecteurs propres (c'est à dire de lignes de la matrice $B_r^{(1)} - B_r^{(2)}$ extraites par l'étape 1 de l'algorithme) sont en fait nuls. La précaution de choisir une valeur q fixe et réduite au regard de r ainsi que la prédominance des premiers axes à fortes covariances en sont les causes. De plus, elle induit une décroissance des nombres de valeurs non nulles par vecteurs consécutifs.

La figure E.4 montre un exemple de ces nombres de valeurs non nulles pour les vecteurs-lignes de la matrice $\mathbf{B}_r^{(1)} - \mathbf{B}_r^{(2)}$ utilisée dans notre expérimentation (décrite plus bas. Comme le montre la figure, leur courbe pour les $r = 500$ dimensions de projection est décroissante et atteint rapidement la valeur zéro. Seule une centaine des lignes de la matrice sont non nulles. De plus, ce graphique présente l'aspect d'un *spectre d'énergie*. Le nombre de valeurs non nulles joue ainsi, dans l'espace binarisé, un rôle équivalent à celui du spectre des valeurs propres.

Il est à noter que la procédure de binarisation a significativement accentué l'énergie principale du spectre initial, ce qui montre un pouvoir de réduction conséquent de la méthode. Cette faculté aide également à déterminer le rang suffisant de projection (dans l'exemple de la figure E.4, à peu près 50) et à ne conserver que les typicalités les plus significatives.

L'algorithme se poursuit donc par :

Etape 2 :

> Calculer le nombre de valeurs n_i non nulles par ligne i de $\mathbf{B}_r^{(1)} - \mathbf{B}_r^{(2)}$
> Conserver les seules lignes dont n_i est supérieur à un seuil.

le seuil pouvant être à 0. Dans tous nos cas d'expérimentation, les lignes supprimées étaient les dernières.

Interprétation du résultat La matrice $r' \times N$ de vecteurs propres obtenue $\left(\mathbf{B}_{r'}^{(1)} - \mathbf{B}_{r'}^{(2)}\right)$, où $r' \leqslant r$, peut être interprétée ainsi :

- $\mathbf{B}_r^{(1)}$ et $\mathbf{B}_r^{(2)}$ sont des matrices binaires,
- chaque ligne de $\mathbf{B}_r^{(1)}$ indique un groupe de spécificités fortement corrélées, i.e. une typicalité,
- chaque ligne de $\mathbf{B}_r^{(2)}$ indique une autre typicalité,
- la soustraction $\mathbf{B}_{r'}^{(1)} - \mathbf{B}_{r'}^{(2)}$ mentionne une "opposition" entre ces deux typicalités.
- chaque ligne de $\left(\mathbf{B}_{r'}^{(1)} - \mathbf{B}_{r'}^{(2)}\right)$ dissocie les spécificités en trois parties $\{-1\}$, $\{0\}$, $\{1\}$ d'une manière similaire à un noeud à trois branches d'un arbre de classification.

E.3.2 Validation des typicalités obtenues par binarisation de vecteurs propres

Elaboration d'une similarité

Nous devons maintenant montrer que les typicalités obtenues par l'algorithme précédent vérifient les propriétés des liens inter-spécificités. Comme expliqué précédemment, une typicalité possède une puissance discriminante et cette puissance doit pouvoir être mesurée par application d'un produit total entre les paires de vecteurs binaires. Pour tester ce fait, nous avons expérimenté la technique NAP avec une similarité de la forme de l'équation E.8 mais où la matrice de variabilité $\mathbf{U}$ est remplacée par les matrices binaires $\mathbf{B}_{r'}^{(1)}$ et $\mathbf{B}_{r'}^{(2)}$ de typicalités. La validité -et la qualité- de ces liens pourra être évaluée par la performance calculée sur notre jeu d'évaluation. Pour obtenir une matrice "prête au scoring", deux points doivent être pris en compte :

- d'une part, les lignes de $\left(\mathbf{B}_{r'}^{(1)} - \mathbf{B}_{r'}^{(2)}\right)$ ne sont pas exactement indépendantes. C'est à dire que la matrice n'est pas exactement orthogonale, comme une matrice habituelle de vecteurs propres. Cependant, si q est réduit au regard du rang r, le très petit nombre de valeurs non nulles autorise la matrice :

$$\left(\mathbf{B}_{r'}^{(1)} - \mathbf{B}_{r'}^{(2)}\right)\left(\mathbf{B}_{r'}^{(1)} - \mathbf{B}_{r'}^{(2)}\right)^t \tag{E.17}$$

à être très proche de la diagonalité (ses vecteurs lignes orthogonaux entre eux).

- d'autre part, les vecteurs propres ne sont pas de norme 1. Normaliser leurs longueurs permet de standardiser leurs effets dans la similarité calculée. Pour cela, on considère la matrice diagonale D de dimensions $r \times r$ définie par :

$$\forall j = 1, ..., r \qquad \mathbf{D}_j = \sum_{k=1}^{N} \left(\left(\mathbf{B}_{r'}^{(1)} \right)_{j,k} + \left(\mathbf{B}_{r'}^{(2)} \right)_{j,k} \right) \tag{E.18}$$

qui relève le nombre de valeurs non nulles pour le $j^{\text{ème}}$ vecteur propre binarisé. La matrice de vecteurs propres binaire devient :

$$\left(\mathbf{B}_{r'}^{(1)} - \mathbf{B}_{r'}^{(2)} \right) \rightarrow \mathbf{D}^{-\frac{1}{2}} \left(\mathbf{B}_{r'}^{(1)} - \mathbf{B}_{r'}^{(2)} \right) \tag{E.19}$$

qui normalise ainsi les lignes.

Utilisant la quasi-orthogonalité de ces vecteurs propres binarisés, la similarité proposée entre deux vecteurs v_1 et v_2 est :

$$S\left(v_1, v_2\right) = v_1.v_2 - v_1.\mathbf{A}v_2 \tag{E.20}$$

où

$$\mathbf{A} = \left(\mathbf{B}_{r'}^{(1)} - \mathbf{B}_{r'}^{(2)} \right)^t \mathbf{D}^{-1} \left(\mathbf{B}_{r'}^{(1)} - \mathbf{B}_{r'}^{(2)} \right). \tag{E.21}$$

Les vecteurs v_1 et v_2 sont binaires, ainsi que les matrices $\mathbf{B}_{r'}^{(1)}$ et $\mathbf{B}_{r'}^{(2)}$. La matrice $\mathbf{D}$ est diagonale, avec des valeurs diagonales positives et entières. Il est à noter que le terme $v_1.\mathbf{A}v_2$ est un produit total. En effet, notant $\mathcal{L}_k^+$ et $\mathcal{L}_k^-$ les deux typicalités "opposées" du $k^{\text{ème}}$ vecteur propre et $card\left(.\right)$ la fonction cardinal, ce terme peut être réécrit :

$$v_1.\mathbf{A}v_2 = \sum_k \frac{\left(\delta_{\mathcal{L}_k^+}.v_1 - \delta_{\mathcal{L}_k^-}.v_1 \right) \times \left(\delta_{\mathcal{L}_k^+}.v_2 - \delta_{\mathcal{L}_k^-}.v_2 \right)}{card\left(\mathcal{L}_k^+ \cup \mathcal{L}_k^- \right)} \tag{E.22}$$

En développant le terme de droite de l'égalité, on constate que $v_1.\mathbf{A}v_2$ est composé de sommes et différences de produits totaux de la forme de l'équation E.2.

Notons que distribuer plus qu'un seul 1 et -1 par spécificité éloigne les vecteurs propres de l'orthogonalité, mais ajoute de nouvelles informations à la matrice.

Résultat fondamental

La validité de la formule mathématique de l'équation E.20 doit être démontrée : d'une part les coordonnées initiales des vecteurs propres ont été discrétisées, d'autre part l'équivalence entre les similarités E.7 et E.8 n'est plus vérifiée, du fait de la non-orthogonalité des nouveaux vecteurs propres. La similarité E.20 est le résultat d'étapes d'échantillonnage et de simplification. Justifier mathématiquement la similarité E.20 ne

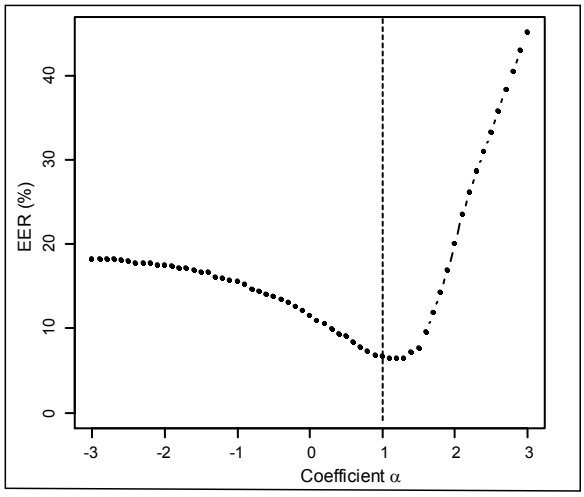

FIGURE *E.5* – *Validation expérimentale de la pertinence de la formule NAP binaire.*

peut se faire qu'en montrant qu'elle approxime à peu près correctement une valeur issue d'une projection vectorielle. C'est à dire que cette similarité n'est pas l'heureux résultat d'une fusion entre deux scores hétérogènes : la similarité initiale $v_1.v_2$ et la compensation $v_1.\mathbf{A}v_2$. Pour procéder, nous avons donc testé la fusion des termes séparés : un score de fusion S_α est produit par variation d'un coefficient réel α entre les deux similarités :

$$S_\alpha(v_1, v_2) = v_1.v2 - \alpha v_1.\mathbf{A}v_2 \tag{E.23}$$

La matrice $\mathbf{A}$ a été calculée à partir de la matrice $\mathbf{U}$ utilisée dans le cadre d'une compensation NAP de l'effet-session.

La figure E.5 montre le résultat de cette fusion pour l'expérience détaillée dans le paragraphe suivante. Les EER en % de S_α sont affichés pour α variant de -3 à 3 (qui s'est avéré l'intervalle contenant les meilleures performances). Comme on le voit, la courbe de performance montre clairement un minimum pour α proche de 1, c'est à dire pour le score E.20 proposé. Ce résultat montre le bon degré d'homogénéïté de la similarité E.20 : celle-ci n'est pas une combinaison de deux mesures distinctes, mais le résultat d'un équivalent dans le champ binaire de la technique NAP usuellement employée dans le domaine du continu.

Annexe F

Formulation du score 4-covariance

F.1 Sous l'hypothèse θ_{tar}

$$P(w_1, w_2|\theta_{\text{tar}}) = \int P(w_1, y_2)\, P(w_2|y_2)\, dy_2 et P(w_1, y_2) \tag{F.1}$$

$$= \int P(w_1, y_1)\, P(y_2|y_1)\, dy_1 \tag{F.2}$$

$$= \int \left(\int P(w_1, y_1)\, P(y_2|y_1)\, dy_1 \right) P(w_2|y_2)\, dy_2 \tag{F.3}$$

$$= \int \left(\int P(w_1|y_1)\, P(y_1)\, P(y_2|y_1)\, dy_1 \right) P(w_2|y_2)\, dy_2 \tag{F.4}$$

Dans le cas de la régression linéaire, on estime $y_2 = A(y_1 - \mu_1) + \mu_2 + \varepsilon$ où μ_2 et μ_1 moyennes de y_2 et y_1 et $\varepsilon \sim \mathcal{N}(0, \mathbf{M})$. La matrice A est obtenue par les moindres carrés (régression linéaire). Sur le fichier d'apprentissage, on calcule les moyennes locuteurs centrées des segments de type 1 (matrice $\overline{Y_1}$ des y_1 **en lignes**) et type 2 (matrice $\overline{Y_2}$ des y_2 **en lignes**).

$$\overline{Y_2} = \overline{Y_1}\mathbf{A} + \varepsilon \Longrightarrow \mathbf{A}_{\text{estimé}} \sim \left(\overline{Y_1}^t \overline{Y_1} \right)^{-1} \overline{Y_1}^t \overline{Y_1} \tag{F.5}$$

D'où $P(y_2|y_1) \sim \mathcal{N}(y_1|\mu_1 + \mathbf{A}(y_1 - \mu_1), cov(Y_2 - Y_1\mathbf{A}))$. On note $\mathbf{M} = cov(Y_2 - Y_1\mathbf{A}) = cov(y_2|y_1)$.

Dans le cas d'une fonction de transfert gaussienne, on écrit de même $P(y_2|y_1) \sim \mathcal{N}(y_2|\mu_2 + \mathbf{A}(y_1 - \mu_1), \mathbf{M})$

$$P\left(w_1, w_2|\theta_{\text{tar}}\right) = \int \left(\int \mathcal{N}\left(w_1|y_1, \mathbf{W}_1\right) \mathcal{N}\left(y_1|\mu_1, \mathbf{B}_1\right) \mathcal{N}\left(y_2|\mu_2 + \mathbf{A}\left(y_1 - \mu_1\right), \mathbf{M}\right) dy_1 \right) \tag{F.6}$$

$$\mathcal{N}\left(w_2|y_2, \mathbf{W}_2\right) dy_2 \tag{F.7}$$

$$P\left(w_1, w_2|\theta_{\text{tar}}\right) = \int \mathcal{A}\mathcal{N}\left(w_2|y_2, \mathbf{W}_2\right) dy_2 \tag{F.8}$$

où $\mathcal{A} = \int \mathcal{N}\left(w_1|y_1, \mathbf{W}_1\right) \mathcal{N}\left(y_1|\mu_1, \mathbf{B}_1\right) \mathcal{N}\left(y_2|\mu_2 + \mathbf{A}\left(y_1 - \mu_1\right), \mathbf{M}\right) dy_1$

$$\mathcal{A} = K \int \exp\left\{-\frac{1}{2}\mathcal{B}\right\} dy_1$$

où

$$\begin{aligned}
\mathcal{B} =& \left(w_1 - y_1\right)^t \mathbf{W}_1^{-1} \left(w_1 - y_1\right) + \left(y_1 - \mu_1\right)^t \mathbf{B}_1^{-1} \left(y_1 - \mu_1\right) && \text{(F.9)}\\
&+ \left(y_2 - \mu_2 - \mathbf{A}\left(y_1 - \mu_1\right)\right)^t \mathbf{M}^{-1} \left(y_2 - \mu_2 - \mathbf{A}\left(y_1 - \mu_1\right)\right) && \text{(F.10)}\\
=& \left(w_1 - y_1\right)^t \mathbf{W}_1^{-1} \left(w_1 - y_1\right) + \left(y_1 - \mu_1\right)^t \mathbf{B}_1^{-1} \left(y_1 - \mu_1\right) && \text{(F.11)}\\
&+ \left(-\mathbf{A}y_1 + y_2 - \mu_2 + \mathbf{A}\mu_1\right)^t \mathbf{M}^{-1} \left(-\mathbf{A}y_1 + y_2 - \mu_2 + \mathbf{A}\mu_1\right) && \text{(F.12)}\\
=& y_1 \left\{\mathbf{W}_1^{-1} + \mathbf{B}_1^{-1} + \mathbf{A}^t\mathbf{M}^{-1}\mathbf{A}\right\} y_1 && \text{(F.13)}\\
&+ \left\{-2w_1^t\mathbf{W}_1^{-1} - 2\mu_1^t\mathbf{B}_1^{-1} - 2\left(y_2 - \mu_2 + \mathbf{A}\mu_1\right)^t \mathbf{M}^{-1}\mathbf{A}\right\} y_1 && \text{(F.14)}\\
&+ \left\{w_1^t\mathbf{W}_1^{-1}w_1 + \mu_1^t\mathbf{B}_1^{-1}\mu_1 + \left(y_2 - \mu_2 + \mathbf{A}\mu_1\right)^t \mathbf{M}^{-1} \left(y_2 - \mu_2 + \mathbf{A}\mu_1\right)\right\} && \text{(F.15)}
\end{aligned}$$

donc $\mathcal{B} = y_1 Q_{11} y_1 + Q_1^t y_1 + Q_0^a$

où

$$\begin{aligned}
Q_{11} &= \mathbf{W}_1^{-1} + \mathbf{B}_1^{-1} + \mathbf{A}^t\mathbf{M}^{-1}\mathbf{A} && \text{(F.16)}\\
Q_1 &= \left\{-2w_1^t\mathbf{W}_1^{-1} - 2\mu_1^t\mathbf{B}_1^{-1} - 2\left(y_2 - \mu_2 + \mathbf{A}\mu_1\right)^t \mathbf{M}^{-1}\mathbf{A}\right\}^t && \text{(F.17)}\\
&= -2\mathbf{W}_1^{-1}w_1 - 2\mathbf{B}_1^{-1}\mu_1 - 2\mathbf{A}^t\mathbf{M}^{-1}\left(y_2 - \mu_2 + \mathbf{A}\mu_1\right) && \text{(F.18)}\\
Q_0^a &= w_1^t\mathbf{W}_1^{-1}w_1 + \mu_1^t\mathbf{B}_1^{-1}\mu_1 + \left(y_2 - \mu_2 + \mathbf{A}\mu_1\right)^t \mathbf{M}^{-1}\left(y_2 - \mu_2 + \mathbf{A}\mu_1\right) && \text{(F.19)}
\end{aligned}$$

d'où

$$\mathcal{A} = K \int \exp\left\{-\frac{1}{2}\left[y_1 Q_{11} y_1 + Q_1^t y_1 + Q_0^a\right]\right\} dy_1 = K \exp\left\{-\frac{1}{2}\left[Q_0^a - \frac{1}{4}Q_1^t Q_{11}^{-1} Q_1\right]\right\}$$

$$\mathcal{A} = K \int \exp\left\{-\frac{1}{2}\left[y_1 Q_{11} y_1 + Q_1^t y_1 + Q_0^a\right]\right\} dy_1 \tag{F.20}$$

$$= K \exp\left\{-\frac{1}{2}\left[Q_0^a - \frac{1}{4} Q_1^t Q_{11}^{-1} Q_1\right]\right\} \tag{F.21}$$

avec

$$Q_1 = -2\mathbf{W}_1^{-1} w_1 - 2\mathbf{B}_1^{-1}\mu_1 - 2\mathbf{A}^t\mathbf{M}^{-1}\left(y_2 - \mu_2 + \mathbf{A}\mu_1\right) \tag{F.22}$$

$$= -2\mathbf{A}^t\mathbf{M}^{-1} y_2 - 2\mathbf{W}_1^{-1} w_1 - 2\mathbf{B}_1^{-1}\mu_1 - 2\mathbf{A}^t\mathbf{M}^{-1}\left(-\mu_2 + \mathbf{A}\mu_1\right) \tag{F.23}$$

donc

$$Q_{11} = \mathbf{W}_1^{-1} + \mathbf{B}_1^{-1} + \mathbf{A}^t\mathbf{M}^{-1}\mathbf{A} \tag{F.24}$$

$$Q_1 = -2\mathbf{A}^t\mathbf{M}^{-1} y_2 - 2\mathbf{W}_1^{-1} w_1 - 2\mathbf{B}_1^{-1}\mu_1 - 2\mathbf{A}^t\mathbf{M}^{-1}\left(-\mu_2 + \mathbf{A}\mu_1\right) \tag{F.25}$$

$$Q_0^a = w_1^t\mathbf{W}_1^{-1} w_1 + \mu_1^t\mathbf{B}_1^{-1}\mu_1 + \left(y_2 - \mu_2 + \mathbf{A}\mu_1\right)^t \mathbf{M}^{-1}\left(y_2 - \mu_2 + \mathbf{A}\mu_1\right) \tag{F.26}$$

$$P\left(w_1, w_2|\theta_{\text{tar}}\right) = \int \mathcal{A}\mathcal{N}\left(w_2|y_2, \mathbf{W}_2\right) dy_2 \tag{F.27}$$

$$= K \int \exp\left\{-\frac{1}{2}\left[Q_0^a - \frac{1}{4} Q_1^t Q_{11}^{-1} Q_1\right]\right\} \mathcal{N}\left(w_2|y_2, \mathbf{W}_2\right) dy_2 \tag{F.28}$$

$$= K \int \exp\left\{-\frac{1}{2}\left[Q_0^a - \frac{1}{4} Q_1^t Q_{11}^{-1} Q_1 + \left(w_2 - y_2\right)^t \mathbf{W}_2^{-1}\left(w_2 - y_2\right)\right]\right\} dy_2 \tag{F.29}$$

$$= K \int \exp\left\{-\frac{1}{2}\mathcal{C}\right\} dy_2 \tag{F.30}$$

où

$$\mathcal{C}=Q_0^a - \frac{1}{4}Q_1^t Q_{11}^{-1} Q_1 + (w_2 - y_2)^t \mathbf{W}_2^{-1} (w_2 - y_2) \tag{F.31}$$

$$=w_1^t \mathbf{W}_1^{-1} w_1 + \mu_1^t \mathbf{B}_1^{-1} \mu_1 + (y_2 - \mu_2 + \mathbf{A}\mu_1)^t \mathbf{M}^{-1} (y_2 - \mu_2 + \mathbf{A}\mu_1) \tag{F.32}$$

$$- \left\{ \mathbf{A}^t \mathbf{M}^{-1} y_2 + \mathbf{W}_1^{-1} w_1 + \mathbf{B}_1^{-1} \mu_1 + \mathbf{A}^t \mathbf{M}^{-1} (-\mu_2 + \mathbf{A}\mu_1) \right\}^t \tag{F.33}$$

$$\left\{ \mathbf{W}_1^{-1} + \mathbf{B}_1^{-1} + \mathbf{A}^t \mathbf{M}^{-1} \mathbf{A} \right\}^{-1} \left\{ \mathbf{A}^t \mathbf{M}^{-1} y_2 + \mathbf{W}_1^{-1} w_1 + \mathbf{B}_1^{-1} \mu_1 + \mathbf{A}^t \mathbf{M}^{-1} (-\mu_2 + \mathbf{A}\mu_1) \right\} \tag{F.34}$$

$$+ (w_2 - y_2)^t \mathbf{W}_2^{-1} (w_2 - y_2) \tag{F.35}$$

$$y_2 \left\{ \mathbf{M}^{-1} - \mathbf{M}^{-1} \mathbf{A} \left\{ \mathbf{W}_1^{-1} + \mathbf{B}_1^{-1} + \mathbf{A}^t \mathbf{M}^{-1} \mathbf{A} \right\}^{-1} \mathbf{A}^t \mathbf{M}^{-1} + \mathbf{W}_2^{-1} \right\} y_2 \tag{F.36}$$

$$+ \left\{ \begin{array}{c} 2 (-\mu_2 + \mathbf{A}\mu_1)^t \mathbf{M}^{-1} - 2 \left\{ \mathbf{W}_1^{-1} w_1 + \mathbf{B}_1^{-1} \mu_1 + \mathbf{A}^t \mathbf{M}^{-1} (-\mu_2 + \mathbf{A}\mu_1) \right\}^t \\ \left\{ \mathbf{W}_1^{-1} + \mathbf{B}_1^{-1} + \mathbf{A}^t \mathbf{M}^{-1} \mathbf{A} \right\}^{-1} \mathbf{A}^t \mathbf{M}^{-1} - 2 w_2^t \mathbf{W}_2^{-1} \end{array} \right\} y_2 \tag{F.37}$$

$$+ \left\{ \begin{array}{c} w_1^t \mathbf{W}_1^{-1} w_1 + \mu_1^t \mathbf{B}_1^{-1} \mu_1 + (-\mu_2 + \mathbf{A}\mu_1)^t \mathbf{M}^{-1} (-\mu_2 + \mathbf{A}\mu_1) \\ + (w_2 - y_2)^t \mathbf{W}_2^{-1} (w_2 - y_2) + w_2^t \mathbf{W}_2^{-1} w_2 \end{array} \right\} \tag{F.38}$$

$$- \left\{ \mathbf{W}_1^{-1} w_1 + \mathbf{B}_1^{-1} \mu_1 + \mathbf{A}^t \mathbf{M}^{-1} (-\mu_2 + \mathbf{A}\mu_1) \right\}^t \left\{ \mathbf{W}_1^{-1} + \mathbf{B}_1^{-1} + \mathbf{A}^t \mathbf{M}^{-1} \mathbf{A} \right\}^{-1} \times \tag{F.39}$$

$$\left\{ \mathbf{W}_1^{-1} w_1 + \mathbf{B}_1^{-1} \mu_1 + \mathbf{A}^t \mathbf{M}^{-1} (-\mu_2 + \mathbf{A}\mu_1) \right\} \tag{F.40}$$

$$=y_1 Q_{22} y_2 + Q_2^t y_2 + Q_0^{b1} + Q_0^{b2} \tag{F.41}$$

où

$$Q_2 = \left\{ \begin{array}{c} 2 (-\mu_2 + \mathbf{A}\mu_1)^t \mathbf{M}^{-1} - 2 \left\{ \mathbf{W}_1^{-1} w_1 + \mathbf{B}_1^{-1} \mu_1 + \mathbf{A}^t \mathbf{M}^{-1} (-\mu_2 + \mathbf{A}\mu_1) \right\}^t \\ \left\{ \mathbf{W}_1^{-1} + \mathbf{B}_1^{-1} + \mathbf{A}^t \mathbf{M}^{-1} \mathbf{A} \right\}^{-1} \mathbf{A}^t \mathbf{M}^{-1} - 2 w_2^t \mathbf{W}_2^{-1} \end{array} \right\}^t \tag{F.42}$$

et donc

$$Q_{22} = \mathbf{M}^{-1} - \mathbf{M}^{-1}\mathbf{A}\left\{\mathbf{W}_1^{-1} + B_1^{-1} + \mathbf{A}^t\mathbf{M}^{-1}\mathbf{A}\right\}^{-1}\mathbf{A}^t\mathbf{M}^{-1} + \mathbf{W}_2^{-1} \tag{F.43}$$

$$Q_2 = 2\mathbf{M}^{-1}\left(-\mu_2 + \mathbf{A}\mu_1\right) - 2\mathbf{M}^{-1}\mathbf{A}\left\{\mathbf{W}_1^{-1} + B_1^{-1} + \mathbf{A}^t\mathbf{M}^{-1}\mathbf{A}\right\}^{-1} \tag{F.44}$$

$$\left\{\mathbf{W}_1^{-1}w_1 + B_1^{-1}\mu_1 + \mathbf{A}^t\mathbf{M}^{-1}\left(-\mu_2 + \mathbf{A}\mu_1\right)\right\} - 2\mathbf{W}_2^{-1}w_2 \tag{F.45}$$

$$Q_0^{b1} = w_1^t\mathbf{W}_1^{-1}w_1 + \mu_1^t B_1^{-1}\mu_1 + \left(-\mu_2 + \mathbf{A}\mu_1\right)^t\mathbf{M}^{-1}\left(-\mu_2 + \mathbf{A}\mu_1\right) + w_2^t\mathbf{W}_2^{-1}w_2 \tag{F.46}$$

$$Q_0^{b2} = -\left\{\mathbf{W}_1^{-1}w_1 + B_1^{-1}\mu_1 + \mathbf{A}^t\mathbf{M}^{-1}\left(-\mu_2 + \mathbf{A}\mu_1\right)\right\}^t\left\{\mathbf{W}_1^{-1} + B_1^{-1} + \mathbf{A}^t\mathbf{M}^{-1}\mathbf{A}\right\}^{-1} \tag{F.47}$$

$$\left\{\mathbf{W}_1^{-1}w_1 + B_1^{-1}\mu_1 + \mathbf{A}^t\mathbf{M}^{-1}\left(-\mu_2 + \mathbf{A}\mu_1\right)\right\} \tag{F.48}$$

d'où

$$P\left(w_1, w_2|\theta_{\text{tar}}\right) = K\int \exp\left\{-\frac{1}{2}\left[y_1 Q_{11} y_1 + Q_1^t y_1 + Q_0^a\right]\right\} dy_1 \tag{F.49}$$

$$= K\exp\left\{-\frac{1}{2}\left[Q_0^a - \frac{1}{4}Q_1^t Q_{11}^{-1} Q_1\right]\right\} \tag{F.50}$$

$$= K\int \exp\left\{-\frac{1}{2}\mathcal{C}\right\} dy_2 = K\exp\left\{-\frac{1}{2}\left[Q_0^{b1} + Q_0^{b2} - \frac{1}{4}Q_2^t Q_{22}^{-1} Q_2\right]\right\} \tag{F.51}$$

$$= K - \frac{1}{2}\left[Q_0^{b1} + Q_0^{b2} - \frac{1}{4}Q_2^t Q_{22}^{-1} Q_2\right] \tag{F.52}$$

$$\sim \frac{1}{4}Q_2^t Q_{22}^{-1} Q_2 - Q_0^{b1} - Q_0^{b2} \tag{F.53}$$

$$Q_{22} = \mathbf{M}^{-1} - \mathbf{M}^{-1}\mathbf{A}\left\{\mathbf{W}_1^{-1} + B_1^{-1} + \mathbf{A}^t\mathbf{M}^{-1}\mathbf{A}\right\}^{-1}\mathbf{A}^t\mathbf{M}^{-1} + \mathbf{W}_2^{-1} \tag{F.54}$$

$$Q_2 = 2\mathbf{M}^{-1}\left(-\mu_2 + \mathbf{A}\mu_1\right) - 2\mathbf{M}^{-1}\mathbf{A}\left\{\mathbf{W}_1^{-1} + \mathbf{B}_1^{-1} + \mathbf{A}^t\mathbf{M}^{-1}\mathbf{A}\right\}^{-1} \tag{F.55}$$

$$\left\{\mathbf{W}_1^{-1}w_1 + \mathbf{B}_1^{-1}\mu_1 + \mathbf{A}^t\mathbf{M}^{-1}\left(-\mu_2 + \mathbf{A}\mu_1\right)\right\} - 2\mathbf{W}_2^{-1}w_2 \tag{F.56}$$

$$Q_0^{b1} = w_1^t\mathbf{W}_1^{-1}w_1 + \mu_1^t\mathbf{B}_1^{-1}\mu_1 + \left(-\mu_2 + \mathbf{A}\mu_1\right)^t\mathbf{M}^{-1}\left(-\mu_2 + \mathbf{A}\mu_1\right) + w_2^t\mathbf{W}_2^{-1}w_2 \tag{F.57}$$

$$Q_0^{b2} = -\left\{\mathbf{W}_1^{-1}w_1 + \mathbf{B}_1^{-1}\mu_1 + \mathbf{A}^t\mathbf{M}^{-1}\left(-\mu_2 + \mathbf{A}\mu_1\right)\right\}^t\left\{\mathbf{W}_1^{-1} + \mathbf{B}_1^{-1} + \mathbf{A}^t\mathbf{M}^{-1}\mathbf{A}\right\}^{-1} \tag{F.58}$$

$$\left\{\mathbf{W}_1^{-1}w_1 + \mathbf{B}_1^{-1}\mu_1 + \mathbf{A}^t\mathbf{M}^{-1}\left(-\mu_2 + \mathbf{A}\mu_1\right)\right\} \tag{F.59}$$

$$Q_2 = 2\mathbf{M}^{-1}\left(-\mu_2 + \mathbf{A}\mu_1\right) - 2\mathbf{M}^{-1}\mathbf{A}\left\{\mathbf{W}_1^{-1} + \mathbf{B}_1^{-1} + \mathbf{A}^t\mathbf{M}^{-1}\mathbf{A}\right\}^{-1} \tag{F.60}$$

$$\left\{\mathbf{W}_1^{-1}w_1 + \mathbf{B}_1^{-1}\mu_1 + \mathbf{A}^t\mathbf{M}^{-1}\left(-\mu_2 + \mathbf{A}\mu_1\right)\right\} - 2\mathbf{W}_2^{-1}w_2 \tag{F.61}$$

$$Q_{22} = \mathcal{B} \tag{F.62}$$

$$Q_2 = 2a - 2\mathcal{C}\left(\mathbf{W}_1^{-1}w_1 + b\right) - 2\mathbf{W}_2^{-1}w_2 \tag{F.63}$$

$$Q_0^{b1} = w_1^t\mathbf{W}_1^{-1}w_1 + c + w_2^t\mathbf{W}_2^{-1}w_2 \tag{F.64}$$

$$Q_0^{b2} = -\left(\mathbf{W}_1^{-1}w_1 + d\right)^t\mathcal{D}\left(\mathbf{W}_1^{-1}w_1 + e\right) \tag{F.65}$$

où

$$\mathcal{B} = \mathbf{M}^{-1} - \mathbf{M}^{-1}\mathbf{A}\left\{\mathbf{W}_1^{-1} + \mathbf{B}_1^{-1} + \mathbf{A}^t\mathbf{M}^{-1}\mathbf{A}\right\}^{-1}\mathbf{A}^t\mathbf{M}^{-1} + \mathbf{W}_2^{-1} \tag{F.66}$$

$$a = \mathbf{M}^{-1}\left(-\mu_2 + \mathbf{A}\mu_1\right) \tag{F.67}$$

$$\mathcal{C} = \mathbf{M}^{-1}\mathbf{A}\left\{\mathbf{W}_1^{-1} + \mathbf{B}_1^{-1} + \mathbf{A}^t\mathbf{M}^{-1}\mathbf{A}\right\}^{-1} \tag{F.68}$$

$$b = \mathbf{B}_1^{-1}\mu_1 + \mathbf{A}^t\mathbf{M}^{-1}\left(-\mu_2 + \mathbf{A}\mu_1\right) \tag{F.69}$$

$$c = \mu_1^t\mathbf{B}_1^{-1}\mu_1 + \left(-\mu_2 + \mathbf{A}\mu_1\right)^t\mathbf{M}^{-1}\left(-\mu_2 + \mathbf{A}\mu_1\right) \tag{F.70}$$

$$d = \mathbf{B}_1^{-1}\mu_1 + \mathbf{A}^t\mathbf{M}^{-1}\left(-\mu_2 + \mathbf{A}\mu_1\right) \tag{F.71}$$

$$\mathcal{D} = \left\{\mathbf{W}_1^{-1} + \mathbf{B}_1^{-1} + \mathbf{A}^t\mathbf{M}^{-1}\mathbf{A}\right\}^{-1} \tag{F.72}$$

et donc

$$\log P\left(w_1, w_2|\theta_{\text{tar}}\right) \sim \frac{1}{4}Q_2^tQ_{22}^{-1}Q_2 - Q_0^{b1} - Q_0^{b2} \tag{F.73}$$

$$\sim \left(a - \mathcal{C}b - \mathcal{C}\mathbf{W}_1^{-1}w_1 - \mathbf{W}_2^{-1}w_2\right)^t\mathcal{B}^{-1}\left(a - \mathcal{C}b - \mathcal{C}\mathbf{W}_1^{-1}w_1 - \mathbf{W}_2^{-1}w_2\right) \tag{F.74}$$

$$- w_1^t\mathbf{W}_1^{-1}w_1 - c - w_2^t\mathbf{W}_2^{-1}w_2 + \left(\mathbf{W}_1^{-1}w_1 + b\right)^t\mathcal{D}\left(\mathbf{W}_1^{-1}w_1 + b\right) \tag{F.75}$$

où

$$\mathcal{B} = \mathbf{M}^{-1} - \mathbf{M}^{-1}\mathbf{A}\left\{\mathbf{W}_1^{-1} + \mathbf{B}_1^{-1} + \mathbf{A}^t\mathbf{M}^{-1}\mathbf{A}\right\}^{-1}\mathbf{A}^t\mathbf{M}^{-1} + \mathbf{W}_2^{-1} \tag{F.76}$$

$$a = \mathbf{M}^{-1}\left(-\mu_2 + \mathbf{A}\mu_1\right) \tag{F.77}$$

$$\mathcal{C} = \mathbf{M}^{-1}\mathbf{A}\left\{\mathbf{W}_1^{-1} + \mathbf{B}_1^{-1} + \mathbf{A}^t\mathbf{M}^{-1}\mathbf{A}\right\}^{-1} \tag{F.78}$$

$$b = \mathbf{B}_1^{-1}\mu_1 + \mathbf{A}^t\mathbf{M}^{-1}\left(-\mu_2 + \mathbf{A}\mu_1\right) \tag{F.79}$$

$$c = \mu_1^t\mathbf{B}_1^{-1}\mu_1 + \left(-\mu_2 + \mathbf{A}\mu_1\right)^t\mathbf{M}^{-1}\left(-\mu_2 + \mathbf{A}\mu_1\right) \tag{F.80}$$

$$\mathcal{D} = \left\{\mathbf{W}_1^{-1} + \mathbf{B}_1^{-1} + \mathbf{A}^t\mathbf{M}^{-1}\mathbf{A}\right\}^{-1} \tag{F.81}$$

Par factorisation des termes quadratique, linéaire et constant, on obtient

$$\log P\left(w_1, w_2|\theta_{\text{tar}}\right) \sim \left(a - \mathcal{C}b - \mathcal{C}\mathbf{W}_1^{-1}w_1 - \mathbf{W}_2^{-1}w_2\right)^t\mathcal{B}^{-1}\left(a - \mathcal{C}b - \mathcal{C}\mathbf{W}_1^{-1}w_1 - \mathbf{W}_2^{-1}w_2\right) \tag{F.82}$$

$$- w_1^t\mathbf{W}_1^{-1}w_1 - c - w_2^t\mathbf{W}_2^{-1}w_2 + \left(\mathbf{W}_1^{-1}w_1 + b\right)^t\mathcal{D}\left(\mathbf{W}_1^{-1}w_1 + b\right) \tag{F.83}$$

$$\sim w_1^t\left\{\mathbf{W}_1^{-1}\mathcal{C}^t\mathcal{B}^{-1}\mathcal{C}\mathbf{W}_1^{-1} - \mathbf{W}_1^{-1} + \mathbf{W}_1^{-1}\mathcal{D}\mathbf{W}_1^{-1}\right\}w_1 \tag{F.84}$$

$$+ w_2^t\left\{\mathbf{W}_2^{-1}\mathcal{B}^{-1}\mathbf{W}_2^{-1} - \mathbf{W}_2^{-1}\right\}w_2 \tag{F.85}$$

$$+ 2w_1^t\left\{\mathbf{W}_1^{-1}\mathcal{C}^t\mathcal{B}^{-1}\mathbf{W}_2^{-1}\right\}w_2 + \left\{-2\left(a - \mathcal{C}b\right)^t\mathcal{B}^{-1}\mathcal{C}\mathbf{W}_1^{-1} + 2b^t\mathcal{D}\mathbf{W}_1^{-1}\right\}w_1 \tag{F.86}$$

$$+ \left\{-2\left(a - \mathcal{C}b\right)^t\mathcal{B}^{-1}\mathbf{W}_2^{-1}\right\}w_2 + \left\{\left(a - \mathcal{C}b\right)^t\mathcal{B}^{-1}\left(a - \mathcal{C}b\right) - c + b^t\mathcal{D}b\right\} \tag{F.87}$$

et l'on en conclut que

$$\log P\left(w_1, w_2|\theta_{\text{tar}}\right) \sim w_1^tN_{11}^{\text{tar}}w_1 + w_2^tN_{22}^{\text{tar}}w_2 + 2w_1^tN_{12}^{\text{tar}}w_2 + n_1^{\text{tar}}w_1 + n_2^{\text{tar}}w_2 \tag{F.88}$$

où

$$N_{11}^{\text{tar}} = \mathbf{W}_1^{-1}\mathcal{C}^t\mathcal{B}^{-1}\mathcal{C}\mathbf{W}_1^{-1} - \mathbf{W}_1^{-1} + \mathbf{W}_1^{-1}\mathcal{D}\mathbf{W}_1^{-1} \tag{F.89}$$

$$N_{22}^{\text{tar}} = \mathbf{W}_2^{-1}\mathcal{B}^{-1}\mathbf{W}_2^{-1} - \mathbf{W}_2^{-1} \tag{F.90}$$

$$N_{12}^{\text{tar}} = \mathbf{W}_1^{-1}\mathcal{C}^t\mathcal{B}^{-1}\mathbf{W}_2^{-1} \tag{F.91}$$

$$n_1^{\text{tar}} = -2\left(a - \mathcal{C}b\right)^t\mathcal{B}^{-1}\mathcal{C}\mathbf{W}_1^{-1} + 2b^t\mathcal{D}\mathbf{W}_1^{-1} \tag{F.92}$$

$$n_2^{\text{tar}} = -2\left(a - \mathcal{C}b\right)^t\mathcal{B}^{-1}\mathbf{W}_2^{-1} \tag{F.93}$$

avec

$$\mathcal{B} = \mathbf{M}^{-1} - \mathbf{M}^{-1}\mathbf{A}\left\{\mathbf{W}_1^{-1} + \mathbf{B}_1^{-1} + \mathbf{A}^t\mathbf{M}^{-1}\mathbf{A}\right\}^{-1}\mathbf{A}^t\mathbf{M}^{-1} + \mathbf{W}_2^{-1} \tag{F.94}$$

$$a = \mathbf{M}^{-1}\left(-\mu_2 + \mathbf{A}\mu_1\right) \tag{F.95}$$

$$\mathcal{C} = \mathbf{M}^{-1}\mathbf{A}\left\{\mathbf{W}_1^{-1} + \mathbf{B}_1^{-1} + \mathbf{A}^t\mathbf{M}^{-1}\mathbf{A}\right\}^{-1} \tag{F.96}$$

$$b = \mathbf{B}_1^{-1}\mu_1 + \mathbf{A}^t\mathbf{M}^{-1}\left(-\mu_2 + \mathbf{A}\mu_1\right) \tag{F.97}$$

$$c = \mu_1^t\mathbf{B}_1^{-1}\mu_1 + \left(-\mu_2 + \mathbf{A}\mu_1\right)^t\mathbf{M}^{-1}\left(-\mu_2 + \mathbf{A}\mu_1\right) \tag{F.98}$$

$$\mathcal{D} = \left\{\mathbf{W}_1^{-1} + \mathbf{B}_1^{-1} + \mathbf{A}^t\mathbf{M}^{-1}\mathbf{A}\right\}^{-1} \tag{F.99}$$

F.2 Sous l'hypothèse θ_{non}

$$P\left(w_1, w_2|\theta_{\text{non}}\right) = \left(\int P\left(w_1|y_1\right)P\left(y_1\right)dy_1\right)\left(\int P\left(w_2|y_2\right)P\left(y_2\right)dy_2\right) \tag{F.100}$$

$$= \int \exp\left(-\frac{1}{2}\left\{\left(w_1 - y_1\right)^t\mathbf{W}_1^{-1}\left(w_1 - y_1\right) + \left(y_1 - \mu_1\right)^t\mathbf{B}_1^{-1}\left(y_1 - \mu_1\right)\right\}\right)dy_1 \tag{F.101}$$

$$\times \int \exp\left(-\frac{1}{2}\left\{\left(w_2 - y_2\right)^t\mathbf{W}_2^{-1}\left(w_2 - y_2\right) + \left(y_2 - \mu_2\right)^t\mathbf{B}_2^{-1}\left(y_2 - \mu_2\right)\right\}\right)dy_2 \tag{F.102}$$

$$\left(w_1 - y_1\right)^t\mathbf{W}_1^{-1}\left(w_1 - y_1\right) + \left(y_1 - \mu_1\right)^t\mathbf{B}_1^{-1}\left(y_1 - \mu_1\right) = y_1\left\{\mathbf{W}_1^{-1} + \mathbf{B}_1^{-1}\right\}y_1 \tag{F.103}$$

$$+ y_1\left\{-2\mathbf{W}_1^{-1}w_1 - 2\mathbf{B}_1^{-1}\mu_1\right\} \tag{F.104}$$

$$+ \left\{w_1^t\mathbf{W}_1^{-1}w_1 + \mu_1^t\mathbf{B}_1^{-1}\mu_1\right\} \tag{F.105}$$

d'où

$$\log\int \exp\left(-\frac{1}{2}\left\{\left(w_1 - y_1\right)^t\mathbf{W}_1^{-1}\left(w_1 - y_1\right) + \left(y_1 - \mu_1\right)^t\mathbf{B}_1^{-1}\left(y_1 - \mu_1\right)\right\}\right)dy_1 = -\frac{1}{2}\times\mathcal{S} \tag{F.106}$$

où

$$\mathcal{S} = \left\{ w_1^t \mathbf{W}_1^{-1} w_1 + \mu_1^t \mathbf{B}_1^{-1} \mu_1 - \left(\mathbf{W}_1^{-1} w_1 + \mathbf{B}_1^{-1} \mu_1 \right)^t \left(\mathbf{W}_1^{-1} + \mathbf{B}_1^{-1} \right)^{-1} \left(\mathbf{W}_1^{-1} w_1 + \mathbf{B}_1^{-1} \mu_1 \right) \right\} \tag{F.107}$$

et

$$\log P\left(w_1, w_2 | \theta_{\text{non}}\right) \sim \left(\mathbf{W}_1^{-1} w_1 + \mathbf{B}_1^{-1} \mu_1 \right)^t \left(\mathbf{B}_1^{-1} + \mathbf{W}_1^{-1} \right)^{-1} \left(\mathbf{W}_1^{-1} w_1 + \mathbf{B}_1^{-1} \mu_1 \right) \tag{F.108}$$

$$- w_1 \mathbf{W}_1^{-1} w_1 - \mu_1^t \mathbf{B}_1^{-1} \mu_1 + \left(\mathbf{W}_2^{-1} w_2 + \mathbf{B}_2^{-1} \mu_2 \right)^t \left(\mathbf{B}_2^{-1} + \mathbf{W}_2^{-1} \right)^{-1} \tag{F.109}$$

$$\left(\mathbf{W}_2^{-1} w_2 + \mathbf{B}_2^{-1} \mu_2 \right) - w_2 \mathbf{W}_2^{-1} w_2 - \mu_2^t \mathbf{B}_2^{-1} \mu_2 \tag{F.110}$$

On en conclut que

$$\log P\left(w_1, w_2 | \theta_{\text{non}}\right) \sim w_1^t N_{11}^{\text{non}} w_1 + w_2^t N_{22}^{\text{non}} w_2 + n_1^{\text{non}} w_1 + n_2^{\text{non}} w_2 + K \tag{F.111}$$

où

$$N_{11}^{\text{non}} = \mathbf{W}_1^{-1} \left(\mathbf{B}_1^{-1} + \mathbf{W}_1^{-1} \right)^{-1} \mathbf{W}_1^{-1} - \mathbf{W}_1^{-1} \tag{F.112}$$

$$N_{22}^{\text{non}} = \mathbf{W}_2^{-1} \left(\mathbf{B}_2^{-1} + \mathbf{W}_2^{-1} \right)^{-1} \mathbf{W}_2^{-1} - \mathbf{W}_2^{-1} \tag{F.113}$$

$$n_1^{\text{non}} = 2 \left(\mathbf{B}_1^{-1} \mu_1 \right)^t \left(\mathbf{B}_1^{-1} + \mathbf{W}_1^{-1} \right)^{-1} \mathbf{W}_1^{-1} \tag{F.114}$$

$$n_2^{\text{non}} = 2 \left(\mathbf{B}_2^{-1} \mu_2 \right)^t \left(\mathbf{B}_2^{-1} + \mathbf{W}_2^{-1} \right)^{-1} \mathbf{W}_2^{-1} \tag{F.115}$$

Annexe G

Formulation et propriété des scorings après normalisation

Nous présentons des re-formulations des expressions de différents scorings (Mahalanobis, two-covariance) après normalisation des vecteurs. Ces nouvelles expressions simplifiées mettent en avant les éléments-clés de ces scorings et permettent également de mieux cerner la pertinence du modèle théorique standard unitaire.

Les différentes formulations de scores que nous avons présentées peuvent être simplifiées après normalisation EFR ou SphN. En dehors de leur aspect fonctionnel, ces réécritures permettent de mieux identifier les types de proximités qui sont mesurées durant leurs calculs. Ces simplifications s'appuient sur le fait suivant : les algorithmes de normalisation font tendre les données vers un modèle théorique. Par exemple, après normalisation EFR, les vecteurs w d'apprentissage vérifient :

- $\|w\| = \sqrt{p}$
- $\mu \to 0$
- $cov\,(w) \to \mathbf{I}$
- $\mathbf{B}$ et $\mathbf{W}$ tendent à former une base commune d'*eigenfactors*.

Ce dernier point peut s'exprimer ainsi : soit $\left\{\lambda_j\right\}_{j=1,\ldots,p}$ le spectre des valeurs propres décroissantes de $\mathbf{B}$. Si l'on exprime les i-vectors dans la base de $\mathbf{B}$ après EFR (en effectuant une rotation suivant la base de ses vecteurs propres), alors on tend vers la configuration suivante, illustrée sur la figure 4.4 :

- $\mathbf{B}$ est une matrice diagonale $\mathbf{\Lambda}$, de diagonale $\left\{\mathbf{\Lambda}_{j,j}\right\}_{j=1,\ldots,p} = \left\{\lambda_j\right\}_{j=1,\ldots,p}$
- $\mathbf{W}$ est la matrice diagonale $\mathbf{I} - \mathbf{\Lambda}$
- Donc $\mathbf{B}^{-1}$ est la matrice diagonale $\mathbf{\Lambda}^{-1}$ et $\mathbf{W}^{-1}$ la matrice diagonale $(\mathbf{I} - \mathbf{\Lambda})^{-1}$

Les scorings de Mahalanobis, two-covariance et, pour certains cas, de la PLDA peuvent alors s'écrire de manière simplifiée. Ces simplifications supposent que les assertions précédentes soient exactement vérifiées. Les appliquer aux données empiriques n'est pas seulement commode : nous supposons qu'elles constituent des estimations

plus fiables des paramètres statistiques.

G.1 Simplification du scoring imposteur après EFR

Dans tous les scorings, la mesure correspondant à l'hypothèse θ_{non} (imposteur :les deux observations n'appartiennent pas au même locuteur) est déterminée par le logarithme de $P(w_1, w_2|\theta_{\text{non}})$ où $P(.)$ est une mesure bayesienne de vraisemblance dépendant du modèle choisi. Les i-vectors w_1 et w_2, étant supposés issus de deux locuteurs distincts, sont considérés comme indépendants. La mesure précédente s'écrit alors :

$$\log P(w_1, w_2|\theta_{\text{non}}) = \log P(w_1) P(w_2) \tag{G.1}$$

Ce résultat utilise le produit des probabilités $P(w_1)$ et $P(w_2)$. Or, après application de l'algorithme de normalisation EFR, il a été montré que les i-vectors tendant vers une loi normale standard sur la surface d'une hypersphère, leurs probabilités tendent vers une valeur constante. On a, pour $i = 1$ ou 2 :

$$P(w_i) \underset{i \to +\infty}{\to} \mathcal{N}(w_i|0, \mathbf{I}) = (2\pi)^{\frac{p}{2}} \exp(-p) \tag{G.2}$$

au fur et à mesure des itérations d'EFR et $\log P(w_1, w_2|\theta_{\text{non}})$ tend à être constant.

La décision repose donc en théorie sur la seule mesure de θ_{tar} (cible : les deux observations appartiennent au même locuteur). La convergence des i-vectors vers le modèle normal standard n'étant pas exacte, nous présenterons dans cette section les résultats d'une comparaison entre le scoring initial du modèle two-covariance et sa version simplifiée.

G.2 Score de Mahalanobis

G.2.1 Expression après EFR

En l'absence de toute information sur la variabilité globale des facteurs locuteurs, ce scoring très simple est pourtant optimal dans un cadre gaussien, après EFR. Considérons un i-vector w dont on veut savoir s'il appartient au locuteur s dont on ne dispose que d'une observation, notée w_{tar}. Considérons d'autre part que nous ne nous disposons d'aucune information sur la distribution des classes-locuteur. On a alors :

$$\begin{aligned} \log P(s|w) &= \log \frac{P(w|s) P(s)}{P(w)} \\ &= \log \frac{\mathcal{N}(w|w_{\text{tar}}, \mathbf{W}) P(s)}{\mathcal{N}(w|0, \mathbf{I})} \end{aligned} \tag{G.3}$$

la seule estimation de la moyenne des vecteurs de s étant fournie par w_{tar}. Comme $\mathcal{N}(w|0, \mathbf{I})$ est une constante et les locuteurs étant équiprobables, l'équation précédente devient :

$$\log P(s|w) = C - \frac{1}{2}(w - w_{\text{tar}})^t \mathbf{W}^{-1}(w - w_{\text{tar}}) \tag{G.4}$$

(C étant une constante), où apparaît le score de Mahalanobis de l'équation 2.58 à une constante près.

Après normalisation EFR, calcul de $\mathbf{W}$ et passage des données dans sa base de vecteurs propres, le score de Mahalanobis peut s'écrire :

$$\begin{aligned} score_{\text{Maha}}(w_1, w_2) &= -(w_1 - w_2)^t \mathbf{W}^{-1}(w_1 - w_2) \\ &= \sum_{j=1}^{p} \frac{(w_{1,j} - w_{2,j})^2}{\lambda_j} \end{aligned} \tag{G.5}$$

où les $w_{1,j}, w_{2,j}$ sont les coordonnées des i-vectors. Les inverses $\left\{\lambda_j^{-1}\right\}_j$ des valeurs propres de $\mathbf{W}$ déterminent seuls les poids associés à chaque dimension de l'espace des i-vectors.

G.2.2 Après la LDA

Nous avons démontré au paragraphe 4.3.4 que les réductions de dimensionnalité par LDA et NAP sont identiques après l'algorithme EFR : les sous-espaces de projection sont ceux des premiers vecteurs propres de $\mathbf{B}$. Appliquer une réduction de dimension (de p à un rang r) revient donc, les i-vectors étant exprimés dans cette base, à ne conserver que leurs r premières coordonnées. Notant $\mathbf{A}_r$ la matrice $r \times p$ de projection, le score de l'équation G.5 précédente devient le nouveau score entre les projetés $\mathbf{A}_r w_1$ et $\mathbf{A}_r w_2$:

$$score_{\text{Maha}}(\mathbf{A}_r w_1, \mathbf{A}_r w_2) = \sum_{j=1}^{r} \lambda_j^{-1} (w_{1,j} - w_{2,j})^2 \tag{G.6}$$

La rotation dans la base de $\mathbf{B}$ n'affectant pas les rapports de proximité entre les vecteurs, le score de Mahalanobis est un simple score de norme pondéré, sur la différence entre deux i-vectors. Il appartient à la gamme de mesures de proximité :

$$\left\{ \phi : (w_1, w_2) \mapsto \sum_{j=1}^{r} \beta_j (w_{1,j} - w_{2,j})^2 ,\ (\beta_j)_j \in \boldsymbol{R}^p \right\} \tag{G.7}$$

Le modèle de Mahalanobis ne s'intéresse pas directement aux vecteurs mais à la différence entre ceux-ci. Sous l'hypothèse d'additivité des effets locuteur et session et d'appartenance de deux vecteurs w_1, w_2 à un même locuteur, cette différence n'est imputable qu'à la variabilité intrinsèque (session). Le score de Mahalanobis propose une pondération de cet écart entre les vecteurs, suivant les poids $\left\{\lambda_j^{-1}\right\}_j$.

G.3 LDA et two-covariance

Un certain nombre de propriétés des i-vectors après normalisation EFR ont été présentées au 4.2. Le graphe spectral de la figure 4.9 montre qu'après un faible nombre d'itérations de l'algorithme, le critère de Rayleigh de détermination géométrique d'axes discriminants (LDA) est quasiment optimal dans la base d'*eigenfactors*. Il va de soi que cette normalisation sera choisie pour préparer les i-vectors à une modélisation suivant le paradigme LDA + modèle two-covariance :

- le modèle two-covariance s'appuie sur des hypothèses gaussiennes et nous avons montré que l'EFR contribuait à gaussianiser à la fois données d'apprentissage et d'évaluation,
- la LDA va dégager un sous-espace de principale variabilité locuteur (qui est aussi celui de la technique NAP et de **B**, cette dernière servant de pivot à l'estimation de la variabilité locuteur du modèle two-covariance).

G.3.1 Simplification du score two-covariance

Après normalisation EFR, la formule de calcul du scoring two-covariance peut être simplifiée. La nouvelle écriture obtenue n'a pas pour seul but de faciliter l'implémentation du calcul. Elle permet également d'observer la nature du calcul de proximité effectuée entre les observations à comparer, de dégager les matrices et sous-espaces les plus utiles à la discrimination et d'envisager de futurs aménagements et variantes.

La simplification s'appuie sur l'hypothèse que les données, qui tendent vers un modèle normal standard, suivent exactement cette loi. Cette approximation, qui allège nettement la formulation, se justifie théoriquement par la considération de l'écart à loi théorique comme imputable aux seules fluctuations d'échantillonnage.

i) D'une part, la moyenne globale μ des données d'apprentissage est nulle, permettant d'appliquer la formule de score two-covariance de l'équation B.27.

ii) D'autre part, dans la base de **B**, en notant $\mathbf{\Lambda}$ la matrice diagonale de ses valeurs propres décroissantes, on a :

$$\begin{cases} \mathbf{B} = \mathbf{\Lambda} \\ \mathbf{I} = \mathbf{B} + \mathbf{W} \text{ où } \mathbf{I} \text{ matrice identité} \\ \mathbf{W} = \mathbf{I} - \mathbf{\Lambda} \\ \mathbf{B}^{-1} = \mathbf{\Lambda}^{-1} \\ \mathbf{W}^{-1} = (\mathbf{I} - \mathbf{\Lambda})^{-1} \end{cases} \tag{G.8}$$

sous l'hypothèse $\mathcal{N}(0, \mathbf{I})$.

D'où, en passant les i-vectors dans la base de $\mathbf{B}$ (la rotation ne modifiant pas la distribution des scores), les parts de scores cible et non-cible deviennent :

$$\begin{aligned} score_{2\text{cov}}(w_1, w_2|\theta_{\text{tar}}) = (w_1 + w_2)^t (\mathbf{I} - \mathbf{\Lambda})^{-1} \left(\mathbf{\Lambda}^{-1} + 2(\mathbf{I} - \mathbf{\Lambda})^{-1}\right)^{-1} (\mathbf{I} - \mathbf{\Lambda})^{-1} (w_1 + w_2) \\ - w_1 (\mathbf{I} - \mathbf{\Lambda})^{-1} w_1 - w_2 (\mathbf{I} - \mathbf{\Lambda})^{-1} w_2 \end{aligned} \tag{G.9}$$

et

$$\begin{aligned} score_{2\text{cov}}(w_1, w_2|\theta_{\text{non}}) = w_1^t (\mathbf{I} - \mathbf{\Lambda})^{-1} \left(\mathbf{\Lambda}^{-1} + (\mathbf{I} - \mathbf{\Lambda})^{-1}\right)^{-1} (\mathbf{I} - \mathbf{\Lambda})^{-1} w_1^{-1} \\ + w_2^t (\mathbf{I} - \mathbf{\Lambda})^{-1} \left(\mathbf{\Lambda}^{-1} + (\mathbf{I} - \mathbf{\Lambda})^{-1}\right)^{-1} (\mathbf{I} - \mathbf{\Lambda})^{-1} w_2 \\ - w_1 (\mathbf{I} - \mathbf{\Lambda})^{-1} w_1 - w_2 (\mathbf{I} - \mathbf{\Lambda}) w_2 \end{aligned} \tag{G.10}$$

Toutes les matrices utilisées sont diagonales et peuvent se traiter dans les calculs comme des scalaires, d'où les simplifications :

$$\begin{aligned} score_{2\text{cov}}(w_1, w_2|\theta_{\text{tar}}) = (w_1 + w_2)^t \mathbf{\Lambda} \left(\mathbf{I} - \mathbf{\Lambda}^2\right)^{-1} (w_1 + w_2) \\ - w_1 (\mathbf{I} - \mathbf{\Lambda})^{-1} w_1 - w_2 (\mathbf{I} - \mathbf{\Lambda})^{-1} w_2 \end{aligned} \tag{G.11}$$

et

$$\begin{aligned} score_{2\text{cov}}(w_1, w_2|\theta_{\text{non}}) = w_1^t \mathbf{\Lambda} (\mathbf{I} - \mathbf{\Lambda})^{-1} w_1 + w_2^t \mathbf{\Lambda} (\mathbf{I} - \mathbf{\Lambda})^{-1} w_2 \\ - w_1 (\mathbf{I} - \mathbf{\Lambda})^{-1} w_1 - w_2 (\mathbf{I} - \mathbf{\Lambda})^{-1} w_2 \end{aligned} \tag{G.12}$$

Concernant le score non-cible, on peut aller plus loin :

$$\begin{aligned} score(w_1, w_2|\theta_{\text{non}}) &= \sum_{i=1}^{2} w_i^t \left(\mathbf{\Lambda} (\mathbf{I} - \mathbf{\Lambda})^{-1} - (\mathbf{I} - \mathbf{\Lambda})^{-1}\right) w_i \\ &= -w_1^t w_1 - w_2^t w_2 = -\|w_1\|^2 - \|w_2\|^2 \\ &= -2p \end{aligned} \tag{G.13}$$

		hommes		femmes	
normalisation	scoring	EER	DCF	EER	DCF
EFRnorm 2 iter.	LDA$_{\mathbf{B},\mathbf{W}}$-2Cov	1.27	0.31	1.89	0.35
EFRnorm 2 iter.	LDA$_{\mathbf{B},\mathbf{W}}$-2Cov simplifié	1.26	0.32	2.07	0.33

TABLE *G.1* – *Performances du scoring LDA-2Covariance et de sa version simplifiée sous l'hypothèse EFR.*

Nous retrouvons le fait, déjà signalé, qu'après normalisation EFR la mesure correspondant à l'hypothèse θ_{non} est constante, donc non-informative. Sur une modélisation où les observations sont équiprobables, la décision ne s'appuie donc que sur la mesure sous l'hypothèse θ_{tar}.

L'algorithme de calcul simplifié des scores se présente ainsi :

- calculer les vecteurs et valeurs propres de $\mathbf{B}$ sur le fichier d'apprentissage.
- exprimer les i-vectors à évaluer dans cette base,
- calculer le score :

$$
\begin{aligned}
score_{2\text{cov}}(w_1, w_2) &= score_{2\text{cov}}(w_1, w_2|\theta_{\text{tar}}) \\
&= (w_1 + w_2)^t \, \mathbf{\Lambda} \left(\mathbf{I} - \mathbf{\Lambda}^2\right)^{-1} (w_1 + w_2) && \text{(G.14)} \\
&- w_1 \left(\mathbf{I} - \mathbf{\Lambda}\right)^{-1} w_1 - w_2 \left(\mathbf{I} - \mathbf{\Lambda}\right)^{-1} w_2 && \text{(G.15)} \\
&= \sum_{j=1}^{p} \left\{ \frac{\lambda_j}{1 - \lambda_j^2} \left(w_{1,j} + w_{2,j}\right)^2 - \frac{1}{1 - \lambda_j} \left(w_{1,j}^2 + w_{2,j}^2\right) \right\} && \text{(G.16)}
\end{aligned}
$$

G.3.2 Performance du score simplifié

Le score simplifié est une approximation du score two-covariance sous les hypothèses du modèle EFR. Il présente l'avantage de permettre une meilleure lisibilité des mesures mises en oeuvre dans ce scoring. Si le modèle EFR décrit mieux les paramètres statistiques de moyenne et covariance que leur estimation empirique, il doit également permettre d'améliorer la qualité de la détection du locuteur. Pour mieux évaluer cette qualité, nous avons expérimenté les deux versions du score, initial puis simplifié sous hypothèses EFR, sur un certain nombre de jeux d'évaluation et comparé les taux d'erreur obtenus.

La table G.1 affiche les résultats de la comparaison des deux modèles et scorings LDA-two-covariance initial et simplifié. Le cadre expérimental est le même qu'à la section 4.5.

Une similitude apparaît dans les résultats de l'évaluation hommes. Par contre, pour les femmes, un écart de performance non négligeable est observable en terme d'EER (2.07% pour 1.89%).

La dégradation d'EER pour l'évaluation femmes indique un défaut de séparabilité plus important dans cette population, celui-la même qui conduit tous les systèmes à s'avérer moins discriminants pour ces dernières. Une part plus forte de la nature même de leur variabilité vocale échappe encore aux modèles.

Annexe H

Perspectives sur l'anisotropie

Nous présentons ici une voie que nous avons explorée, induite par les observations du chapitre 6. Celle-ci n'a pas abouti à une progression significative en terme de performance, mais peut entamer un inventaire de techniques d'adaptation locale des métaparamètres. Ici, nous étudions l'adaptation du score two-covariance à l'existence de relations mathématiques explicites entre ces métaparamètres , en l'occurrence suivant la position considérée dans l'espace.

H.1 Adaptation du score two-covariance

Nous rappelons que les scores du modèle two-covariance sous les hypothèses θ_{tar} :"même locuteur" et θ_{non} :"pas le même locuteur" sont :

$$\log P\left(w_1, w_2|\theta_{\text{tar}}\right) = \log \int \mathcal{N}\left(w_1|y, \mathbf{W}\right) \mathcal{N}\left(w_2|y, \mathbf{W}\right) \mathcal{N}\left(y|\mu, \mathbf{B}\right) dy \tag{H.1}$$

et

$$\log P\left(w_1, w_2|\theta_{\text{non}}\right) = \log \int \mathcal{N}\left(w_1|y, \mathbf{W}\right) \mathcal{N}\left(y|\mu, \mathbf{B}\right) dy \tag{H.2}$$

$$+ \log \int \mathcal{N}\left(w_2|y, \mathbf{W}\right) \mathcal{N}\left(y|\mu, \mathbf{B}\right) d \tag{H.3}$$

Nous avons vu (équation 6.11) que les vraisemblances de classe vis à vis des métaparamètres $\mathbf{B}$ et $\mathbf{W}$ suivent à peu près exactement une relation linéaire. Dans le score correspondant à θ_{tar}, sous l'hypothèse de généralisation de cette relation, cela signifie que le produit de densités $\mathcal{N}\left(w_1|y, \mathbf{W}\right) \mathcal{N}\left(w_2|y, \mathbf{W}\right)$ a pour moyenne géométrique :

$$E\left[\mathcal{N}\left(w_1|y, \mathbf{W}\right) \mathcal{N}\left(w_2|y, \mathbf{W}\right)\right] = \left(E_{w \in s}\left[\mathcal{N}\left(w|y, \mathbf{W}\right)\right]\right)^2 \approx e^{2\beta_0} P\left(y|\mu, \mathbf{B}\right)^{2\beta_1} \tag{H.4}$$

où β_0 et β_1 sont des réels empiriquement évalués sur le fichier d'apprentissage des métaparamètres. On a donc :

$$E\left[\frac{\mathcal{N}\left(w_1|y,\mathbf{W}\right)\mathcal{N}\left(w_2|y,\mathbf{W}\right)}{\mathcal{N}\left(y|\mu,\mathbf{B}\right)^{2\beta_1}}\right] \approx e^{2\beta_0} \tag{H.5}$$

Ce terme, qui tend à être constant, constitue donc une estimation de la probabilité conjointe des deux observations w_1 et w_2 supposées indépendantes sous l'hypothèse d'appartenance à la même classe s. L'ajout du dénominateur $\mathcal{N}\left(y|\mu,\mathbf{B}\right)^{2\beta_1}$ rend donc ces probabilités conjointes indépendantes de l'effet anisotropique du métaparamètre $\mathbf{W}$ suivant la localisation dans l'espace du centre y_2 présumé et combat ainsi partiellement l'anisotropie. Le score sous l'hypothèse θ_{tar} devient :

$$\log P\left(w_1,w_2|\theta_{\text{tar}}\right) = \log \int \mathcal{N}\left(w_1|y,\mathbf{W}\right)\mathcal{N}\left(w_2|y,\mathbf{W}\right)\left(\mathcal{N}\left(y|\mu,\mathbf{B}\right)\right)^{1-2\beta_1} dy \tag{H.6}$$

à une constante près, indépendante de w_1 et w_2.

De même, on peut normaliser la densité de l'observation w_1 ou w_2 dans le score θ_{non} en divisant cette densité par le facteur $\mathcal{N}\left(y|\mu,\mathbf{B}\right)^{\beta_1}$, pour redresser l'anisotropie des métaparamètres. Le score sous l'hypothèse θ_{non} devient :

$$\begin{aligned}\log P\left(w_1,w_2|\theta_{\text{non}}\right) = &\log \int \mathcal{N}\left(w_1|y,\mathbf{W}\right)\mathcal{N}\left(y|\mu,\mathbf{B}\right)^{1-\beta_1} dy \\ &+ \log \int \mathcal{N}\left(w_2|y,\mathbf{W}\right)\mathcal{N}\left(y|\mu,\mathbf{B}\right)^{1-\beta_1} dy\end{aligned} \tag{H.7}$$

Il a été vu à l'équation B.27 que ces scores peuvent se calculer explicitement. Dans le cas où la moyenne globale μ est nulle, ce qui est approximativement le cas après itérations d'EFR, l'expression explicite du score two-covariance de l'équation B.27 en tenant compte de ce régulateur d'anisotropie consiste simplement à redéfinir la matrice $\mathbf{B}$. En effet :

$$\log \mathcal{N}\left(y|\mu,\mathbf{B}\right)^{\alpha} = K - \frac{1}{2}\left(y-\mu\right)^t \left(\frac{1}{\alpha}\mathbf{B}\right)^{-1}\left(y-\mu\right) \tag{H.8}$$

où K est une constante., ce qui revient à effectuer les opérations suivantes dans l'équation B.27 :

– $\mathbf{B} \rightarrow \mathbf{B}_{\text{tar, reg}} = \dfrac{1}{1-2\beta_1}\mathbf{B}$ dans le score θ_{tar}

et

EER %				EFR LDA2Cov	SphN G-PLDA	EFR LDA2Cov (anisotropique)
LIA	2008	hommes	det6	4.93	4.80	**4.77**
LIA	2008	hommes	det7	1.53	1.58	**1.45**
BUT	2008	hommes	det6	3.09	**2.74**	3.10
BUT	2008	hommes	det7	1.06	0.90	**0.89**
LIA	2010	hommes	det5Ext	2.36	2.28	**2.25**
BUT	2010	hommes	det5Ext	1.27	**1.04**	1.06
BUT	2010	femmes	det5Ext	1.94	**1.73**	1.89
		moyenne	EER	2.31	2.15	2.20

TABLE *H.1* – *Performances en terme d'EER de systèmes basés sur les modèles PLDA ou LDA-two-covariance avec ou sans prise en compte de l'anisotropie.*

– $\mathbf{B} \rightarrow \mathbf{B}_{\text{non, reg}} = \dfrac{1}{1-\beta_1}\mathbf{B}$ dans le score θ_{non}

On obtient donc le nouveau score two-covariance, déduit du score initial après normalisation de l'équation B.27 et de cette régularisation des matrices $\mathbf{B}$:

$$\begin{aligned} score\,(w_1, w_2) = {} & (w_1 + w_2)^t \mathbf{W}^{-1} \left[(1-2\beta_1)\,\mathbf{B}^{-1} + 2\mathbf{W}^{-1}\right]^{-1} \mathbf{W}^{-1}\,(w_1 + w_2) \\ & - w_1^t \mathbf{W}^{-1} \left[(1-\beta_1)\,\mathbf{B}^{-1} + 2\mathbf{W}^{-1}\right]^{-1} \mathbf{W}^{-1} w_1 \qquad \text{(H.9)} \\ & - w_2^t \mathbf{W}^{-1} \left[(1-\beta_1)\,\mathbf{B}^{-1} + 2\mathbf{W}^{-1}\right]^{-1} \mathbf{W}^{-1} w_2 \qquad \text{(H.10)} \end{aligned}$$

H.2 Performances comparées

La table H.1 présente les performances comparées, en terme d'EER, des trois systèmes :

- normalisation EFR puis LDA et scoring two-covariance

- normalisation SphN puis PLDA gaussienne (initialisée non aléatoirement suivant le procédé décrit en 4.6.7)

- normalisation EFR puis LDA et scoring two-covariance avec prise en compte de l'anisotropie par la méthode décrite au paragraphe précédent,

ce pour les jeux de données LIA et BUT et diverses conditions NIST-SRE 2008 ou 2010, hommes ou femmes, déjà présentées précédemment.

La méthode basée sur le modèle two-covariance avec prise en compte de l'anisotropie parvient à surpasser la version two-covariance standard et à concurrencer la

PLDA gaussienne . Elle dépasse cette dernière dans 4 conditions sur 7, toutefois les EER moyens montrent une supériorité de la PLDA (2.15% pour 2.20%). La méthode probabiliste basée sur maximum de vraisemblance est concurrencée par la méthode déterministe, pourtant considérée comme moins précise. La mise en place de techniques probabilistes plus fines reste à réaliser.

Quant à l'anisotropie des métaparamètres, inévitable sur la surface non-linéaire où s'étendent les vecteurs, cette étude montre qu'elle pénalise les systèmes actuels et qu'existent des solutions pour l'atténuer. La méthode que nous avons proposée est purement empirique et postérieure à la construction du modèle. L'élaboration de modèles probabilistes de type PLDA qui prennent en compte la contrainte d'isotropie des métaparamètres (dans la fonction objective à maximiser, donc durant les itérations de l'algorithme EM) apparaît comme un sujet d'investigation important.

Liste des illustrations

Liste des tableaux

Bibliographie

Anguera, X. and Bonastre, J.-F. (2010). Novel binary key representation for biometric speaker recognition. In *International Conference on Speech Communication and Technology*.

Atal, B. (1974). Effectiveness of linear prediction of the speech wave for automatic speaker identification and verification. *Journal of Statistic Society of America (JASA)*, 55 :pp 1304–1312.

Auckenthaler, R., Carey, M., and Lloyd-Thomas, H. (2000). Score Normalization for Text-Independent Speaker Verification System. *Digital Signal Processing*, 1(10) :42–54.

Bartholomew, D. (1987). *Latent Variable Models and Factor Analysis*. London : Charles Griffith and Co. Ltd.

Belkin, M. and Niyogi, P. (2003). Laplacian eigenmaps for dimensionality reduction and data representation. *Neural co*, 15(6) :pp 1373–1396.

Bishop, C. M. (2006). *Pattern recognition and machine learning*, volume 4. Springer.

Boë, L.-J. and Bonastre, J.-F. (2012). L'identification du locuteur : 20 ans de témoignage dans les cours de justice. le cas du lipsadon « laboratoire indépendant de police scientifique ». In *Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 1 : JEP*.

Bogert, B., Healy, M., , and Tukey, J. (1963). *The quefrency analysis for time-series for echoes : cepstrum, pseudoautocovariance, cross-cepstrum and shape cracking*. New York : J. Wiley.

Bonastre, J.-F., Anguera, X., Sierra, G. H., and Bousquet, P.-M. (2011a). Speaker modeling using local binary decisions. In *International Conference on Speech Communication and Technology*, pages 485–488.

Bonastre, J.-F., Bousquet, P.-M., Matrouf, D., and Anguera, X. (2011b). Discriminant binary data representation for speaker recognition. In *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, pages 5284–5287.

Bonastre, J.-F., Morin, P., and Junqua, J.-C. (2003). Gaussian dynamic warping (GDW) method applied to text-dependent speaker detection and verification. In *European Conference on Speech Communication and Technology (Eurospeech)*, pages 2013–2016.

Bousquet, P.-M. and Bonastre, J.-F. (2012). Typicality extraction in a speaker binary keys model. In *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*.

Bousquet, P.-M., Larcher, A., Matrouf, D., Bonastre, J.-F., and Ma, B. (2011a). Application of new i-vector Conditioning Algorithm and Scoring Method to NIST Speaker Recognition Evaluation 2010. In *NIST SRE Analysis Workshop 2011*.

Bousquet, P.-M., Larcher, A., Matrouf, D., Bonastre, J.-F., and Plchot, O. (2012a). Variance-Spectra based Normalization for I-vector Standard and Probabilistic Linear Discriminant Analysis. In *Speaker and Language Recognition Workshop (IEEE Odyssey)*.

Bousquet, P.-M., Matrouf, D., and Bonastre, J.-F. (2011b). Intersession compensation and scoring methods in the i-vectors space for speaker recognition. In *International Conference on Speech Communication and Technology*, pages 485–488.

Bousquet, P.-M., Matrouf, D., and Bonastre, J.-F. (2012b). Bayesian extension to the two-covariance model for mixed utterances in speaker recognition. In *submitted to Interspeech 2012*.

Bousquet, P.-M., Matrouf, D., and Bonastre, J.-F. (2012c). LIA system description for NIST SRE 2012. In *NIST Speaker Recognition Evaluation Workshop*.

Breiman, L. (2001). Random forests. *Machine Learning*, Volume 45, Issue 1 , :pp 5–32.

Brummer, N. and de Villiers, E. (2010). The speaker partitioning problem. In *Speaker and Language Recognition Workshop (IEEE Odyssey)*.

Burget, L., Fapso, M., Hubeika, V., Glembek, O., Karafiat, M., Kockmann, M., Matejka, P., Schwarz, P., and Cernocky, J. H. (2009). BUT system for NIST 2008 speaker recognition evaluation. In *International Conference on Speech Communication and Technology*, pages 2335–2338. International Speech Communication Association.

Burget, L., Matejka, P., and Cernocky, J. (2006). Discriminative training techniques for acoustic language identification. In *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, volume 1.

Burget, L., Matejka, P., Schwarz, P., Glembek, O., and Cernocky, J. (2007). Analysis of feature extraction and channel compensation in GMM speaker recognition system. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(7) :1979–1986.

Burget, L., Plchot, O., Cumani, S., Glembek, O., Matejka, P., and Brummer, N. (2011). Discriminatively trained probabilistic linear discriminant analysis for speaker verification. In *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, pages 4832–4835.

Campbell, J., Shen, W., Campbell, W., Schwartz, R., Bonastre, J., and Matrouf, D. (2009). Forensic speaker recognition. *IEEE Signal Processing Magazine*, 26 n°2 :95–103.

Campbell, W., Campbell, J., Reynolds, D., Jones, D. A., and Leek, T. (2004). Phonetic speaker recognition with support vector machines. *Advances in Neural Information Processing Systems*, (16) :pp 1377–1384.

Campbell, W. M., Sturim, D., Borgstrom, B. J., Dunn, R., McCree, A., Quatieri, T. F., and Reynolds, D. A. (2012). Exploring the impact of advanced front-end processing on NIST speaker recognition microphone tasks. In *Speaker and Language Recognition Workshop (IEEE Odyssey)*.

Campbell, W. M., Sturim, D. E., Reynolds, D. A., and Solomonoff, A. (2006). SVM based speaker verification using a GMM supervector kernel and NAP variability compensation. In *Proc. ICASSP*, volume 1, pages 97–100.

Carey, M. and Parris, E. (1992). Speaker verification using connected words. *Proc. Institute of Acoustics*, 14(6) :96–100.

Celeux, G. and Diebolt, J. (1985). The sem algorithm : a probabilistic teacher algorithm derived from the em algorithm for the mixture problem. *Comp. Statis. Quaterly*, 2 :pp 73–82.

Celeux, G. and Diebolt, J. (1986). L'algorithme sem : un algorithme d'apprentissage probabiliste pour la reconnaissance de mélanges de densités. *Revue de Statistique Appliquée*, 34 (2) :pp 35–52,.

Collet, M. (2006). *Mesures de similarité robustes dans un espace de locuteurs d'ancrage. Application pour l'indexation de documents audio*. PhD thesis, Université de Rennes.

Cooper, G. and Herskovitz, E. (1992). A bayesian method for the induction of probabilistic networks for data. *Machine Learning*, 9 :pp 309–347.

Cumani, S., Brummer, N., Burget, L., Laface, P., Plchot, O., and Vasilakakis, V. (2013). Pairwise discriminative speaker verification in the x28 ;-vector space. *IEEE Transactions on Audio, Speech & Language Processing*, 21(6) :1217–1227.

Davis, S. B. and Mermelstein, P. (1980). Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences. *IEEE Transactions on Audio, Speech, and Language Processing*, 28(4) :357–366.

DeGroot, M. (1970). *Optimal Statistical Decisions*.

Dehak, N. (2009). *Discriminative and Generative Approaches for Long-and Short-term Speaker Characteristics Modeling : Application to Speaker Verification*. PhD thesis, École de technologie supérieure.

Dehak, N., Dehak, R., Glass, J., Reynolds, D., and Kenny, P. (2010). Cosine similarity scoring without score normalization techniques. In *Speaker and Language Recognition Workshop (IEEE Odyssey)*. Odyssey.

Dehak, N., Dehak, R., Kenny, P., Brummer, N., Ouellet, P., and Dumouchel, P. (2009). Support Vector Machines versus Fast Scoring in the Low-Dimensional Total Variability Space for Speaker Verification. In *International Conference on Speech Communication and Technology*, pages 1559–1562.

Dehak, N., Kenny, P., Dehak, R., Dumouchel, P., and Ouellet, P. (2011). Front-End Factor Analysis for Speaker Verification. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4) :788–798.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society*, 39(1) :1–38.

Doddington, G., Liggett, W., Martin, A., Przybocki, M., and Reynolds, D. A. (1998). Sheep, goats, lambs and wolves : A statistical analysis of speaker performance in the NIST 1998 speaker recognition evaluation. In *Proceedings International Conference on Spoken Language Processing, ICSLP*. ISCA.

Eide, E. and Gish, H. (1996). A parametric approach to vocal tract length normalization. In *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, Atlanta (USA).

Enqing, D., Guizhong, L., Yatong, Z., and Xiaodi, Z. (2002). Applying support vector machines to voice activity detection. In *International Conference on Signal Processing*, volume 2.

Ferrer, L., Bratt, H., Burget, L., Cernocky, H., Glembek, O., Graciarena, M., Lawson, A., Lei, Y., Matejka, P., Plchot, O., and Scheffer, N. (2011). Promoting robustness for speaker modeling in the community : the prism evaluation set. In *NIST SRE Analysis Workshop 2011*.

Fine, S., Navratil, J., and Gopinath, R. A. (2001). A Hybrid Gmm/Svm Approach To Speaker Identification. In *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, volume 1, pages 417–420, Salt Lake City (USA).

Fredouille, C. (2000). *Approche Statistique pour la Reconnaissance Automatique du Locuteur : Informations Dynamiques et Normalisation Bayesienne des Vraisemblances*. PhD thesis, Université d'Avignon, Avignon, FRANCE.

Furui, S. (1981). Cepstral analysis technique for automatic speaker verification. *IEEE Transactions on Acoustics, Speech, and Signal Processing [see also IEEE Transactions on Signal Processing]*, 29(2) :254–272.

Garcia-Romero, D. and Espy-Wilson, C. Y. (2011). Analysis of i-vector length normalization in speaker recognition systems. In *International Conference on Speech Communication and Technology*, pages 249–252.

Gauvain, J.-L. and Lee, C.-H. (1994). Maximum a posteriori estimation for multivariate gaussian mixture observations of Markov chains. In *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, volume 2, pages 291–298.

Glembek, O., Burget, L., Dehak, N., Brummer, N., and Kenny, P. (2009). Comparison of Scoring Methods used in Speaker Recognition with Joint Factor Analysis. In *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, Taipei (Taiwan).

Hamsici, O. C. and Martinez, A. M. (2007). Spherical-homoscedastic distributions : The equivalency of spherical and normal distributions in classification. *The Journal of Machine Learning Research*, 8 :1583–1623.

Hatch, A. O., Kajarekar, S., and Stolcke, A. (2006). Within-Class Covariance Normalization for SVM-based Speaker Recognition. In *International Conference on Speech Communication and Technology*, pages 1471–1474.

Heckerman, D. (1995). A tutorial on leraning with bayesian networks. *Microsoft research*, (Msr-tr-95-06).

Hermansky, H. (1990). Perceptual linear predictive (PLP) analysis of speech. *Journal of Acoustic Society of America*, 87 :1738–1752.

Hermansky, H., Morgan, N., Bayya, A., and Kohn, P. (1991). Compensation for the effect of the communication channel in auditory-like analysis of speech (RASTA-PLP). In *European Conference on Speech Communication and Technology (Eurospeech)*. ISCA.

Hermansky, H., Morgan, N., Bayya, A., and Kohn, P. (1992). RASTA-PLP speech analysis technique. In *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, volume 1.

Higgins, A. L., Bahler, L., and Porter, J. (1991). Speaker verification using randomized phrase prompting. *Digital Signal Processing*, 1 :89–106.

Hyvärinen, A., Karhunen, J., and Oja, E. (2001). *Independent Component Analysis*. John Wiley and Son.

Ikedo, J. (1998). Voice activity detection using neural network. *IEICE Transactions on Communications*, 81(12) :2509–2513.

Jiang, Y., Lee, K. A., Tang, Z., Ma, B., Larcher, A., and Li, H. (2012). Plda modeling in i-vector and supervector space for speaker verification. In *International Conference on Speech Communication and Technology*.

Jolliffe, I. (2002). Principal component analysis. *Encyclopedia of Statistics in Behavioral Science*.

Kahn, J. (2011). *Parole de locuteur : performance et confiance en identification biométrique vocale*. PhD thesis, University of Avignon.

Kahn, J., Rossato, S., and Bonastre, J.-F. (2010). Intra-speaker variability effects on speaker verification performance. In *Speaker and Language Recognition Workshop (IEEE Odyssey)*.

Kanagasundaram, A., Dean, D., Sridharan, S., and Vogt, R. (2012). Plda based speaker verification with weighted lda techniques. In *Speaker and Language Recognition Workshop (IEEE Odyssey)*.

Kato, N., Omachi, S., Aso, H., and Nemoto, Y. (1999). A handwritten character recognition system using directional element feature and asymmetric mahalanobis distance. *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE*, 21(3) :pp258–262.

Kenny, P. (2010). Bayesian speaker verification with heavy-tailed priors. In *Speaker and Language Recognition Workshop (IEEE Odyssey)*.

Kenny, P. (2012). A small footprint i-vector extractor. In *Speaker and Language Recognition Workshop (IEEE Odyssey)*.

Kenny, P., Boulianne, G., and Dumouchel, P. (2005). Eigenvoice modeling with sparse training data. *IEEE Transactions on Speech and Audio Processing*, 13(3) :345–354.

Kenny, P., Boulianne, G., Ouellet, P., and Dumouchel, P. (2007). Joint factor analysis versus eigenchannels in speaker recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(4) :1435.

Kenny, P., Mihoubi, M., and Dumouchel, P. (2003). New MAP estimators for speaker recognition. In *European Conference on Speech Communication and Technology (Eurospeech)*. ISCA.

Kuhn, R., Nguyen, P., Junqua, J.-C., Goldwasser, L., Niedzielski, N., Fincke, S., Field, K., and Contolini, M. (1998). Eigenvoices for speaker adaptation. In *Proceedings International Conference on Spoken Language Processing, ICSLP*, pages 1771–1774, Sydney (Australia).

Kumar, N. and Andreou, A. (1998). Heteroscedastic Discriminant Analysis and Reduced Rank HMMs for Improved Speech Recognition. In *Speech Communication*, volume 26, pages 283–297.

L. Ferrer and, N. S. and et E. Shriberg (2010). A comparison of approaches for modeling prosodic features in speaker recognition. In *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*.

Ladefoged, P. (2005). *Vowels and consonants an Introduction to the sounds of Languages*. Oxford : Wiley-Blackwell.

Larcher, A. (2009). *Modèles acoustiques à structure temporelle renforcée pour la vérification du locuteur embarquée.* PhD thesis, University of Avignon.

Larcher, A., Bousquet, P.-M., Lee, K. A., Matrouf, D., Li, H., and Bonastre, J.-F. (2012a). I-vectors in the context of phonetically-constrained short utterances for speaker verification. In *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*.

Larcher, A., Bousquet, P.-M., Matrouf, D., and Bonastre, J.-F. (2012b). Analyse en Composante Principale pour l'extraction des i-vecteurs en vérification du locuteur (Principal Component Analysis for i-vector extraction in speaker verification.) [in French]. In ATALA/AFCP, editor, *Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 1 : JEP*, pages 297–304.

Larcher, A., Lee, K. A., Ma, B., and Li, H. (2013). Phonetically-constrained plda modeling for text-dependent speaker verification with multiple short utterances. In *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*.

Larcher, A., Lévy, C., Matrouf, D., and Bonastre, J.-F. (2010). LIA NIST-SRE'10 systems. In *NIST Speaker Recognition Evaluation Workshop*.

Lebart, L., Piron, M., and Morineau, A. (2000). *Statistique exploratoire multidimensionnelle*. Dunod, Paris.

Lee, J. A. and Verleysen, M. (2009). Simbed : Similarity-based embedding.

Magrin-Chagnolleau, I. (1997). *Approches statistiques et filtrage vectoriel de trajectires spetrales pour l'identification du locuteur indépendante du texte*. PhD thesis, ENST.

Mahalanobis, P. C. (1936). On the generalized distance in statistics. *Proceedings National Institute of Science, India*, 2(1) :pp. 49–55.

Marchal, A. (2007). *La production de la parole*. Paris : Hermès.

Martin, A. and Przybocki, M. (2000). The NIST 1999 speaker recognition evaluation - An overview. *Digital Signal Processing*, 10(1-3) :1–18.

Martin, A. F., Doddington, G. R., Kamm, T., Ordowski, M., and Przybocki, M. A. (1997). The DET Curve in Assessment of Detection Task Performance. In *European Conference on Speech Communication and Technology (Eurospeech)*. ISCA.

Mason, J. S., Oglesby, J., and Xu, L.-Q. (1989). Codebooks to optimise speaker recognition. In *First European Conference on Speech Communication and Technology*. ISCA.

Matejka, P., Glembeck, O., Castaldo, F., Alam, M., Plchot, O., Kenny, P., Burget, L., and Cernocky, J. (2011). Full-covariance UBM and heavy-tailed PLDA in i-vector speaker verification. In *International Conference on Speech Communication and Technology*, pages 4828–4831.

Matrouf, D. and Bonastre, J.-F. (2009). Session effects on speaker modeling. *Encyclopedia of Biometrics*, pages 1164–1169.

Matrouf, D., Bonastre, J.-F., and Mezaache, S. (2008). Factor analysis multi-session training constraint in session compensation for speaker verification. In *International Conference on Speech Communication and Technology*.

Matrouf, D., Scheffer, N., Fauve, B., and Bonastre, J.-F. (2007). A straightforward and efficient implementation of the factor analysis model for speaker verification. In *International Conference on Speech Communication and Technology*.

McLachlan, G. J. (1999). Mahalanobis distance. *Resonance*, June :pp 20–26.

McLaren, M. and Leeuwen, D. A. V. (2011). Source-normalised and weighted LDA for robust speaker recognition using i-vectors. In *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, pages 5456–5459.

Merlin, T., Bonastre, J.-F., and Fredouille, C. (1999). Non directly acoustic process for costless speaker recognition and indexation. In *International Workshop on Intelligent Communication Technologies and Applications, with emphasis on Mobile Communications*, Neuchâtel (Switzerland).

Pelecanos, J. and Sridharan, S. (2001). Feature warping for robust speaker verification. In *Speaker and Language Recognition Workshop (IEEE Odyssey)*.

Plannerer, B. (2005). *An Introduction to Speech Recognition*. plannerer@ieee.org.

Prince, S. J. and Elder, J. H. (2007). Probabilistic linear discriminant analysis for inferences about identity. In *International Conference on Computer Vision*, pages 1–8. IEEE.

Prince, S. J., Warrell, J., Elder, J., and Felisberti, F. (2008). Tied factor analysis for face recognition across large pose differences. *IEEE transactions on Pattern Analysis and Machine intelligence*, 30(6) :970–984.

Rabiner, L. R. (1989). A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proceedings of the IEEE*, 77(2) :257–286.

Rao, R. (1948). The utilization of multiple measurements in problems of biological classification. *Journal of the Royal Statistical Society, Series B 10 (2)*, pages pp 159–203.

Reynolds, D. A. (1995). Speaker identification and verification using Gaussian mixture speaker models. *Speech Communication*, 17(1-2) :91–108.

Reynolds, D. A., Quatieri, T. F., and Dunn, R. B. (2000). Speaker Verification Using Adapted Gaussian Mixture Models. *Digital Signal Processing*, 10 :19–41.

Rosenberg, A. E., DeLong, J., Lee, C.-H., Juang, B.-H., and Soong, F. K. (1992). The use of cohort normalized scores for speaker verification. In *Proceedings International Conference on Spoken Language Processing, ICSLP*, pages 599–602.

Rouvier, M. and Meignier, S. (2012a). A global optimization framework for speaker diarization. In *Speaker and Language Recognition Workshop (IEEE Odyssey)*.

Rouvier, M. and Meignier, S. (2012b). Nouvelle approche pour le regroupement des locuteurs dans des émissions radiophoniques et télévisuelles. In *Proceedings of journees d'etudes sur le parole, JEP*.

Sarkar, A. K., Matrouf, D., Bousquet, P.-M., and Bonastre, J.-F. (2012). Study of the effect of i-vector modeling on short and mismatch utterance duration for speaker verification. In *International Conference on Speech Communication and Technology*.

Scheffer, N. and Bonastre, J.-F. (2006). UBM-GMM Driven discriminative approach for speaker verification. In *Speaker and Language Recognition Workshop (IEEE Odyssey)*.

Scheffer, N., Ferrer, L., Graciarena, M., Kajarekar, S., Shriberg, E., and Stolcke, A. (2011). The SRI NIST 2010 speaker recognition evaluation system. In *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, pages pp 5292–5295.

Shriberg, E. and Stolcke, A. (2008). The case for automatic higher-level features in forensic speaker recognition. In *interspeech*, pages pp 1509–1512.

Singer, E., Torres-Carrasquillo, P., Reynolds, D., McCree, A., Richardson, F., Dehak, N., and Sturim, D. (2012). The MITLL NIST LRE 2011 Language Recognition System. In *Speaker and Language Recognition Workshop (IEEE Odyssey)*.

Soong, F. K. P., Rosenberg, A., Rabiner, L., and Juang, B. (1985). A Vector Quantization Approach to Speaker Recognition. In *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, volume 10, pages 387–390, Tampa (USA).

Starpert, R. P. and Mason, J. S. (2001). A segmental mixture model for speaker recognition. In *European Conference on Speech Communication and Technology (Eurospeech)*, volume 4, pages 2509–2512, Aalborg (Denmark).

Stevens, S., Volkmann, J., and Newman, E. (1937). The mel scale equates the magnitude of perceived differences in pitch at different frequencies. *Journal of Acoustic Society of America*, (8(3)) :pp 185–190.

Tenenbaum, J. B., de Silva, V., and Langford, J. C. (2002). A global geometric framework for nonlinear dimensionality reduction. *SCIENCE VOL 295 4 JANUARY 2002*, 290 :pp 2319–2322.

Teunen, R., Shahshahani, B., and Heck, L. (2000). A model-based transformational approach to robust speaker recognition. In *Proceedings International Conference on Spoken Language Processing, ICSLP*. ISCA.

Tipping, M. E. and Bishop, C. M. (1999a). Mixtures of probabilistic principal component analyzers. *Neural computation*, 11(2) :443–482.

Tipping, M. E. and Bishop, C. M. (1999b). Probabilistic principal component analysis. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 61(3) :611–622.

Turk, M. A. and Pentland, A. P. (1991a). Eigenfaces for Face Detection/Recognition. *Journal of Cognitive Neuroscience*, 3 :71–86.

Turk, M. A. and Pentland, A. P. (1991b). Face recognition using eigenfaces. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 586–591, Hawaï (USA).

van Leeuwen, D. A. and Brummer, N. (2013). The distribution of calibrated likelihood-ratios in speaker recognition. In *International Conference on Speech Communication and Technology*.

Vapnik, V. N. (1979). *Estimation of Dependences Based on Empirical Data*. [in Russian Nauka, Moscow (English translation : Springer Verlag, New York, 1982).

Vapnik, V. N. (1995). *The nature of Statistical Learning Theory*. Statistics for Engineering and information Science. Springer-Verlag.

Vapnik, V. N. (1998). *Statistical learning theory*. John Wiley & Sons.

Wan, V. and Campbell, W. M. (2000). Support Vector Machines for Speaker Verification and Identification. In *IEEE Signal Processing Society Workshop Neural Networks for Signal Processing*, volume 2, pages 775–784, Sydney (Australia).

Weinberger, K. Q. and Saul, L. K. (2006). Unsupervised learning of image manifolds by semidefinite programming. *International Journal of Computer Vision*, 70(1) :pp 77–90.

Yaman, S., Pelecanos, J., and Omar, M. K. (2011). Boosting Speaker Recognition Performance with Compact Representations. In *International Conference on Speech Communication and Technology*, pages 381–384.