UNIVERSIDAD AUTÓNOMA DE MADRID

ESCUELA POLITÉCNICA SUPERIOR
DEPARTAMENTO DE TECNOLOGÍA ELECTRÓNICA
Y DE LAS COMUNICACIONES

SESSION VARIABILITY COMPENSATION IN AUTOMATIC SPEAKER AND LANGUAGE RECOGNITION

–TESIS DOCTORAL–

COMPENSACIÓN DE VARIABILIDAD DE SESIÓN EN RECONOCIMIENTO AUTOMÁTICO DE LOCUTOR E IDIOMA

Author: Javier González Domínguez
(Ingeniero en Informática,
Universidad Autónoma de Madrid)

A Thesis submitted for the degree of:

Doctor of Philosophy

Madrid, November 2011

Colophon

This book was typeset by the author using LaTeX2e. The main body of the text was set using a 11-points Computer Modern Roman font. All graphics and images were included formatted as Encapsuled Postscript ($^{\text{TM}}$ Adobe Systems Incorporated). The final postscript output was converted to Portable Document Format (PDF) and printed.

Department:	Tecnología Electrónica y de las Comunicaciones Escuela Politécnica Superior Universidad Autónoma de Madrid (UAM) SPAIN
PhD Thesis:	Session Variability Compensation in Automatic Speaker and Language Recognition Systems
Author:	**Javier González Domínguez** Ingeniero en Informática (Universidad Autónoma de Madrid)
Advisor:	**Joaquín González Rodríguez** Doctor Ingeniero de Telecomunicación (Universidad Politécnica de Madrid) Universidad Autónoma de Madrid, SPAIN
Year:	2011
Committee:	President:
	Secretary:
	Vocal 1:
	Vocal 2:
	Vocal 3:

The research described in this Thesis was carried out within the Biometric Recognition Group – ATVS at the Dept. of Tecnología Electrónica y de las Comunicaciones, Escuela Politécnica Superior, Universidad Autónoma de Madrid (from 2006 to 2011). The project was partially funded by a PhD scholarship from Ministerio de Educación y Ciencia.

The author was awarded with a PhD scholarship from Ministerio de Educacion y Ciencia between 2007 and 2011, which partially supported the research summarized in this Dissertation.

The author was awarded with the Microsoft Research Best Student Paper Award at international conference SIG-IL/Microsoft 2009, for one publication originated from this Dissertation: "Multilevel and Channel-compensated Language Recognition: ATVS-UAM systems at NIST LRE 2009", Javier Gonzalez-Dominguez, Ignacio Lopez-Moreno, Javier Franco-Pedroso, Daniel Ramos, Doroteo T. Toledano e Joaquin Gonzalez-Rodriguez, Proc. of SIG-IL/Microsoft 2009, Porto Salvo, Portugal, September 2009.

Derived from partial work of this dissertation, the author was granted with an intern scholarship from Google Inc. This internship was entirely developed at the speech recognition technologies department of Google Research, NY, U.S.A., under the supervision of Pedro Moreno and Eugene Weinstein during the period July-October 2010.

Derived from partial work of this dissertation, the author has been invited to give talks in the area in prestigious speech labs as SRI international, Menlo Park, U.S.A and ICSI, Berkeley, U.S.A.

Abstract

Robust and accurate automatic speaker and language recognition, through the voice signal, remains a challenge for the scientific community mainly due to an old and well-known 'enemy': the session variability, defined as the set of variations among recordings belonging to a same identity (either speaker or language respectively).

During the past decades the issue of compensating/removing undesired variability effects has been broadly accepted as one of the biggest challenges in the field, giving rise to a number of publications full of new manners of somehow avoiding or cleaning the distortions present in the speech signal. However, major advances in the field have not been achieved until the development of new schemes based on Factor Analysis (FA) modelling. This fact responds to the conjunction of several ideas, properly combined in FA, which can be roughly summed up in two key points. First, exploiting prior knowledge in order to model session variability rather than directly removing it; and second, considering session variability as a continuous source rather than a discrete one.

This Ph.D. Thesis is focused on the study, analysis and development of new forms to palliate in a proper way the effects of the session variability problem through recent compensation schemes based on classical FA. In this sense, an extent analysis of the use and mathematical background of FA-based techniques, from the eigen-channels approach to more sophisticated schemes such as Joint Factor Analysis has been conducted.

Further, a special focus has been placed on the use of FA techniques applied to challenging scenarios, as those where the available background data is far from target conditions or the amount of train/test speech is very limited. This is a common case in the increasingly relevant forensic speaker recognition area. Regarding the experimental framework, well-defined and challenging recent automatic speaker and language recognition evaluations (SRE'08 and LRE'09 respectively) have been employed to assess the proposed and studied methods.

A mis padres y a mi hermano.
A Verónica, por supuesto.

¿No sientes ruido?
Mayor desdicha sospecho.
¿Si me podré levantar?
La voz es de mi señor. ¡Señor!.

–Lope de Vega, *El Arenal de Sevilla*, 1603.

Acknowledgements

> "La villanía es perdonable, la ingratitud no."
>
> "Villainy is forgivable, ingratitude is not."
>
> –Benito Pérez Galdós.

Más por vergüenza que por disposición y porque desde esta primera línea ya me persigue el recuerdo de todos aquellos a quienes injustamente olvido, bien habría yo aliviado esta página con un sincero "A todos". *Dios libre al lector de largos prólogos y malos epítetos*, así como agradecimientos que se enmadejan en letanías y desencadenan en solemnes epitafios.

Agradezco por delante, y ya me lanzo, a mi tutor Prof. Joaquín González por su encomiable esfuerzo y ayuda en llevar esta tesis a buen puerto. Una puerta siempre abierta con entrada de luz a la vorágine de dudas y sombras en el trayecto. Asimismo, soy también deudor de la experiencia y tutela de los Profesores Javier Ortega y Doroteo Torre Toledano.

Agradezco, con el sombrero en las manos, a los que son y seguirán siendo mis maestros dondequiera me halle. Gracias a Daniel Ramos, Alberto Montero y Julián Fiérrez por un sinfín de consejos que desde el principio desbordaron las palabras, para arremangados, convertirse siempre en ejemplos.

Ya en la trinchera, donde sobran los motivos, guardo especial lugar a mis fieles compañeros de batallas, no todas ellas necesariamente científicas; al tercio español, donde ando enrolado tiempo ha junto a Javier Franco, Ignacio López y Javier Galbally, y que tuvo en suerte la capitanía de Daniel Ramos (ora maestro, ora Spinola). Dumas estaría orgulloso, voto a Dios. Salud!

En estricto orden alfabético, ellas primero, agradezco profundamente a todos los actores de este Teatro con mayúsculas que damos en llamar laboratorio y donde he tenido la suerte de realizar esta tesis. Gracias a Alicia Beisner, Almudena Gilpérez, María Puertas, Marta Gómez, Miriam Moreno, Virginia Ruiz, Alberto Harriero, Alejandro Abejón, Daniel Hernández, Danilo Spada, Fernando Alonso, Fernando García, Ismael Mateos, Javier Simón, Juan Bonillo, Lucas Pérez, Manuel Freire, Pedro Tomé, Rubén Vera, Sergio Lucas y Víctor González. De sobra sabéis ya todos *que amistades que son ciertas nadie las puede turbar*. Y sí, olvido aquí a una compañera, soy consciente.

De cruzar otros mares me vienen otras deudas en otras lenguas y otras latitudes, que no acierto a ver como saldar algún día, caso de que esté en mi mano. Durante la elaboración de esta tesis he tenido la inmensa suerte de viajar del edificio A al C, de moverme de Brisbane a New York. Doy gracias por ello, pues son lugares que uno lleva en el camino como suyos, pero que carecerían de recuerdo sin las gentes que hicieron de la hospitalidad su bandera y cuya amistad no entiende de distancias.

I must thank to my Australians friends Robert Vogt and Brendan Baker their hospitality, support and understanding, even when their broad Australian accent insisted on mismatching with my excellent Spanish. Thanks mates!

I am also particularly indebted with Prof. David van Leeuwen, tireless worker, brilliant scientific and excellent person, who kindly welcomed me in Utrecht, and with whom I have had the pleasure to share ideas, variability subspaces and some beers. Thanks a lot.

I also must thank Pedro Moreno and Eugene Weinstein their praiseworthy efforts to integrate me into Google as one of them and make me part of a project where the word 'team' does not end on the covers of books; where good work is done in short-pants or not, by scooter or not, through the engineers, from the whiteboards to the users.

And of course, I must thank Patrick Lucey for something more than just a thesis, as I never would have imagined that I had a brother in Australia. Thanks Noreen and Daniel, well done!

Volviendo ya a Madrid, y con el regusto amargo de saber que me queda mucha gente en el cajón, hago dispendio de felicitaciones a todos los amigos que han estado y están en el camino. Entre ellos, a toda la banda de ex-futbolistas que, por supuesto, tuvieron fino toque en otra época, cuando hacían balones de verdad, a mis primos que son hermanos y a aquellos que están cuando tienen que estar. Gracias a Chema, Peter, Pablo, Darío, Sergio, Iván, Breza y el resto del elenco de grandes. Bien saben ellos quienes son.

Finalmente, nada tendría validez, si no agradeciera esto a mis padres y a mi hermano, a los que debo sencillamente lo que soy. Con y sinrazones éstas que, porque se me emborronan y atropellan infancias y recuerdos, se las tengo que ahorrar al lector para compartirlas con ellos.

Y de Machados a Cernudas, a mi compañera incansable, de Australia a cualquier confín, a Verónica. 222 veces y las que hagan falta, *seguiremos derecho, derecho, derecho, otra vez hasta La Dorada.*

Javier González Domínguez
Madrid, November 2011

Mathematical Notation

A consistent mathematical notation has been tried throughout this Dissertation, sometimes at the expense of usual or original conventions in other fields or works. Following symbols denote corresponding definitions:

x	Scalar.
$\boldsymbol{x}$	Multidimensional column vector.
$\boldsymbol{X}$	Matrix.
$\boldsymbol{X}^T$	The transpose of matrix $\boldsymbol{X}$.
$\boldsymbol{X}^{-1}$	The inverse of matrix $\boldsymbol{X}$.
$diag(\boldsymbol{X})$	Diagonal of matrix $\boldsymbol{X}$.
$\mathrm{tr}(\boldsymbol{X})$	Trace of matrix $\boldsymbol{X}$.
$(\mathrm{x}_1, ...x_D)^T$	Column vector of D elements.
$\boldsymbol{I}_D$	$D \times D$ identity matrix (abbreviated to I if there is no ambiguity).
$\boldsymbol{x}_1, ..., \boldsymbol{x}_N$	N samples (multidimensional column vectors).
$\boldsymbol{o}_1, ..., \boldsymbol{o}_N$	N speech observations (feature vectors in columns form).
$\mathcal{X}$	Data or observed space.
$\mathcal{Z}$	Latent Space.
$\lambda = \{w_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^K$	GMM model of K mixtures, being the kth mixture defined by the mean vector $\boldsymbol{\mu}_k$, the covariance matrix $\boldsymbol{\Sigma}_k$ and weight w_k.
$\boldsymbol{\mu}$	Mean supervector formed as the concatenation of K mixtures.
$\boldsymbol{\mu_k}$	Either the mean vector belonging to Gaussian k or the part corresponding to the Gaussian k within the supervector $\boldsymbol{\mu}$.
$\boldsymbol{\mu}^a$	Mean supervector originated by utterance or model a (used just if there is ambiguity).
$\boldsymbol{\Theta}$	Set of parameters of a given problem.
$\boldsymbol{\Theta^{(t)}}$	Set of parameters of a given problem in time or step t.
$\mathrm{E}_x[f(x, y)]$	Expectation of function $f(x, y)$ with respect to variable x (the suffix is omitted if there is no ambiguity).
$\mathcal{L}(\Theta)$	The likelihood function of some density model given a certain sample defined by a set of parameters Θ.
$\mathcal{L}c_\Theta$	The complete-data likelihood function of some density model given a certain sample. defined by a set of parameters ϕ.
$\cdot$	Scalar product.
$\bigtriangledown f$	Gradient of function f.

Contents

Abstract viii

Acknowledgements xii

MathematicalNotation xiv

List of Figures xix

List of Tables xxi

1. Introduction 1

1.1. Automatic Speaker and Language Recognition: Definitions and Applications . . 3

1.1.1. Automatic Speaker Recognition 3

1.1.2. SV Applications 4

1.1.3. Automatic Spoken Language Recognition 5

1.1.4. SLR Applications 5

1.2. Motivation of the Thesis 6

1.3. The Thesis 7

1.4. Objectives 7

1.5. Outline 7

1.6. Research Contributions 8

2. Automatic Speaker and Language Recognition 11

2.1. Introduction 11

2.2. Identity Information in the Speech Signal 12

2.3. Systems Architecture 12

2.3.1. Feature extraction 13

2.3.2. Modelling stage 17

2.3.3. Scoring normalization 18

2.3.4. Fusion 19

2.3.5. Calibration 19

2.4. Acoustic Systems 20

2.4.1. GMM 20

2.4.2. SVM . . . 23
2.4.3. SVM GMM-supervector . . . 24
2.5. High-Level Systems . . . 25
2.5.1. Statistical modelling . . . 25
2.5.2. Phone SVM . . . 26
2.5.3. Prosodic systems . . . 26
2.5.4. Phonotactic systems . . . 26
2.6. A Need for a Session Variability Compensation Approach . . . 27
2.6.1. Taxonomy of compensation approaches . . . 29
2.6.2. A historic and discrete view . . . 29
2.7. Summary . . . 31

3. A Continuous Approach to Variability Modelling: The Joint Factor Analysis Model **33**
3.1. Introduction . . . 33
3.2. From Eigenfaces to Joint Factor Analysis Model . . . 34
3.2.1. Eigenfaces . . . 34
3.2.2. Eigenvoices . . . 36
3.2.3. Eigenchannels . . . 37
3.2.4. The Joint Factor Analysis model . . . 38
3.3. Factor Analysis: The Model . . . 41
3.3.1. Latent variables models . . . 41
3.3.2. Factor analysis . . . 44
3.4. Factor Analysis on Gaussian Mixture Models . . . 47
3.4.1. Sufficient statistics . . . 48
3.4.2. MAP revisited . . . 48
3.4.3. Latent factors and hyperparameters estimation . . . 49
3.5. Factor Analysis on Speaker/Language Recognition . . . 51
3.5.1. Joint versus disjoint estimation . . . 51
3.5.2. Initialization of variability subspaces . . . 51
3.6. Summary . . . 53

4. Factor Analysis applied to SV and SLR systems: PART I (algorithmics) **55**
4.1. Introduction . . . 55
4.2. FA: Where and How . . . 56
4.2.1. FA in the model domain . . . 56
4.2.2. FA in the feature domain . . . 61
4.2.3. FA in the statistics domain . . . 62
4.3. The Linear Scoring Approach . . . 63
4.4. Toward Efficient and Robust Text-Independent Speaker and Language Recognition Acoustic Systems . . . 65

4.4.1. An efficient JFA based GMM-UBM systems for SV 66
4.4.2. An efficient JFA based GMM-UBM systems for SLR 67
4.5. Summary 67

5. Experimental Framework 71
5.1. Introduction 71
5.2. Databases 72
5.2.1. Automatic speaker recognition databases 72
5.2.2. Automatic language recognition databases 73
5.3. Evaluation of Performance 74
5.3.1. Detection Error Trade-off curve 74
5.3.2. Detection cost 76
5.4. Protocol and Tasks Definition 77
5.4.1. Automatic speaker recognition task definition 77
5.4.2. Automatic spoken language recognition task definition 79
5.5. Summary 81

6. Factor Analysis applied to SV and SLR systems: Part II (experimental) 83
6.1. Introduction 83
6.2. Joint Factor Analysis applied in SV Systems 84
6.2.1. GMM-UBM with standard MAP adaptation [MAP] 85
6.2.2. MAP adaptation with session variability compensation [MAP-SVC] . . . 85
6.2.3. Eigenvoice adaptation with session variability compensation [EV-SVC] . . 88
6.2.4. JFA modelling 91
6.2.5. JFA modelling with D trained on data 92
6.3. Joint Factor Analysis applied in SLR Systems 94
6.3.1. ATVS SLR submitted to NIST LRE 2009 94
6.3.2. Configuration of spectral systems 94
6.3.3. Configuration of high-level systems 95
6.3.4. Fusion and calibration 96
6.3.5. Performance of JFA-based spectral systems 98
6.3.6. Performance of global system 98
6.4. Summary 99

7. Factor Analysis in challenging SV and SLR scenarios 101
7.1. Introduction 101
7.2. Facing the database mismatch problem via FA 102
7.2.1. Joining Matrices 102
7.2.2. Pooled Sufficient Statistics 103
7.2.3. Scaling Statistics 103
7.2.4. Results 104

7.3. Facing the short duration problem via FA 108
7.3.1. Broad phonetic classes defined 109
7.3.2. Effect of phonetic composition in the speaker variability subspace 110
7.4. Summary 111

8. Conclusions and Future Work **113**
8.1. Conclusions 113
8.2. Future Work 116

Appendices **118**

A. Factor Analysis **119**

B. Linear Scoring **125**

C. Extended Results **127**

List of Figures

1.1. Dependence among the different chapters in this Dissertation. 10

2.1. Identity levels in the speech signal (adapted from [Kinnunen and Li, 2009]). . . . 13
2.2. Modular representation of typical training and test phases of a SV or SLR system. 14
2.3. Short-term feature extraction process. 15
2.4. Bank of classical Mel-filters on the MFCC feature extraction. 17
2.5. A GMM of four 3-Dimensional Gaussians (a) and their contours (b). 21
2.6. A Maximum a Posteriori Adaptation process representation where a speaker model (right) is adapted from a UBM model (left). 22
2.7. Representation of the SVM underlying idea (left) and basic elements of the SVM approach (right). 24
2.8. GMM to means supervector representation. 25
2.9. Scheme of a Phonotactic Language Modelling for recognition. 28

3.1. Global scheme of a eigenfaces for classification approach. 36
3.2. Representation of the speaker supervector decomposed in the speaker and session variability components. 39
3.3. Illustration of a point mapping process from a 2-dimensional latent space to a 3-dimensional observed space. The prior of the latent variables is assumed to be normally distributed $N(O,I)$ (density contours in the left side of the figure). The non-linear mapping function generates a manifold in the observed space where the point is mapped and then modified by a normal distributed noise, also considered $N(O,I)$ (grey sphere). 43
3.4. Graphical model representation of Factor Analysis. 45

4.1. Linear Scoring Representation. The scoring is computed as an approximation of the likelihood function over the point defined by the UBM mean supervector. . . 66

5.1. Representation of False Reject (FR) and False Acceptance (FA) errors in a biometric recognition system. 76
5.2. Example of DET curve. System 1 and 2 are compared in terms of FR and FA errors. 77

6.1. Results on SRE'08 tel-tel conditions for a GMM-UBM system with/without session variability compensation applied to first order statistics, in both training and test recordings. a) pooled scores. b) separate male and female scores. 86

6.2. a) Evolution of the Equal Error Rate for a session variability compensated system in function of the eigenchannels considered. b) Eigenvalues associated to the eigenchannels estimated for a session variability subspace sorted in descending order. 88

6.3. Effect of EM iterations on the ML refinement of the session variability subspace. a) comparison of male and female results in funcion of the EM iterations. b) comparison of the effect of the EM iterations in the female part by using or not data very similar to test data. 89

6.4. a) Effect of varying the number of eigenvoices in the EV-SVC system. b) Effect of the size of scoring normalization cohort in both the MAP-SVC and the EV-SVC system. 90

6.5. Results on SRE'08 tel-tel conditions for all the systems considered namely, MAP, MAP-SVC, EV-MAP, JFA and JFA with D matrix trained on data. a) male condition, b) female condition 93

6.6. Fusion scheme for ATVS2 submitted system. 95

6.7. Effect of session variability compensation on SVM-SV and FA-GMM-LS systems. Results on ATVS-Dev09 using VOA models and U_{VOA}. 97

6.8. Pooled DETs per ATVS submitted systems on development (ATVS-Dev09) and evaluation (LRE'09) per all target test segment durations (3, 10 and 30 seconds). 98

6.9. Comparision of ATVS submitted systems on both, development (ATVS-Dev09) and evaluation (LRE'09) datasets for 30 seconds test duration segments. 100

6.10. ATVS primary system performance on both, development (ATVS-Dev09) and evaluation (LRE'09) datasets (3, 10 and 30s). 100

7.1. Performance of the several proposed methods for the diverse conditions considered, *1conv4w-1conv4w*, *1conv4w-1mic*, *1mic-1conv4w* and *1mic-1mic*. 106

7.2. DET curves of the several proposed methods for the diverse conditions considered, *1conv4w-1conv4w*, *1conv4w-1mic*, *1mic-1conv4w* and *1mic-1mic*. 107

7.3. Representation of the Equal Error Rate in the telephone part of SRE08 in function of the training and test recordings duration. 108

List of Tables

3.1. JFA model components description, (equation 3.5). 40

3.2. Classification of Latent Variable Models according to the nature of the observed and latent variables (continuous or discrete). 42

3.3. Mathematical model definition of Factor Analysis. 44

3.4. Summary of JFA model parameters estimation. 50

4.1. Original Joint Factor Analysis integration within classical GMM-UBM framework. 57

4.2. General main advantages/disadvantages of applying FA in the different domains of an acoustic SV or SLR system. 63

4.3. Execution times for acoustic JFA and SVM supervector with session variability compensation systems. Numbers in brackets means already compute through the other system. 68

4.4. A robust and efficient acoustic GMM system for speaker verification. 69

4.5. A robust and efficient acoustic GMM system for spoken language recognition. . . 70

5.1. Information about languages/dialect involved as target languages in LRE series. 75

5.2. Composition of development datasets. 78

5.3. Composition of the development dataset C (”real-world” scenario). 79

5.4. Alphabetical list of languages used as development for LRE’09 evaluation. In bold, LRE’09 target languages. 80

6.1. UBM data distribution and main properties used in the GMM-UBM baseline system configuration. 85

6.2. Results on SRE’08 tel-tel conditions for a GMM-UBM system with/without session variability compensation. 87

6.3. Results on SRE’08 female tel-tel condition for MAP, MAP-SVC and EV-SVC systems. The number of eigenvoices considered in the EV-SVC system ranges from 50 to 300. A number of 50 eigenchannels was considered regarding the session variability subspace. 92

6.4. Results on SRE'08 male tel-tel condition for MAP, MAP-SVC and EV-SVC systems. The number of eigenvoices considered in the EV-SVC system ranges from 50 to 300. A number of 50 eigenchannels was considered regarding the session variability subspace. 92

6.5. Results on SRE'08 tel-tel (female) condition for the four evaluated systems, MAP, MAP-SVC, EV-SVC and JFA. 93

6.6. Results on SRE'08 tel-tel (male) condition for the four evaluated systems, MAP, MAP-SVC, EV-SVC and JFA. 93

6.7. Distribution of data used for training Universal Background Models and Session Variability Subspaces. 96

6.8. ATVS submitted systems performance (meanCavg x 100) on development and evaluation datasets. 97

7.1. Performance under restricted MIC data conditions in U training. 104

7.2. Performance using the joint matrices subspaces estimation approach. 104

7.3. Performance under restricted microphone data conditions when statistics are pooled with *devTel*. 105

7.4. Performance using scaled statistics during ML estimation. Results using 3 mic utterances per speaker. 105

7.5. Broad phonetic decomposition. 109

7.6. Broad phonetic classes originated by joining other phonetic classes. 110

7.7. Results on 10s-10s SRE'08 conditions by estimating the speaker variability subspace constrained to different phonetic classes. 111

7.8. Results on 10s-10s SRE'08 condition by estimating the speaker variability subspace constrained to different phonetic classes. 111

7.9. Eigenvoices used to compose five different speaker variability subspaces. 112

7.10. Results on 10s-10s SRE'08 condition for five different phonetic-class composition of the speaker variability subspace. 112

C.1. System: MAP-SVC. Results on SRE'08 female tel-tel condition by using different ML iterations on training the session variability subspace U as well as different number of eigenchannels. 128

C.2. System: EV-SVC. Results on SRE'08 female tel-tel condition by using different ML iterations on training the speaker variability subspace V as well as different number of eigenvoices. U is fixed over all the experiments and was trained via PCA keeping 50 eigenchannels. 129

C.3. System: JFA. Results on SRE'08 female tel-tel condition by using different ML iterations on training the speaker variability subspace V as well as different number of eigenvoices. U is fixed over all the experiments and was trained via PCA keeping 50 eigenchannels. 130

C.4. System: JFA D trained on data. Results on SRE'08 female tel-tel condition by using different ML iterations on training the speaker variability subspace V as well as different number of eigenvoices. U is fixed over all the experiments and was trained via PCA keeping 50 eigenchannels. 131

C.5. System: MAP-SVC. Results on SRE'08 male tel-tel condition by using different ML iterations on training the session variability subspace U as well as different number of eigenchannels. 132

C.6. System: EV-SVC. Results on SRE'08 male tel-tel condition by using different ML iterations on training the speaker variability subspace V as well as different number of eigenvoices. U is fixed over all the experiments and was trained via PCA keeping 50 eigenchannels. 133

C.7. System: JFA. Results on SRE'08 male tel-tel condition by using different ML iterations on training the speaker variability subspace V as well as different number of eigenvoices. U is fixed over all the experiments and was trained via PCA keeping 50 eigenchannels. 134

C.8. System: JFA D trained on data. Results on SRE'08 male tel-tel condition by using different ML iterations on training the speaker variability subspace V as well as different number of eigenvoices. U is fixed over all the experiments and was trained via PCA keeping 50 eigenchannels. 135

C.9. LRE'09 ATVS submitted systems performance (meanCavg x 100) on development dataset. 136

C.10.LRE'09 ATVS submitted systems performance (meanCavg x 100) on evaluation dataset. 137

Chapter 1

Introduction

This Ph.D Thesis is focused on building robust and efficient automatic speaker and language recognition systems. In particular, the Thesis is intended to provide a better understanding of the session variability problem and how this can be mitigated via techniques based on classical Factor Analysis (FA) modelling.

Automatic speaker and language recognition technologies have historically gone and still go hand by hand due to sharing numerous similarities in their problem formulation [Bimbot *et al.*, 2004; Campbell *et al.*, 2006a; Castaldo *et al.*, 2007; Torres-Carrasquillo *et al.*, 2002]. From the input speech signals to the final decisions given about the identity (speaker or language), a wide set of similarities can be found among the different approaches used in the several phases, which conform speaker and language recognition systems. Modules as voice activity detection, feature extraction or even modelling and classification stages are often identical, based on the same strategies or slightly modified to the specificities of one task to the other.

In this context, it is not surprising that as the same manner that they share similar components, they suffer from similar problems. Among all of them, the session variability problem requires special attention. Session variability, understood as the set of differences among recordings belonging to a same identity (either speaker or language depending on the corresponding task), has long been identified as the main cause of performance degradation in both fields [Bimbot *et al.*, 2004; Kinnunen and Li, 2009; Reynolds, 1996, 2002]. Channel distortions or effects produced by using different devices (landline, GSM) can be considered among the most relevant of factors that include session variability within the speech signal, but reducing session variations to the distortions produced by the channels is a naive approximation. Actually, a myriad of factors cause recordings to be different irrespective of the contained identity; for instance, the environment acquisition conditions at different locations (home, office, street, park, restaurant) or the emotional status of the speaker (calm, stressed, angry, happy) produce also session variations. Even in *controlled* acquisition environments, as smart-rooms, where session variations are intended to be minimized, slight variations such as an opened/closed window or changes in the speaker and acquisition terminal distance may lead to significant differences in resulting speech signals and consequent to performance degradation. [Sturim *et al.*, 2007].

During the last three decades numerous techniques had been proposed to palliate the session variability problem, those being based either on blind solutions [Furui, 1981; Hermansky and Morgan, 1994; Pelecanos and Sridharan, 2001] or based on quantifications of variability types [Reynolds, 2003; Teunen *et al.*, 2000]. Blind strategies, although desirable from a resource optimization perspective as non additional data is required to train them, fail to fit to the specificities of particular session conditions as no any prior information is taken into account. On the other hand, discrete strategies, even thought allowing a better adjustment, they still are an approximation to the session variability problem, as a proper quantification of variability sources or types is extremely difficult.

Extrapolated to our day to day, far from *controllable* conditions, the problem is scaled to acquire a major dimension. In a world flooded by an overwhelming number of devices able to capture and delivering speech, the application scenarios, and therefore the possible session variability sources, of speaker and language recognition technologies tends to be unquantifiable. According to Gartner information technology research reports, mobile connections are forecast to reach 7.4 billion from the current 5.6 billion in 2011 [1], supported by the rising trend of mobile sales, driven in turn, by the emergence of the smart-phones and tablets market. Further, the increasing availability of broadband lines in the different countries also predicts an explosion of the voIP (voice over internet protocol) use. In this context, an ever-growing need to cope with the session variability problem in a disparate number of scenarios is critical.

Driven by these needs, the design paradigm of session variability strategies has been recently redefined, and built on two main pillars or principles. First, to treat the variability as a continuous source; and second to exploit as much as possible prior information about possible session conditions encountered in target (operational) data. In 2004, the work conducted by Patrick Kenny [Kenny and Dumouchel, 2004b] succeeded in joining these two principles under a modelling strategy based on the classical Latent Variable Model, FA [Bartholomew, 1987]. From this pioneer work in the field, which was highly influenced by several advances in other related fields, such as face or speech recognition, a huge number of works have followed in a relative short period of time [Campbell *et al.*, 2006c; Kenny *et al.*, 2005b; Kenny and Dumouchel, 2004b; Kenny *et al.*, 2008b; Vair *et al.*, 2006; Vogt and Sridharan, 2008]

This Ph.D Thesis addresses the use of FA based methods to palliate the session variability problem, with the main objective of clarifying the grounds of this modelling strategy and how it is incorporated to achieve more robust and efficient speaker and language recognition systems.

Regarding the organization, this Dissertation begins by reviewing the foundations of the state-of-the-art speaker and language automatic recognition technology as well as the most successful techniques, which have arisen to deal with session effects before Factor Analysis[2]. Both acoustic and high level based systems will be detailed following the standard global scheme used in both speaker and language recognition fields. Also, a taxonomy of the different techniques to

[1] http://www.gartner.com/it/page.jsp?id=1759714.

[2] Advanced readers in the area could consider to skip this introduction part to automatic speaker and language recognition systems.

face the session variability will be exposed considering diverse criteria. Then, we will go into the development of techniques based on Factor Analysis by carefully detailing the main motivations, ideas, related studies, as well as the underlying mathematical framework which sustain them. A detailed exposition of *where and how* those techniques are integrated within the speaker and language recognition systems in an efficient manner, besides some of the original contributions of this Ph.D. Thesis will be exposed in this part of the Dissertation.

The experimental part starts then evaluating the inclusion of FA in speaker verification systems to later support the benefits achieved in the field of language recognition. To this aim, the widely accepted Speaker and Language Recognition Evaluations (LRE, SRE) (databases and protocols) conducted by the American National Institute of Standards and Technology (NIST) have been adopted as the experimental set-up. The databases used for such evaluations constitute challenging corpora presenting many different variability factors.

The use of Factor Analysis in the field of forensic speaker recognition field is then treated. One of the challenges of this Ph.D. thesis has been to adapt the use of Factor Analysis techniques to challenging scenarios, as those found in forensic speaker recognition [Gonzalez-Rodriguez *et al.*, 2007b; Leeuwen and Brümmer, 2007; Ramos, 2007], where the available background data is far from target conditions or the amount of train/test speech is very limited.

Finally, future work and conclusions are exposed. The research work described in this Dissertation has led to novel contributions which are mainly focused on three areas, namely, i) improving automatic speaker and language discrimination of state-of-the-art systems, ii) studying and developing new efficient ways to include techniques based on Factor Analysis to deal with session variability, and iii) facing the session variability problem in forensic speaker recognition. Moreover, some literature reviews has been derived from this Dissertation.

1.1. Automatic Speaker and Language Recognition: Definitions and Applications

Even thought automatic and language speaker recognition systems are deeply studied in Chapter 2, it is convenient at this point to define basics concepts of both fields, as well as their application framework, in order to properly introduce the motivation of this Dissertation.

1.1.1. Automatic Speaker Recognition

Speaker recognition is defined as the task of recognizing persons from their voice and it has a history extending back to the 1960s [Atal, 1972, 1976; Bricker and Pruzansky, 1966]. Among other biometrics, the voice has two main desirable characteristics that have made it an attractive trait. First, voice acquisition does not generate an *intrusive perception*, as other traits such as iris or fingerprint, being on contrary, captured from the individual in a natural and familiar way. Second, there is not need of using specialized technology, as telephone network, either landline, GSM or voIP, provides an excellent channel to obtain and delivering speech.

As any other biometric system, speaker recognition system can operate in two different modes [Bimbot *et al.*, 2004; Gonzalez-Rodriguez *et al.*, 2007c; Kinnunen and Li, 2009]:

- **Identification.** Identification is the task of determining an unknown speaker's identity among a group of known identities. This mode can be, in turn, divided in two subsets:

  - *open-set*. In this case the systems has to decide is the unknown test identity is or not among the speaker stored identities.
  - *closed-set*. Unlike the above case, here, the system is forced to identify one of the stored speakers with the identity of the unknown test recording, as the test identity is expected to be in the database.

  Identification systems usually returns a ranked list of similarities (in decreased order of similarity) extracted from a 'one to many', 1:N, matching process, where the input speech signals features of the unknown test recording are faced versus all the models stored in the database. As expected, the *open-set* condition is, in general, more challenging that the *closed-set* one, as a global threshold for final decision has to be properly defined.

- **Verification**. Speaker verification is defined as deciding if a speaker is who claims to be. In this case, a 'one-to-one', 1:1 matching process where the testing recording is compared to the enrolled model associated with the claimed identity is carried out. As a verification process just an affirmative or negative answer is possible, being this decided in function of a global threshold defined in the system.

Other traditional classification divides speaker recognition systems in function of the constraints imposed to the allowed spoken text within the recordings. Those being

- *text-dependent*. In this mode, the speaker usually pronounces a text or pass-phrase text-prompted in the testing phase.

- *text-independent*. In this case, no any restrictions to the text within the recordings in both training and test phases is required.

Unless otherwise stated, throughout this Thesis, the terminology "speaker recognition or verification", short-handed by the acronym SV, will be indistinctly used to refer to the verification mode (also known as authentication).

1.1.2. SV Applications

Speaker verification technologies have a broad number of scenarios of application such as voice dialling, on-line banking, tele-commerce, database access service, voice mail, security control for confidential information etc. Those can broadly classified in the following three groups:

- **Speaker Recognition for Authentication**. As a biometric modality one of the main application of the speaker recognition technologies is authentication. Access control applications for banking or e-commerce are examples of those applications [A. and S., 2006; James *et al.*, 1997; Zhang, 2002].

- **Speaker Recognition for Surveillance.** The ever-growing penetration of multimedia web-portals such as Youtube or Facebook, and in general of applications where large multimedia repositories are stored, have led to an increasing demand of data-indexing applications. In this context, automatic speaker recognition systems are a powerful tool to properly classify multimedia content by speakers [Tsekeridou and Pitas, 1998; Viswanathan *et al.*, 2000-06-01].

- **Forensic Speaker Recognition.** The confluence of accuracy in the technology [Przybocki *et al.*, 2007] and a more comprehensive study about the role of automatic speaker recognition in forensic science [Gonzalez-Rodriguez *et al.*, 2007b] has led to a increasing interest for the use of automatic speaker verification in forensics.

1.1.3. Automatic Spoken Language Recognition

Language recognition refers to identify the spoken language within a speech signal and its origin has a history dating back some decades [Atkinson, 1968; Muthusamy *et al.*, 1993, 1994; Zissman, 1996; Zissman and Singer, 1994]. As above mentioned, language recognition share many similarities with speaker recognition, mostly due to both being based on the same biometric trait: the voice.

Similar to speaker recognition, language recognition systems operate as either language identification or language verification tasks. Throughout this Thesis, terms ”spoken language recognition or identification”, short-handed by the acronym SLR, will be used to refer to the identification task.

1.1.4. SLR Applications

Although language recognition has been latent for nearly 40 years [Atkinson, 1968], it has not been up to the last decade when systems have experienced a major research development [NIST, 2009]. Those advances have favoured the used of automatic language recognition technologies in several areas and different applications. Among them, the following are highlighted:

1. **Call-Centres**. One of the most intuitive domains of applications for automatic language recognition technologies is to automatically route an incoming call to a fluent operator or automated agent in the call language. This type of services gains importance in security or health fields, but also can be extended to commerce services or in general, any kind of phone service [Zissman and Berkling, 2001].

2. **Audio indexing**. As in the case of speaker verification, the ever-growing increase of applications based on large repositories of multimedia data (eg. Youtube, MySpace), demands efficient tools to index the data in function of several parameters, among them, the language [Makhoul *et al.*, 2000].

1.2. Motivation of the Thesis

Understanding automatic speaker and language recognition systems as valuable tools for different industry and scientific applications, which embrace critical fields as security or forensic apart from others useful applications above mentioned such as data-indexing; and after identifying session variability as the main cause of the performance degradation of this systems, the main motivation of this Thesis is clear: improving automatic speaker and language recognition systems by dealing with session variability. But, more precisely, three observations from the state-of-the-art have mainly motivated the work conducted in this Dissertation. Those being:

- Due to its great ability of dealing with the problem of session variability, a high proliferation of Factor Analysis based methods applied to SV and SLR systems has taken place in a short period of time. However, whereas the basic concepts and hypothesis of the Factor Analysis are widely extended, a small amount of work has been published to deep review the mathematical foundations of Factor Analysis modelling, as well as to clarify the necessary modifications to its integration into SV and SLR fields. This fact has often conducted to a certain obscurity about the implementation process and also to the use of FA tools in a *black-box* mode, without a deep understanding of them.

- Related to the above observation, despite some valuable efforts such as those conducted in specific workshop in the field as JHU 2008[1] and Bosaris 2010[2], little research has been published to put on the same context the different manners to incorporate Factor Analysis into speaker and language verification systems. Even although same protocols or databases are often used by the scientific community, some other systems differences in the configuration parameters among published works (different number of Gaussians, type of features) hinder a fair comparison between the different forms of Factor Analysis.

- The increasing interest in forensic speaker recognition and the need of finding appropriate solutions to the often very adverse session variability conditions associated to this field. Due to the confluence of more robust and accurate systems as well as a better understanding in the field [Gonzalez-Rodriguez *et al.*, 2007b; Ramos, 2007], the interest to integrate automatic speaker recognition in the forensic field, to adequately supplement the labour carried out by the expert (eg. phoneticians) has rapidly expanded in recent years. In that sense, a little amount of research [Gonzalez-Dominguez *et al.*, 2010a; Ramos *et al.*, 2010,

[1] http://www.clsp.jhu.edu/workshops/ws08/groups/rsrovc/

[2] http://speech.fit.vutbr.cz/en/workshops/bosaris-2010

2008] has been conducted to explore, analyse and deal with the multiple problems and specificities encountered in the area.

1.3. The Thesis

The Thesis developed in this Dissertation can be stated as follows:

> *Exploiting prior knowledge about speech signal variability, conceived this as a continuous source, to properly include and adapt it to the particular characteristics of the target scenarios is essential to build robust and reliable automatic speaker and language recognition technology.*

1.4. Objectives

This Dissertation pursuits the following two prime objectives to a major benefit of the automatic SV and SLR systems:

- Provide insight about Factor Analysis as a powerful and efficient tool to deal with the session variability problem in automatic speaker and language recognition.
- Explore and propose different ways to incorporate FA into speaker and language recognition systems, able to obtain significant gains even in very adverse scenarios conditions.

1.5. Outline

The Dissertation is structured according to a *traditional complex* type [Paltridge, 2002] with background theory, literature review, theoretical and practical methods and three experimental studies in which the methods are applied. Essentially, chapters are structured as follows:

- Chapter 1 has introduced the basics of automatic speaker and language recognition topics, a description of the session variability problem, main cause of system performance degradation, and the motivation of this Dissertation. Research contributions originated from this Thesis will also be exposed at the end of this introductory chapter.
- Chapter 2 reviews state-of-the-art speaker and language recognition systems, placing special interest in most successful approaches adopted by the scientific community. Previous techniques to the appearance of Factor Analysis to palliate session variability effects are also presented in the final part of this chapter.
- Chapter 3 presents a deep analysis of Factor Analysis mathematical foundations besides the studies performed in related fields which motivated its application in speaker and language recognition tasks.

- Chapter 4 details proposed and existent methods to efficiently incorporate Factor Analysis into both speaker and language recognition systems, presenting different algorithms to get robust but also efficient acoustic systems.

- Chapter 5 describes the speech databases and protocols used to evaluate and provide empirical support to the different proposed methods and strategies exposed along this Dissertation.

- Chapter 6 opens the experimental part of this Thesis with a wide set of experiments to support Factor Analysis as an efficient and powerful tool to deal with the session variability problem. Experiments on the challenging NIST speaker and language evaluations 2008 and 2009 respectively are conducted and deeply analysed with that aim.

- Chapter 7 addresses main problems that hinder the deployment of SV and SLR systems in "real-world" applications as forensic speaker recognition. Specifically, the *database mismatch* and the *short durations* problems are analysed . Several novel contributions to deal with those problems are then presented and evaluated.

- Chapter 8 concludes the Dissertation summarizing the main results obtained and outlining future research lines.

  The dependence among the chapters is depicted in Figure 1.1.

Some methods developed in this PhD Thesis are strongly based on classical approaches coming from pattern recognition literature. The reader is referred to standard texts for a background on the topic [Bishop, 2007; Duda *et al.*, 2001; Fukunaga, 1990; Theodoridis and Koutroumbas, 2003]. More specific readings as automatic speaker [Bimbot *et al.*, 2004; Kinnunen and Li, 2009; Reynolds, 2002] and language recognition tutorials are also advised to get a broader vision of the field, despite this is intended in Chapter 2. It would be also useful to consult some algebra notes [Lay, 1997; Strang, 2003] and specialized bibliography about Factor Analysis [Bartholomew, 1987; Bartholomew *et al.*, 2011; Loehlin, 2004; Rubin and Thayer, 1982] for a deep understanding of concepts addressed in Chapter 3.

1.6. Research Contributions

The research contributions of this Ph.D. Thesis are the following (some publications are repeated in different items of the list):

Literature reviews.

1. Feature extraction for automatic speaker verification. [Ramos *et al.*, 2009]

2. Analysis of the speech signal for automatic speaker verification. [Toledano *et al.*, 2009]

3. Speaker verification systems. [Gonzalez-Dominguez *et al.*, 2010b; Gonzalez-Rodriguez *et al.*, 2007a; Montero-Asenjo *et al.*, 2006]

4. Spoken language recognition systems. [Gonzalez-Dominguez *et al.*, 2010d, 2009; Montero-Asenjo *et al.*, 2006]

Novel methods.

1. Novel methods in robust speaker verification [Gonzalez-Dominguez *et al.*, 2010c; Montero-Asenjo *et al.*, 2006; Perez-Gomez *et al.*, 2010; Ramos *et al.*, 2008].
2. Novel methods in robust spoken language recognition. [Gonzalez-Dominguez *et al.*, 2010d; Toledano *et al.*, 2007].
3. Novel methods for the use of high level features in language recognition [Montero-Asenjo *et al.*, 2006; Toledano *et al.*, 2007].
4. Novel methods for the use of automatic speaker recognition for forensic identification [Gonzalez-Dominguez *et al.*, 2010a; Ramos *et al.*, 2010, 2008].

Improvements in speaker recognition discrimination.

1. Contributions to the improvement of ATVS-UAM automatic speaker recognition systems [Gonzalez-Dominguez *et al.*, 2010a,b; Ramos *et al.*, 2010, 2008].
2. Contributions to the improvement of ATVS-UAM automatic speaker recognition systems in data sparse scenarios [Gonzalez-Dominguez *et al.*, 2010a; Ramos *et al.*, 2010, 2008].

Improvements in spoken language recognition discrimination.

1. Contributions to the improvement of ATVS-UAM automatic language recognition system. [Gonzalez-Dominguez *et al.*, 2010d, 2009]
2. Contributions to the improvement of ATVS-UAM automatic language recognition system based on high levels features. [Montero-Asenjo *et al.*, 2006; Toledano *et al.*, 2007]

Advances in forensic speaker recognition.

1. Studies on real forensic databases. [Ramos *et al.*, 2008]
2. Novel methods to apply Factor Analysis in forensic speaker recognition scenarios. [Gonzalez-Dominguez *et al.*, 2010a]

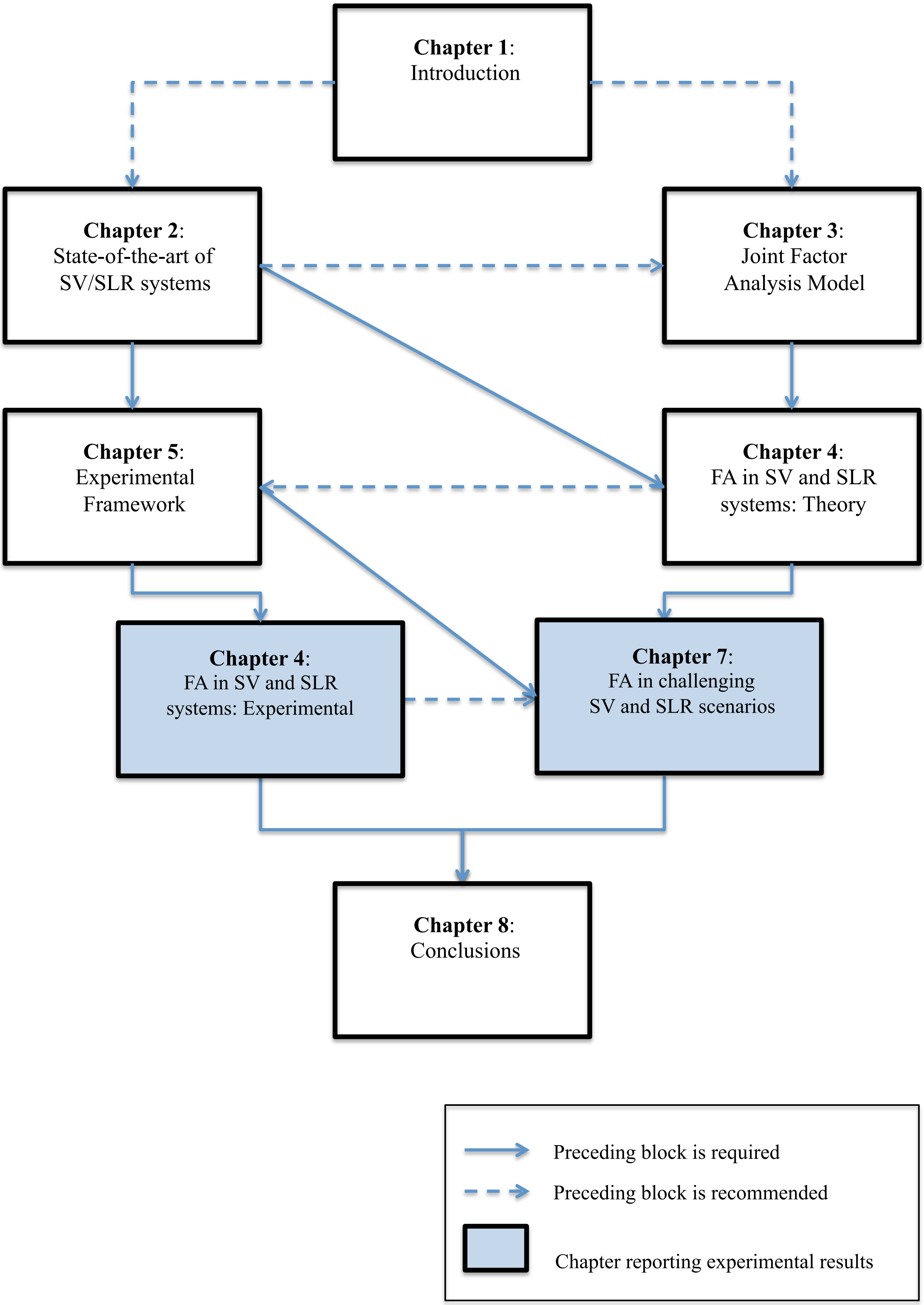

Figure 1.1: *Dependence among the different chapters in this Dissertation.*

Chapter 2

Automatic Speaker and Language Recognition

This chapter provides a holistic overview of automatic SV and SLR systems from the process of analysing/extracting the information within the speech signal to taking decisions concerning identity (speaker or language).

2.1. Introduction

A SV or SLR system can be seen as a process divided into two clear and distinct phases namely, the training phase and the test phase. Each of them are, in turn, composed by a sequence of independent modules which mainly includes the following three main modules i) feature extraction, ii) modelling and iii) scoring (computing similarity)/decision.

This chapter analyses this modular based architecture of SV and SLR systems, from the feature extraction process to the final decisions taken about identity (speaker or language) [1] through the detailed description of the main modules. The most successful approaches in each stage of the global systems with emphasis to those which nowadays conform the state of the art in the field, are highlighted.

The remainder of this chapter is organized as follows. First, an introductory analysis of the levels of information in the speech signal and the global architecture of SV and SLR systems is presented. Then, most successful acoustic and high level systems are detailed. In the final part of this chapter the focus is placed on the set of techniques previous to Factor Analysis (FA) conceived to palliate session variability effects.

[1] For the sake of clarity and due to the high degree of similarity between SV and SLR systems, this chapter has been written in terms of SV systems. Nonetheless, in those parts where differences or specificities between both systems exist, they will be explicitly specified

2.2. Identity Information in the Speech Signal

A speech signal is the result of a complex process that involves a large number of factors, which to a greater or lesser extent *print* a *trace* into it [Deller *et al.*, 1999; Huang *et al.*, 2001; Rabiner and Schafer, 1978; Ramos *et al.*, 2009] and are susceptible to be retrieved in order to formulate hypothesis about identity. Apart from the numerous *physical* factors implicated, other factors such as the *behavioural* factors (i.e socio-economic status, place of birth, etc.), inherent to the speaker, or the *environmental* factors (i.e place, acquisition channel, noise sources, etc.) add specific information into the speech signal. The aim of SV and SLR systems is to take advantage of the different sources of information available in the speech signal, combining them in the best possible way [Doddington, 2001; Reynolds *et al.*, 2003].

In the field of SV and SLR all this information is broadly classified into the so-called high-level (linguistic) and low-level (spectral) characteristics as follows:

- **Spectral level.** The information about the identity is extracted from the spectrum of the speech signal, analysed in short-time windows. The spectrum of the speech signal is directly related to the dynamic configuration of the vocal tract, which presents speaker-dependent specificities.

- **Higher levels.** Several sub-levels can be found here. For instance, at the phonotactic level, the information about the identity of the speaker is embedded in the particular use of the phones and syllables and their realizations. At the prosodic level, parameters like instantaneous energy, intonation, speech rate and unit durations are analysed, which are known to be speaker-dependent. At the idiolectal level, the information about speaker identity relies in the particular use of the words and language in general, which not only depends on the speaker, but in many other sociolinguistic conditions.

Figure 2.1 illustrates the different identity information levels found on the speech signal besides their main advantages/shortcomings.

2.3. Systems Architecture

As it has been mentioned before, a SV or SLR system can be seen as a two-phase (training and test) sequential, modular system which is primarily formed by three modules; the feature extraction, the modelling and the scoring/decision module, as depicted in Figure 2.2.

The feature extraction module is concerned with the extraction from the speech signal of adequate measurements which emphasize speaker (language) specificities while diminish statistical redundancies. Those measurements, better known as *features*, are somehow modelled in the training phase to produce a mathematical *model* which represents the given speaker or language. In the test phase, features extracted from the unknown recording are then compared with the set of available models in order to reach a similarity measure. Those measures are then used to produce a final decision about identity.

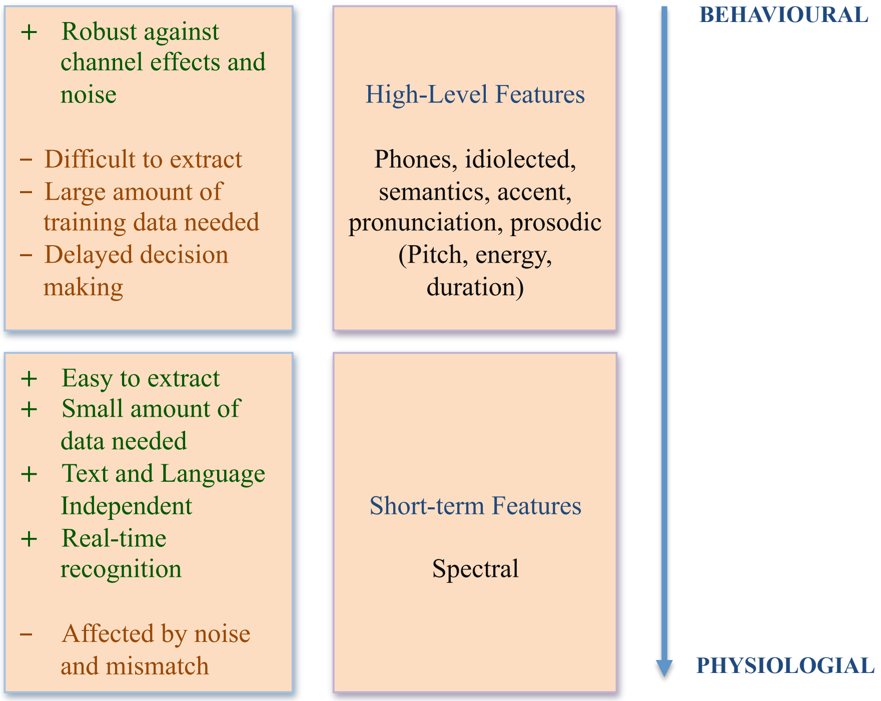
+ Robust against channel effects and noise
– Difficult to extract
– Large amount of training data needed
– Delayed decision making
High-Level Features
Phones, idiolected, semantics, accent, pronunciation, prosodic (Pitch, energy, duration)
+ Easy to extract
+ Small amount of data needed
+ Text and Language Independent
+ Real-time recognition
– Affected by noise and mismatch
Short-term Features
Spectral
BEHAVIOURAL
PHYSIOLOGIAL

Figure 2.1: *Identity levels in the speech signal (adapted from [Kinnunen and Li, 2009]).*

2.3.1. Feature extraction

Speaker (or language) features are measurements extracted from the speech signal with the objective of representing the specific information identity (either speaker or language) contained in it. Features are chosen to meet two fundamental criteria i) emphasize speaker (or language) specific properties and ii) suppress as much as possible statistical redundancy.

Ideally, they should have the following desirable properties [Kinnunen and Li, 2009; Ramos *et al.*, 2009]

a) maximize between-speaker/language variability and minimize within-speaker/language variability.

b) be robust against noise and distortion.

c) occur frequently and naturally in speech.

d) be easy to measure.

e) be hard to impersonate.

f) be robust respect intra-speaker/language variations.

Usually, different measures or observation of a same set of features are taken at different moments of the speech recording, giving rise to several feature vectors from a same recording;

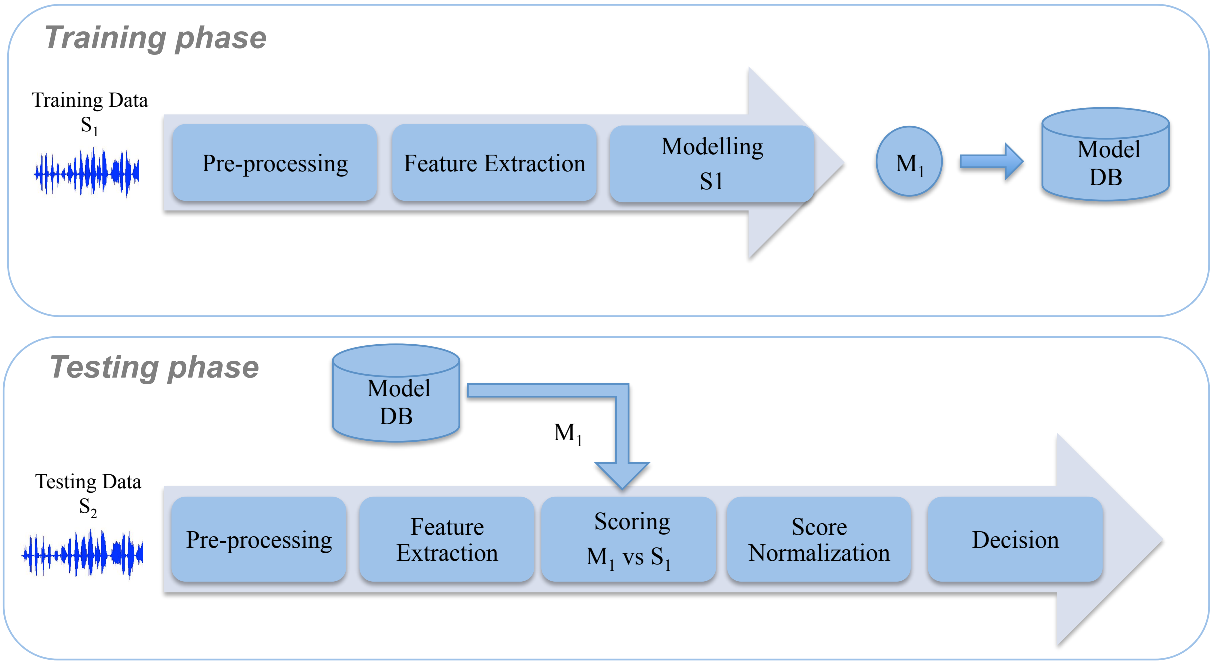

Figure 2.2: *Modular representation of typical training and test phases of a SV or SLR system.*

through this Dissertation the set of N feature vectors extracted from a given recording, also called the *observations* vectors, will be denoted as $\boldsymbol{O} = \boldsymbol{o}_1, ...\boldsymbol{o}_N$, being o_t a D-dimensional vector measured at time t.

2.3.1.1. Short-term spectral features

The analysis at spectral level of the speech signal is based on classic Fourier analysis. However, an exact definition of Fourier transform cannot be directly applied because speech signal cannot be considered stationary due to constant changes in the articulatory system within each speech utterance.

To solve these problems, speech signal is split into a sequence of short segments in such a way that each one is short enough to be considered pseudo-stationary. The length of each segment, also called window or *frame*, ranges between 10 and 40 milliseconds (in such a short-time period our articulatory system is not able to significantly change). Finally, a feature vector will be extracted from the short-time spectrum in each window. The whole process, known as *short-term analysis*, is depicted in Figure 2.3.

Signal representation or coding from short-term spectrum into a feature vector is one of the most important steps in a automatic speaker or language recognition system and it continues being subject of research. Many different techniques have been proposed in the literature and generally they are based on speech production models or speech perception models. Most widely-used techniques in the state of the art are described below.

- **Linear Predictive Coding (LPC)** method, introduced in [Makhoul and Wolf, 1973], is

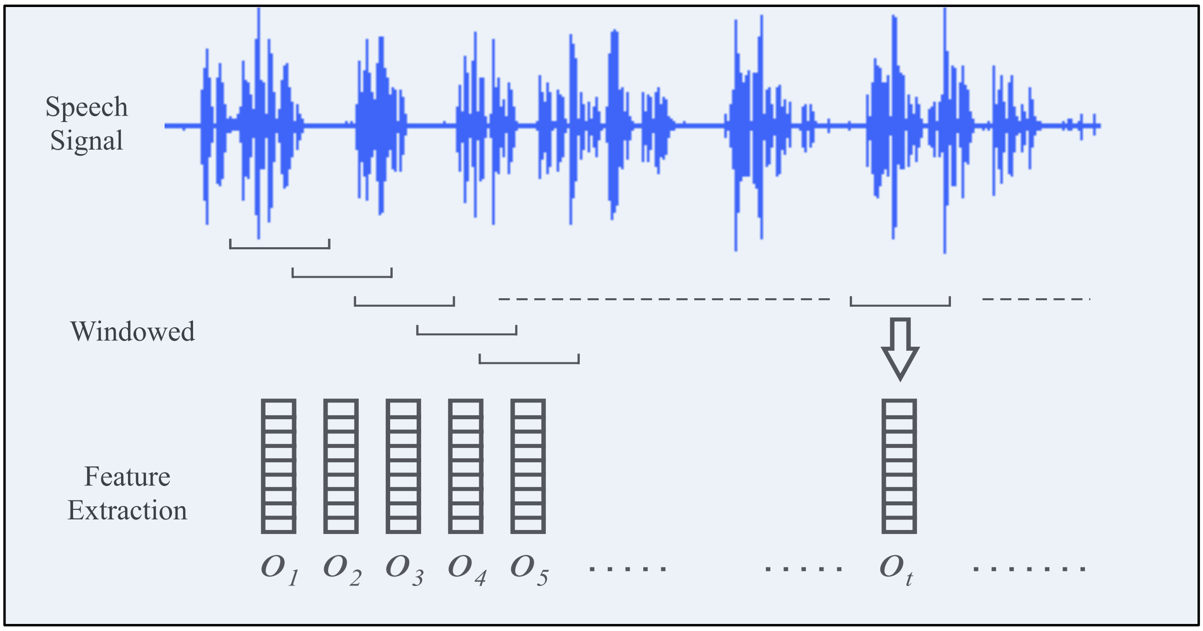

Figure 2.3: *Short-term feature extraction process.*

based on the assumption that a speech sample can be approximated by a linearly weighted summation of a determined number of preceding samples. In time domain, this can be represented as

$$s^*[n] = \sum_{k=0}^{p} a[k]\, s[n-k] \tag{2.1}$$

Here, $s^*[n]$ is the approximation, or *prediction*, of the speech signal, and $a[k]$ are the LPC coefficients calculated to minimize the total square error

$$E = \sum_{n} e[n]^2 \tag{2.2}$$

where $e[n]$ is the error between the real signal value $s[n]$ and predicted value $s^*[n]$, defined as

$$e[n] = s[n] - s^*[n] = s[n] - \sum_{k=1}^{p} a[k]\, s[n-k] \tag{2.3}$$

In the domain of the z-transform, $a[k]$ parameters define an all-pole filter $H(z)$, as defined in [Huang *et al.*, 2001; Makhoul and Wolf, 1973].

$$H(z) = \frac{1}{1 - \sum_{k=1}^{p} a[k]\, z^{-k}} \tag{2.4}$$

LPC has proved to be a valid way to compress the spectral envelope in an all-pole model with just 10 to 16 coefficients [Deller *et al.*, 1999; Huang *et al.*, 2001]. However, LPC coefficients are strongly correlated among them, which is an undesirable characteristic. Therefore, cepstrum transform [Deller *et al.*, 1999; Furui, 1981] has been proposed in order to obtain pseudo-orthogonal *cepstral* coefficients, yielding Linear Prediction Cepstral Coefficients (LPCC).

- **Mel-Frequency Cepstral Coefficients (MFCC)** proposed in [Bridle and Brown, 1974] are the most extensively used parameters at the spectral level in automatic speaker recognition systems. The MFCC method first uses a mel-scale filterbank in order to obtain some coefficients from the power spectrum of the speech window. The main aim of mel filtering is to mimic the human hearing behaviour by emphasizing lower frequencies and penalizing higher frequencies. Thus, a mel filterbank analyses the power spectrum using a logarithmic scale. First, a transformation is applied according to the following formula:

  $$f_m = 2595 * log\left(1 + f/700\right) \tag{2.5}$$

  where f is the linear frequency. Second, a filterbank is applied to the amplitude of the mel-scaled spectrum f_m in order to obtain a vector of outputs from each filter.

  Figure 2.4 shows a typical mel filterbank in the frequency domain. The centres $f[m]$ of the filters $H_m\left[k\right]$ are uniformly spaced in the mel scale. Using a DFT of the input signal with N points each filter $H_m\left[k\right]$ is given by

  $$H_m\left[k\right] = \begin{cases} 0 & k < f\left[m-1\right] \\ \frac{(k-f[m-1])}{(f[m]-f[m-1])} & f\left[m-1\right] \leqslant k \leqslant f\left[m\right] \\ \frac{(f[m+1]-k)}{(f[m+1]-f[m])} & f\left[m\right] \leqslant k \leqslant f\left[m\right] \\ 0 & k > f\left[m+1\right] \end{cases}$$

  where $0 < k < N$.

  Once filtering is carried out, cepstrum transform is applied to the filter outputs in order to obtain mel frequency cesptrum coefficients.

- **Perceptual Linear Prediction (PLP)** was proposed in [Hermansky *et al.*, 1985]. Here, speaker features are calculated in a similar way as LPC coefficients, but previous transformations are carried out in the spectrum of each window aiming at introducing knowledge about human hearing behaviour. Details can be found in [Hermansky *et al.*, 1985].

- **Shifted Delta Cepstral(SDC)** was introduced in [Torres-Carrasquillo *et al.*, 2002], and arise as a means of incorporating additional temporal information about the speech into

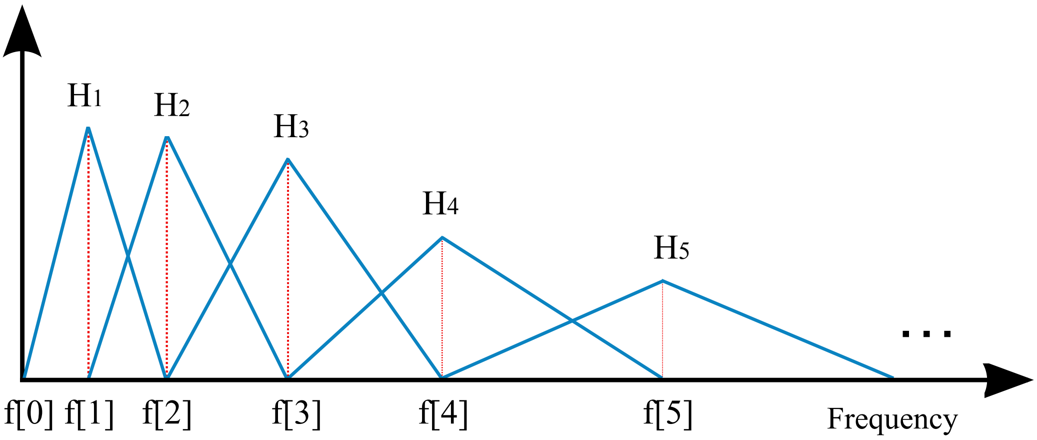

Figure 2.4: *Bank of classical Mel-filters on the MFCC feature extraction.*

the feature vector. In that sense, they are of particular interest in language recognition where units which embrace temporal information (several frames) have proved to be useful. SDC are built by stacking delta cepstral across multiple speech frames.

As mentioned above, the main aim of the described methods is to extract a feature vector for each frame or window. However, in this independent analysis possible useful information such as co-articulation can be lost. In order to take this kind of information into account, velocity (Δ) and acceleration ($\Delta\Delta$) coefficients are usually obtained from the static window-based information. This Δ and $\Delta\Delta$ coefficients model the speed and acceleration of the variation of cepstral feature vectors across adjacent windows.

2.3.2. Modelling stage

Once feature vectors are extracted from a given speaker or language, these are used to *train* a speaker or language model, which it will be stored in a database to be subject of comparison with independent test sample sources.

The generated models can be classified attending to different criteria. A common classification is to divide them into two broad groups i) *non-parametric* models and ii) *parametric* models, also known as *template models* or *stochastic models.* Through template models feature vectors belonging to training samples and testing samples are somehow directly compared, being their degree of similarity representing by the distortion encountered between them. Vector quantification (VQ) [Soong and Rosenberg, 1987] and Dynamic Time Warping (DTW) [Sakoe, 1978] are examples of this type of models for text-independent and text dependent recognition, respectively. By using stochastic models each speaker or language is assumed to follow an unknown but fixed probability density function. The parameters of this probability density function are then estimated in a training stage, while in the test stage, the degree of similarity is computed as the likelihood of the test utterance with respect to the model. Gaussian Mixture Models

(GMM) [Reynolds and Rose, 1995] and Hidden Markov Models (HMM) [Rabiner and Juang, 1986] are examples of this type of modelling for text-independent and text-dependent speaker recognition, respectively.

Regarding the training paradigm other common classification scheme is to divide models into i) *generative models* and ii) *discriminative models*. Generative models such as GMM or VQ estimate the feature distribution of each speaker or language without considering the rest of speaker/languages, while discriminative models are intended to model boundaries between speaker/languages. Support Vector Machines [Campbell *et al.*, 2006a] and Artificial Neural Networks [Farrell *et al.*, 1994] are the most popular modelling approach of discriminative models.

2.3.3. Scoring normalization

A common stage in SV and SLR systems is to normalise the similarity measures, *scores*, obtained from a given pair of test recording and target model, so as to scores from different speakers/languages share a similar range. Thus, the misalignment among non-target distributions for several speakers/language, is diminished and a common/unique threshold can be set in order to take decisions about identity.

The most common form of this type on normalization in score domain follows the form

$$\hat{s} = \frac{s - \mu_{imp}}{\sigma_{imp}} \tag{2.6}$$

where the new score $\hat{s}$ is derived by normalizing the output score through the parameters of a non-target distribution assumed to be normally distributed with mean $\boldsymbol{\mu}_{imp}$ and standard deviation $\boldsymbol{\sigma}_{imp}$. This distribution is generated via an impostor cohort of models or test recordings. The basic idea of this approach consists of modifying the non-target scores distributions to be standard normalized $N(O, I)$, with the main aim of aligning the scores distributions among different speakers.

According to how the impostor distribution are obtained to estimate μ_{imp} and σ_{imp}, there exists different ways to perform scoring normalization. The three most widely extended are

- **z-norm** or zero normalization [Auckenthaler *et al.*, 2000]. In z-norm a cohort of impostor test utterances is faced versus all the target models in the given task, deriving for each, the model-specific impostor statistics μ_{imp_λ} and σ_{imp_λ}. Then, the corresponding statistics of the model λ are used to normalize, via Equation 2.6, the set of system scores where the model λ is involved.

- **t-norm** or test normalization [Auckenthaler *et al.*, 2000]. On contrary z-norm, in t-norm a cohort of impostor models is utilised to generate the impostor score distribution. Again, μ_{imp_t} and σ_{imp_t} impostor statistics are estimated and then are applied to normalize the set of scores where the test utterance t is involved.

- **zt-norm**. z-norm and t-norm can be jointly employed given rise to the zt-norm approach following the equation

$$\hat{s}_{zt-norm}(\lambda,t) = \frac{\frac{s(\lambda,t)-\mu_{t_{znorm}}}{\sigma_{\lambda_{znorm}}} - \mu_{t_{tnorm}}}{\sigma_{t_{tnorm}}} \tag{2.7}$$

  where z-norm scores are t-normalized. Note that impostor score distributions to compute impostor t-norm statistics must be previously z-normalized to keep consistency.

2.3.4. Fusion

From the fact that different levels of information are present in the speech signal and specific systems are built to exploit a determined information level, the *fusion* of several of those systems has been shown to increase the performance of global SV and SLR systems [Brümmer *et al.*, 2007; Lopez-Moreno *et al.*, 2008]. As widely believed, the more uncorrelated information the more effective results the fusion, but it has been proven that also some improvement can be obtained by combining similar systems [Brümmer *et al.*, 2007].

The fusion approaches can be carried out at different levels of a SV or SLR system. A common and easy scheme is to perform fusion at the scoring level, that is, combining scoring coming from different systems. The simplest form is just combining the scores via a weighted sum, where a confidence in form of a weight is deposited in each system involved. This approach allows to combine totally different recognition architectures even though those are based on very different features or modelling concepts. More sophisticated approaches include to estimate/train those weights via training data, such as the fusion approach proposed in [Brümmer and du Preez, 2006] where weights are estimated via logistic regression.

Other extended method to combine systems is the *back-end* approach. The back-end approach is based on considering outputs coming from different classifier as another random variable to then using a back-end classifier to exploit the information delivered for every single system. A SVM trained via scores vectors belonging to target and non-target scores is commonly used as back-end classifier.

2.3.5. Calibration

In forensic evidence reporting, simple classic scores output from speaker verification systems are not adequate [Brümmer and du Preez, 2006; Gonzalez-Rodriguez *et al.*, 2007b; Ramos, 2007]. Instead, scores should provide the interpretation of a likelihood ratio (LR) in a forensic sense. That is, a ratio between *prosecution* and *defence* propositions defined as:

- θ_p (prosecution hypothesis). The speech recording recovered in crime scene comes from the suspect.

- θ_d (defence hypothesis). The speech recording recovered in crime scene does not come from the suspect

In the literature, this likelihood ratio is commonly presented as

$$LR = \frac{P(E \mid \Theta_p, I)}{P(E \mid \Theta_d, I)} \tag{2.8}$$

where E denotes the available evidence, which includes a recovered sample from an unknown origin and a control sample whose origin is known, and I refers to other information relevant for the case.

By using likelihood ratios a fact finder (judge or jury) is able then to compute posteriors odds, taking into account other prior information coming from other different evidences by following

$$\frac{P(\Theta_p \mid E, I)}{P(\Theta_d \mid E, I)} = LR \frac{P(\Theta_p \mid I)}{P(\Theta_d \mid I)} = \frac{P(E \mid \Theta_p, I)}{P(E \mid \Theta_d, I)} \frac{P(\Theta_p \mid I)}{P(\Theta_d \mid I)} \tag{2.9}$$

A very important fact in this sense, made clear from this formulation, is the role of the scientist, which must be limited to compute and report the likelihood term without considering prior odds.

The process of converting scores to proper likelihood ratios is referred as calibration and it is commonly a difficult task, key in the analysis of speaker recognition systems applied to forensic scenarios. Among the different proposed methods to calibrate systems, a widely adopted is a linear transformation of scores as performed in [Brümmer and du Preez, 2006] via logistic regression (*FoCal* toolkit implements this type of calibration [1]). There, this transformation is trained on background data to minimize the following cost, the so-called C_{llr}

$$C_{llr} = \frac{1}{N_{\theta_p}} \sum_{i=1}^{N_{\theta_p}} log_2(1 + \frac{1}{LR_i}) + \sum_{j=1}^{N_{\theta_d}} log_2(1 + \frac{1}{LR_j}) (2.10)$$

where N_{θ_p} and N_{θ_d} are the number of comparison available of both *prosecutor* and *defence* hypotheses.

The C_{llr} cost function deep detailed in [Brümmer and du Preez, 2006; Leeuwen and Brümmer, 2007], gives an estimation of the calibration error over all possible priors; giving an scalar measure of goodness of the total decision system.

2.4. Acoustic Systems

2.4.1. GMM

The state of the art in text-independent speaker recognition has been widely dominated during the past decade by the Gaussian Mixture Model (GMM) approach working at the short-term spectral level introduced by Reynolds *et al.* [2000]. This scheme can be seen as a likelihood ratio detector between a GMM target model and a speaker-independent GMM model, the so-called Universal Background Model (UBM). The UBM model is trained with speech (features

[1] Toolkit for Evaluation, Fusion and Calibration of statistical pattern recognizers. http://sites.google.com/site/nikobrummer/focal

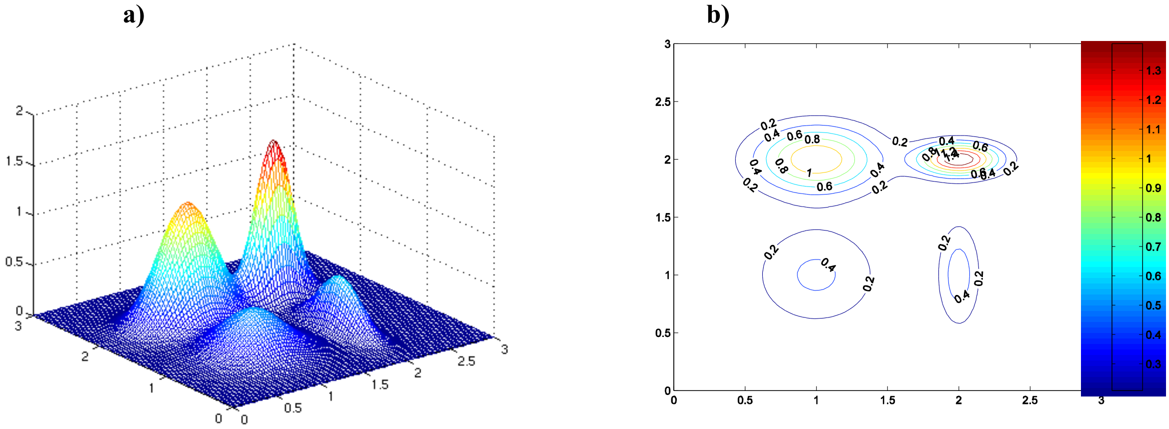

Figure 2.5: A GMM of four 3-Dimensional Gaussians (a) and their contours (b).

vectors) belonging to different speakers to represent as much as possible the speaker-independent distribution of the feature vectors, and it is used as a prior to obtain specific target GMM models via Maximum a Posteriori Adaptation (MAP). In order to obtain a similarity measure between test feature vectors and a given target model, a likelihood ratio is established between the likelihoods ratios obtained versus the target and the UBM model.

2.4.1.1. Definition

A GMM (λ) is a stochastic model composed by a weighted sum of K finite mixture of D-multivariate Gaussian densities as given by the equation,

$$p(\boldsymbol{o_t} \mid \lambda) = \sum_{i=1}^{K} w_k p_k(\boldsymbol{o}_t) \tag{2.11}$$

where $\boldsymbol{o}_t$ is a D-dimensional vector (i.e feature vector), $\{\boldsymbol{w}\}_{i=1}^{K}$ the mixture weights and $p_k(\boldsymbol{o}_t)$ is a shorthand of $N(\boldsymbol{o}_t \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, that is, a D-variate normal distribution, with probability density function of the form

$$p_k(\boldsymbol{o}_t) = N(\boldsymbol{o}_t \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = (2\pi)^{-\frac{d}{2}} \mid \boldsymbol{\Sigma}_k^{-\frac{1}{2}} \mid \exp(-\frac{1}{2}(\boldsymbol{o}_t - \boldsymbol{\mu}_k)'\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{o}_t - \boldsymbol{\mu}_k)) \tag{2.12}$$

Figure 2.5.a shows a 3D GMM formed by four Gaussian and its contours in 2D (Figure 2.5.b).

2.4.1.2. MAP adaptation

Training a GMM model consists of estimating the parameters $\lambda = \{w_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^{K}$ from a set of training observations. In order to do that a Maximum Likelihood (ML) process implemented via an Expectation-Maximization algorithm EM [Dempster *et al.*, 1977] is commonly

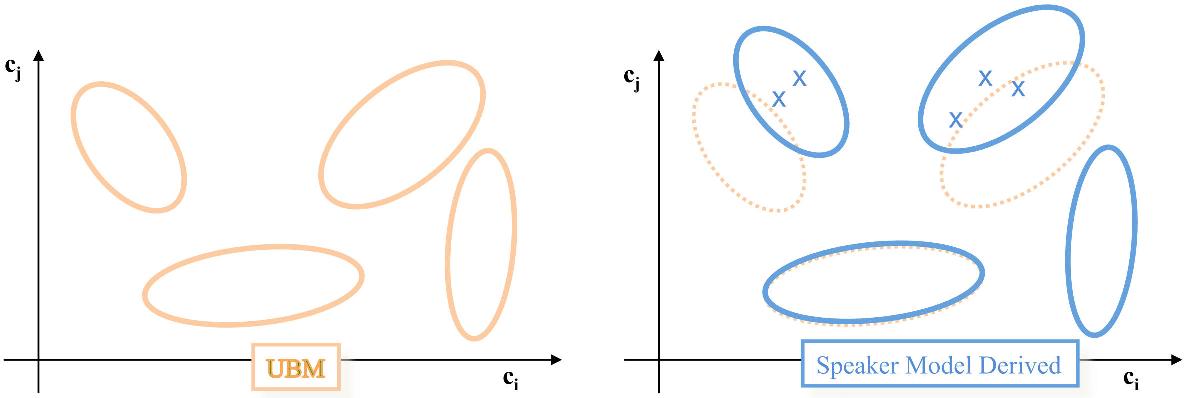

Figure 2.6: *A Maximum a Posteriori Adaptation process representation where a speaker model (right) is adapted from a UBM model (left).*

used [Bishop, 2007]. Previously, a clustering stage via K-Means (KM) [Linde *et al.*, 2003] is normally performed so as to favour a quick convergence of the EM algorithm.

However, frequently, the available speaker samples from specific speakers are not enough to robustly generate a GMM target model via clustering and ML steps. To counteract this drawback the approach GMM-UBM was proposed in Reynolds *et al.* [2000]. The underlying idea of the GMM-UBM framework lies on the fact that once a well-trained speaker-independent model is generated, this can be utilised as a prior when training specific target models. Mathematically, this step suppose turn the ML procedure to estimate new target models into a Maximum Posteriori Adaptation one [Gauvain and Lee, 1994] where the prior is represented by the UBM model.

Given the enrolment observations, $\boldsymbol{O} = \boldsymbol{o}_1, ..., \boldsymbol{o}_N$, and the UBM model, λ_{UBM}, the adapted mean new vectors are derived, as a trade-off between the UBM model means, $\boldsymbol{\mu}_k$, and the new data in the form

$$\boldsymbol{\mu}'_k = \alpha_k \frac{1}{n_k} \boldsymbol{f}_k + (1 - \alpha_k)\boldsymbol{\mu}_k \tag{2.13}$$

where

$$\alpha_k = \frac{n_k}{n_k + \tau} \tag{2.14}$$

$$n_k = \sum_t P_{kt} \tag{2.15}$$

$$\boldsymbol{f}_k = \sum_t P_{kt} \boldsymbol{o}_t \tag{2.16}$$

$$P_{kt} = \frac{w_k p_k(\boldsymbol{o}_t)}{\sum\limits_{k=1}^{K} w_k p_k(\boldsymbol{o}_t)} \tag{2.17}$$

being n_k and $\boldsymbol{f}_k$ the so-called 0th and 1st-order statistics respectively, P_{kt} the Gaussian occupation probability and τ the relevance MAP factor, which controls the importance of training samples and the UBM within the adaptation procedure. Note that the defined statistics, n_k and $\boldsymbol{f}_k$, are computed in relation to the UBM model since $p_k(\boldsymbol{o}_t)$ is defined as a normal distribution with mean $\boldsymbol{\mu}_k$ and variance $\boldsymbol{\Sigma}_k$ as in equation 2.12.

Alike, an update formula for the covariance matrices can be derived. However this has not proved to significantly outperform the global performance whilst slowing the process. For this reason, usually the covariance matrix belonging to the UBM model is shared by all the GMM models. On the other hand, it is common to do the UBM training gender dependent, that is, to estimate two different UBMs, female and male, as it has shown to be advantageous.

2.4.1.3. Log-Likelihood ratio

In the recognition stage, the final score produced from a test observations set $\boldsymbol{O} = \boldsymbol{o}_1, ..., \boldsymbol{o}_T$ and a target model λ_t is computed as a likelihood ratio between the target model, λ_t, and the UBM model, λ_{UBM}. Taking logs this takes the form

$$\mathcal{L}(\boldsymbol{O}, \lambda_t, \lambda_{UBM}) = \frac{1}{N}\sum_{t=1}^{T}\{log\ p(\boldsymbol{o_t} \mid \lambda_t) - log\ p(\boldsymbol{o}_t \mid \lambda_{UBM})\} \tag{2.18}$$

Thus, the difference of the target and the background model in generating the observations $\boldsymbol{O}$ are measured, doing comparable the score ranges of different speakers.

2.4.2. SVM

Support Vector Machines [Cortes and Vapnik, 1995; Perez-cruz and Bousquet, 2004] (SVM) are a discriminative learning technique based on minimum risk optimization, which aims at establishing a high-dimensional optimal separation boundary between two classes. Because of their flexibility and their good performance in a variety of problems, they have been widely used in the last years, both with spectral [Campbell *et al.*, 2006a] and high level features [Campbell *et al.*, 2004b; Shriberg *et al.*, 2005].

The SVM approach is based on the idea that features, which are non-linearly separable in its original space can be linearly separable in a much higher dimension by means of a hyperplane, Figure 2.7.a. The expansion to this high dimensional space is carried out by a *kernel function* $K(.,.)$, which is designed to meet the *Mercer's condition* [Burges, 1998], and therefore it can be expressed as an inner product of a mapping function θ in the form:

$$K(\boldsymbol{x}, \boldsymbol{y}) = \langle\Phi(\boldsymbol{x})\Phi(\boldsymbol{y})\rangle \tag{2.19}$$

In order to avoid the need of explicitly performing operations in the high-dimensional space, the kernel function is selected to allow the inner-product operations in the original and low-dimensional space without knowing nor caring what $\Phi(.)$ looks like, this shortcut is commonly known as *the kernel trick*.

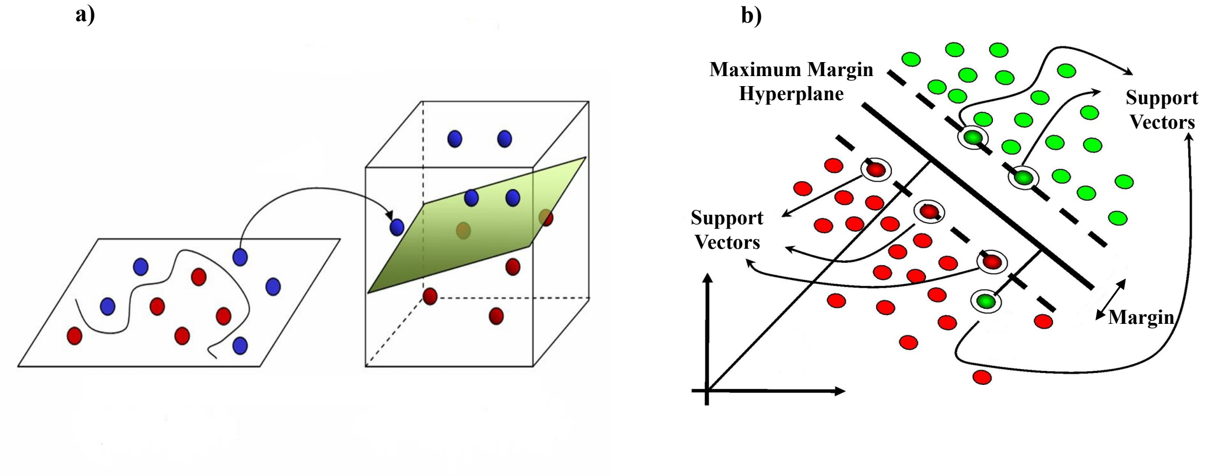

Figure 2.7: Representation of the SVM underlying idea (left) and basic elements of the SVM approach (right).

Obtaining the maximum margin hyperplane MMH is a quadratic programming problem which can be solved with classical optimization techniques. The discriminant SVM function can be expressed as:

$$f(x) = \sum_{i}^{k} \alpha_i t_i k(\boldsymbol{x}, \boldsymbol{x}_i) + d \tag{2.20}$$

where t_i are the ideal output values +1, -1, $\boldsymbol{x}_i$ are the *support vectors* associated to the MMH, α_i their corresponding weights and d the bias term estimated from the optimization process. Figure 2.7.b depicted the basic elements of the SVM approach.

2.4.3. SVM GMM-supervector

The success of MAP adaptation in conjunction with the fact that in practice only means are adapted from the UBM, derived in a new form to represent models; the means *supervector* or just *supervector*. The means supervector is formed by stacking the multivariate means vectors of a GMM Gaussians in a single and large vector as depicted in Figure 2.8.

Due to a supervector synthesises the information about a given speaker or language, it can be considered as a feature vector and as such is susceptible to be modelled. This fact was exploited in [Campbell *et al.*, 2006b], where speaker supervectors served as inputs of a SVM system, resulting in a kernel of the form:

$$K(\boldsymbol{O}_a, \boldsymbol{O}_b) = \sum_{i=1}^{n} w_i N(\boldsymbol{\mu}_i^a - \boldsymbol{\mu}_i^b;\ \boldsymbol{0}, 2\boldsymbol{\Sigma}_i) \tag{2.21}$$

where $\boldsymbol{\mu}_i^a$ and $\boldsymbol{\mu}_i^b$ are the ith mixture component of the mean speaker supervectors estimated via the observations belonging to utterances a and b respectively.

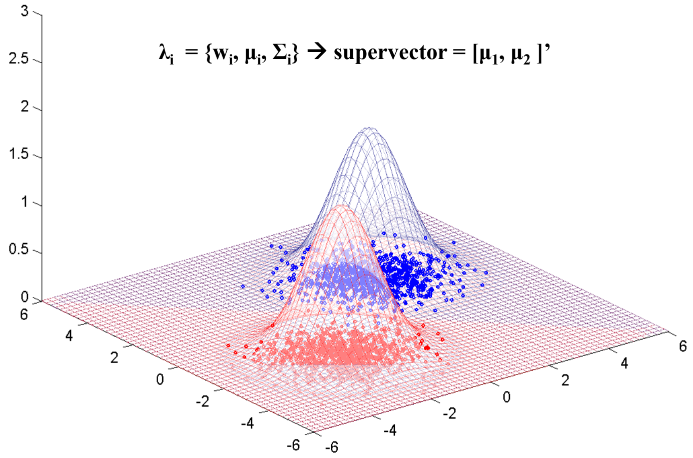

Figure 2.8: *GMM to means supervector representation.*

2.5. High-Level Systems

High-level modelling schemes can be generally divided into two steps: i) the tokenization process and ii) the statistical modelling or discriminant back-end of the extracted tokens. The nature of the modelled tokens as well as the utilised approach to get a measure of similarity define the type of the system. Next section sketches the most common statistical modelling scheme commonly used and discriminative approach based on SVMs, whilst the rest of the section describes the basis of main prosodic and phonotactic systems.

2.5.1. Statistical modelling

The most common modelling technique for tokens sequences is statistical modelling, where the probability of a sequence given a language model is used as the basis for scoring. Given a sequence of M tokens (words, phones, prosodic tokens, data driven units, etc.)

$$\boldsymbol{s} = (w_1, ..., w_M)$$

the probability of occurrence can be decomposed as a product of conditional probabilities

$$P(w_1, ..., w_M) \simeq \prod_{i=1}^{M} P(w_i \mid w_1, ..., w_{i-1}) \tag{2.22}$$

Usually equation 2.24 is approximated by limiting the context:

$$P(w_1, ..., w_m) \simeq \prod_{i=1}^{M} P(w_i \mid w_{i-N+1}, ..., w_{i-1}) \tag{2.23}$$

for some $N \geq 1$. Due to reasons of data sparsity N is usually selected in the range of 1 to 4. Estimates of probabilities in n-gram models are commonly based on maximum likelihood estimates – that is, by counting events in context on some given training text:

$$P(w_{i-N+1}, ..., w_M) = \frac{C(w_{i-N+1}, ..., w_i)}{C(w_{i-N+1}, ..., w_{i-1})} \tag{2.24}$$

where $C(.)$ is the count of a given word sequence in the training text. For robust estimation, probability smoothing techniques can be applied.

2.5.2. Phone SVM

Instead of a generative statistical modelling, a discriminative approach to manage extracted tokens (either prosodic or phonotactic) was proposed by Campbell *et al.* [2004a]. This approach is based on using a SVM to separate high-level *supervectors*, being those created by concatenating the (uni-, bi-, tri-) grams frequencies into a single vector. As showed in [Campbell *et al.*, 2004a] those frequencies can be normalized in function of a background set in order to obtain more reliable results.

2.5.3. Prosodic systems

A prosodic system essentially consists of two main building blocks: the prosodic tokenizer, which analyses the prosody features, and represents it as a sequence of prosodic labels or tokens and the N-gram statistical language modelling stage (Section 2.5.1), which models the frequencies of prosodic tokens and their sequences for each particular speaker.

A typical tokenization process usually consists of two stages. Firstly, for each speech utterance, both temporal trajectories of the prosodic features, (fundamental frequency or pitch and energy) are extracted. Secondly, both contours are segmented and labelled by means of a slope quantification process.

The slope quantification process is then performed as follows: first, a finite set of tokens is defined using level-based quantization of the slopes (e.g fast-rising, slow-rising, fast-falling, slow-falling) for both energy and pitch contours [Adami *et al.*, 2003]. Thus, the combination of levels generate different tokens when combined pitch and energy contours are considered.

Second, both contours are segmented using the start and end of voicing and the maximums and minimums of the contours. These points are detected as the zero-crossings of the contours derivatives using a frame span (typically ±2). Thus, each segment is converted into a set of tokens which describe the joint-dynamic variations of slopes. Utterances with different sequences of tokens contain different prosodic information.

2.5.4. Phonotactic systems

Phonotactic systems use phonetic transcribers to convert speech into a sequence of tokens where each token is a phone. A typical phonotactic speaker recognition system consists of two main building blocks: the phonetic decoders, which transform speech into a sequence of phonetic labels and the n-gram statistical language modelling stage (Section 2.5.1), which models the frequencies of phones and phone sequences for each particular speaker/language. The phonetic decoders can either be taken from a pre-existing speech recognizer or trained ad hoc. Any speech recognition technology can be used, but usually phonetic decoders are based on HMM and null grammars. One of the most common technique for SLR is an extension of phonotactic systems called Parallel Phone Recognition and Language Modelling (PPRLM) [Zissman, 1996]. Basically, it consists on the fusion of several phonotactic systems as described above, related to phonetic decoders in several languages, not necessarily related to the target ones [Gonzalez-Rodriguez *et al.*, 2007a; Montero-Asenjo *et al.*, 2006; Toledano *et al.*, 2007]. Using the transcriptions, statistical grammars are applied and the scoring process is performed in the same way as for speaker recognition. Sum fusion is the most commonly applied fusion technique. In order to train each of the underlying phonetic recognisers, multilingual speech corpus are required, but they do not need to contain labelled speech in the target language. The only requirement is to have labelled in a certain number of language (and in the appropriate amount to train a phonetic recogniser).

From a SV perspective (it would be similar for SLR), once a phonetic decoder is available, the phonetic decodings of many sentences from many different speakers can be used to train a Universal Background Phone Model (UBPM) that models all the possible speakers. These models are then adapted to the characteristics of a particular speaker using the UBPM and several phonetic decodings of that particular speaker to generate a Phone Model (PM_i). This process is more robust than training the speaker model from scratch because the speech available to train a speaker model is often limited. The amount of data available to perform this adaptation as well as the complexity of the N-gram modelling influences the optimal weight of the UBPM in the adaptation process, which has to be adjusted for each particular decoder. Once the statistical language models are trained, the procedure to verify a test utterance against a speaker model PM_i is represented in Figure 2.9. The first step is to produce its phonetic decoding, X, in the same way as the decodings used to train PM_i and UBPM. Then, the phonetic decoding of the test utterance, X, and the statistical models (PM_i, UBPM) are used to compute the likelihoods of the phonetic decoding, X, given the speaker model PM_i and the background model UBPM. The recognition score is the log of the ratio of both likelihoods (Figure 2.9), where the higher the score the higher the similarity between training and test speech. This process may be repeated for different phonetic decoders (e.g., different languages or complexities) and the different recognition scores simply added or fused for better performance.

2.6. A Need for a Session Variability Compensation Approach

It is widely agreed that the main cause of performance degradation in both SV and SLR systems is due to *session variability* [Bimbot *et al.*, 2004; Kenny and Dumouchel, 2004b; Kinnunen and Li, 2009], defined this as the set of differences between recordings belonging to a

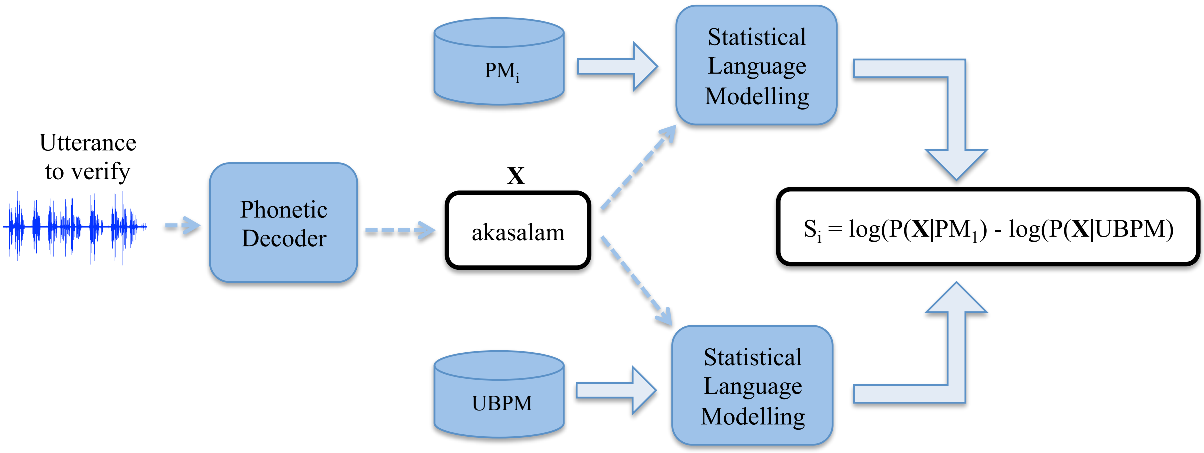

Figure 2.9: *Scheme of a Phonotactic Language Modelling for recognition.*

same identity (speaker or language according to the task). Session variability, although often referred as *channel* variability, it is caused by numberless factors that go beyond the acquisition channel such as environmental factors (e.g speech recorded in different places or situations), the speech style (e.g formal or informal speech, conversational or interview speech) etc. Indeed, as mentioned before, any variation between two recordings of the same speaker (or language) can be considered session variability and it strongly hinders the recognition task as this variation is *entangled* with the actual discriminative information.

In order to be precise and to disambiguate among the different commonly used terms in the literature, it is convenient, at this point, to define the possible types of variability that can be found within a speech signal:

1. **inter-session variability**. The inter-session variability is the set of differences between two recordings belonging to a same identity either. It can be caused by a myriad of different factor such as the channel acquisition, the environment noise, the speech style etc.

2. **intra-session variability**. The intra-session variability term is used to embrace the set of differences within a same recording, such as those produced by a change in the vocal effort of the speaker, a noise produced in some part of a recording etc.

3. **inter-speaker (or language) variability** The inter-speaker variability refers to the set of differences between recordings belonging to different identities due just to dissimilarities among them. As such, it represents the discriminative information exploited by the SV or SLR systems in order to perform the recognition task.

4. **intra-speaker (or language) variability** The intra-speaker variability is the set of differences between one or several recordings belonging to a same identity just due to changes related to the identity (e.g age variation, phone variation, speech style etc.)

For the sake of clarity, through this Dissertation we refer to the union of points 1, 2 and 4 under the term of session variability while point 3 will be referred as speaker variability.

2.6.1. Taxonomy of compensation approaches

During the last three decades a broad number of different techniques to palliate the harmful effects of session variability have appeared. All of them can be classified attending to the following three criteria:

- **Domain of Application.** Session variability compensation can be performed at different levels of the whole SV and SLR system. In particular three domain, namely the feature, the model and the hybrid statistic domain have been largely the focus of the session variability techniques.
- **Need of training data.** Other interesting aspect and classification criteria lies on the fact that whether the technique demands or not training data to be somehow *trained* before its application. Techniques that do not need training data are known as *blind* techniques and they have the main advantage that can be applied in any scenario, no matter if training data is available. On the other hand *trained* techniques have the advantage of yielding a better adaptation to the scenario conditions.
- **Need of labelled data.** Apart from the need of having training data, some techniques can demand to have available the labels associated to this data. Those labels are then used to better exploit the specificities of some kind of session variability. This is the case of some techniques which exploits the type of channel acquisition labels as it will be shown in next section.

2.6.2. A historic and discrete view

Until the development of techniques based on Factor Analysis (FA), as it will be extensively discussed in next chapters, the session variability compensation techniques were designed under two prime principles:

1. Suppress the session variability.
2. Treat the variability as a combination of discrete sources rather than continuous.

Below, the most successful techniques dealing with session variability are listed in chronological order of appearance.

2.6.2.1. Cepstral Mean Subtraction

Cepstral Mean Subtraction [Furui, 1981], also known as *cepstral mean normalization*, is one of the earliest and most widely extended methods employed to ameliorate the effects of intersession variability in ASR, SV and SLR systems.

As it is well-known, a convolutional distortion in the time domain, such as that introduced by a channel, corresponds to an additive bias component in the cepstral domain. Denoting the signal $y(n)$ as the convolution of a clean $s(n)$ and a noisy $h(n)$ sources, their cepstral features are tied by the following relation

$$y(n) = s(n) * h(n) \Leftrightarrow c_y = c_s + c_h \tag{2.25}$$

Under the assumption that the channel signal $h(n)$ does not significantly vary over the duration of the utterance, CMS aims to remove the effect of $h(n)$ in the cepstral domain by removing from each feature vector, $\{c_y\}_{i=1}^{N}$, the arithmetic mean of those. Thus, consistent additive noise in the cepstral domain is eliminated.

2.6.2.2. RASTA

Often, the temporal properties of environmental effects are quite different from the temporal properties of the speech. The *RelAtive SpecTrAl* (RASTA) filtering approach [Hermansky and Morgan, 1994] attempts to exploit these differences in order to produce robust representations for speech recognition and signal enhancement.

Specifically, RASTA works under the assumption that the rate of change of non-linguistic components in speech does not match typical rate change of the vocal tract shape, and therefore highly varying and slowly varying components should be removed.

This process is performed by applying the following band-pass filter on the time trajectories of the features vectors

$$H(z) = 0.1z^4 * \frac{2 + z^{-1} - z^{-3} - 2z^{-4}}{1 - 0.98z^{-1}} \tag{2.26}$$

where components under or over the low and high cut-off defined frequencies are removed as considered non-speech components. RASTA can be seen as an evolved version of CMS where not only the continuous noise component is removed but also those components whose rate of change lies outside of that considered for speech.

2.6.2.3. Warping

Feature Warping, [Pelecanos and Sridharan, 2001], aims to eliminate channel distortions by conditioning and conforming the individual cepstral feature vectors to follow a Gaussian distribution over a window of speech frames. It is supposed here that *clean* or *true* cepstral features follow a determined distribution (Gaussian), which the additive noise and channel distortions modify.

The aim of Feature Warping is to retrieve this original form by *warping* cepstral features to a Gaussian distribution. This process is carried out by locating and reordering the original cepstral features according to a Gaussian distribution into a sliding window of frames of typically 3 seconds.

2.6.2.4. Feature Mapping

Feature Mapping, [Reynolds, 2003], extends the Speaker Model Synthesis (SMS) approach [Teunen *et al.*, 2000] to the feature domain, with the main advantage that once feature vectors are clean, any recognition structure or modelling approach can be used.

The mapping process can be summarised as follows:

1. A channel independent GMM background model, λ_{ci}, is trained using data from different channels.

2. Channel dependent GMM background models, λ_{cd}, are then trained by adapting the λ_{ci} model from channel dependent data via MAP adaptation.

3. Model parameters differences (mean supervector differences) between each λ_{cd} model and the λ_{ci} one indicate how the feature space distributions between two spaces are related. This information is used to mapped feature vectors of each utterance to channel independent space.

4. Given a utterance, first the most likely channel dependent model is detected and then all its feature vectors mapped to the channel independent space following the form

$$\hat{\boldsymbol{o}}_t = \boldsymbol{o}_t - \sum_{i=1}^{k} (\boldsymbol{\mu}_i^{cd} - \boldsymbol{\mu}_i^{ci}) \tag{2.27}$$

   where $\boldsymbol{\mu}_i^{cd}$ and $\boldsymbol{\mu}_i^{ci}$ are the ith component of the mean channel dependent and channel independent supervector respectively.

2.7. Summary

In this chapter a global vision of SV and SLR systems from the very beginning stage of information extraction to the decisions concerning identity has been presented. The different levels of information presented in the speech signal has been highlighted and the most successful feature extraction approaches which aim to take advantage of all this information detailed.

The SV and SLR systems are sequential, modular-based systems, which are nowadays the result of the accumulated efforts of numberless researches which with specific contributions on some parts of the global systems have contributed to yield more and more efficient and accurate systems day by day. Most success techniques in each stage namely, feature extraction, modelling and scoring (computing the similarity) has been detailed.

The final part of the chapter has been devoted to present the most successful existing techniques to deal with session variability before Factor Analysis. The inter-session, or simply, session variability problem is largely considered the prime cause of performance degradation of both, speaker and language recognition systems and main motivation of this Dissertation. The techniques presented in this chapter are based on i) suppress session variability and ii) treat variability as a combination of discrete sources. It will be work of the next chapter to refute those principles exploring the new techniques based on Factor Analysis in order to palliate the session variability problem.

Chapter 3

A Continuous Approach to Variability Modelling: The Joint Factor Analysis Model

This chapter presents and details the grounds of the Joint Factor Analysis modelling approach applied to SV and SLR.

3.1. Introduction

Despite their relative success, the techniques described in the previous chapter designed to deal with session variability suffer from one or both of the following major deficiencies: i) categorize the session conditions and/or ii) suppress the undesired variability according to a general rule, rather than modelling the specific variability within a given recording.

The former (i) clearly does not fit the true nature of the problem. Even though some careful and conscientious recordings classification could be performed regarding several global traits, such as the acquisition channel or type of speech, real session conditions are, from a practical point of view, difficult to quantify. Feature Mapping or Speaker Model Synthesis fall into this group. The latter (ii) goes a step further, questioning the manner in which the session variability issue is addressed. Since each recording is generated under specific and usually non-controllable circumstances, intuitively, inferring general rules in order to suppress session variability in a global manner should be less beneficial than considering the variability associated to a given recording as unique. Techniques such as CMN, Rasta or Feature Warping fall into this group.

These arguments motivated researchers to find a new methodology supported by more ambitious principles, designed to somehow counteract the aforementioned drawbacks. In this context, the techniques based on Factor Analysis (FA), main focus of this chapter, emerged.

The FA modelling approaches break with the established manner of conceiving the variability associated to a speech signal when recognizing speakers or languages by embracing the following two principles:

- Considering variability as a continuous source rather than discrete.
- Explicitly modelling both session and inter-speaker/language variability.

Apart from these two pillars, another fundamental idea, formulated initially as a hypothesis, define the FA based approaches. This hypothesis can be stated as:

- **Much of the variability associated to a given recording lies within subspaces of a much lower dimensionality than the original space (i.e, the model space).**

That is, it is possible to find both speaker/language and session variability subspaces, so that they act as priors in order to disclose the specific variability contained in a given recording.

This chapter is intended to give an in-depth vision of the grounds of FA-based approaches designed to deal with variability into SV and SLR systems. The first part of this chapter chronologically traces the history of the use of subspaces as a powerful tool to manage variability from its origins into some related fields, such as face or speech recognition, to its inclusion in SV by means of FA, focussing on the key papers or research milestones that has led to the current state-of-the-art SV and SLR systems. The second part is devoted to provide a mathematical understanding of the FA model from its generic form to its adaptation to be incorporated to SV and SLR systems.

3.2. From Eigenfaces to Joint Factor Analysis Model

Linked by a common set of problems, it is not surprising that some work performed in one of the related fields of face, speech, speaker and language recognition have been mirrored or inspired among them. This is the case of the FA approach applied to SV or SLR, which took much of its basis lines from the *eigenvoice* technique previously used in speech recognition, and which in turn was inspired by the *eigenfaces* approach in face recognition. The remainder of this section aims to guide the reader to the use of variability subspaces in SV and SLR from the previous studies and success in nearby and related areas.

3.2.1. Eigenfaces

The *eigenfaces* approach, introduced within the automatic face recognition field by Turk and Pentland [1991], is based on the assumption that an *unknown* face image may be approximated by the combination of other set of *known* face images. Specifically, to represent a face image, the eigenface approach proposes a linear combination of a *few relevant directions* extracted from the analysis of the variance in a background bank of images. Thus, a face image is defined by the weights associated to these fixed directions.

This idea, introduced first, into the pattern recognition domain by Watanabe [1965] and, later on, extended by Kirby and Sirovich [1990], within their work on the characterization of human faces, is stated as the assumption of *low-dimensionality of the face variation*: the dimensionality of the face space, defined as the space of variation of face images with same orientation and scale, is much smaller than the dimensionality of a single face considered as an arbitrary 2-D image. Put another way, there is a low-dimensional space that *embeds* the variation of face images with same orientation and scale and from which any face image can be approximated.

Although at a first glance could not seem intuitive, the *eigenfaces* concept links well with the human manner of recognizing faces. If we take some time to think about how we are able to perceive and discern faces, we will quickly realize that it is intrinsic to our reasoning when describing a human face to make references to some components of other faces familiar to us: "She's got the same eyes that my friend . . . ", or "her nose is similar to that actress". So, the idea of reassembling a human face as a set of *elemental pieces* coming from our own background bank of previous human faces seems not to be far from our inherent human faces pattern recognition machinery.

From a pattern recognition machine the approach has evident advantages. First, the face images are represented in a compact way leading to a major benefit of the computational requirements. Second just a few number of free parameters must be estimated in order to train a model of each face image, so the requirements of training data is also greatly diminished.

Other interesting relation can be also established within the information theory field. Representing an image as a linear combination of the *principal face components* can be seen as an efficient manner to encode the image information and therefore does minimize the necessary number of bits to represent the image whilst avoiding undesired noise.

Regarding the estimation procedure to estimate the variation subspace, that is, finding the principal elements of variation of a given background dataset, the Principal Component Analysis approach (PCA, [Pearson, 1901]) utilised in [Turk and Pentland, 1991] is a well-known candidate.

The complete classification process by the eigenfaces approach can be seen then as a two-encoding based procedure divided into three stages, as depicted in Fig. 3.1. In the development stage, PCA is applied over a the set of M available background images $\mathbf{B}$, being images represented as *points* in a D-dimensional space (usually high dimensional). The top K ($D >> K$) eigenvectors of the covariance matrix $\mathbf{C} = (\mathbf{B}\mathbf{B}^T)$, those corresponding to the largest K eigenvalues are then retained yielding the variation subspace $\mathbf{A}$. Then the T training images denoted by $\mathbf{M}$ are projected into this subspace to obtain a proper low-dimensional representation in the training stage; this projection is denoted by $\mathbf{A}^T\mathbf{M}$. In order to classify a test image $\mathbf{t}$, several variants can be followed. The most simple and used in [Turk and Pentland, 1991] is to project it into the subspace as in the same manner than training images to finally compute a euclidean distance $\mathbf{s}$ of this projection with each of the projected training images; the distance function is denoted in the figure by $d(\mathbf{M}, \mathbf{T})$ while $\mathbf{S}$ represents the matrix of distances. Finally, a threshold Θ for every class will mark the decision of acceptance or reject.

In order to reconstruct projected images from the low to high-dimensional space as it is

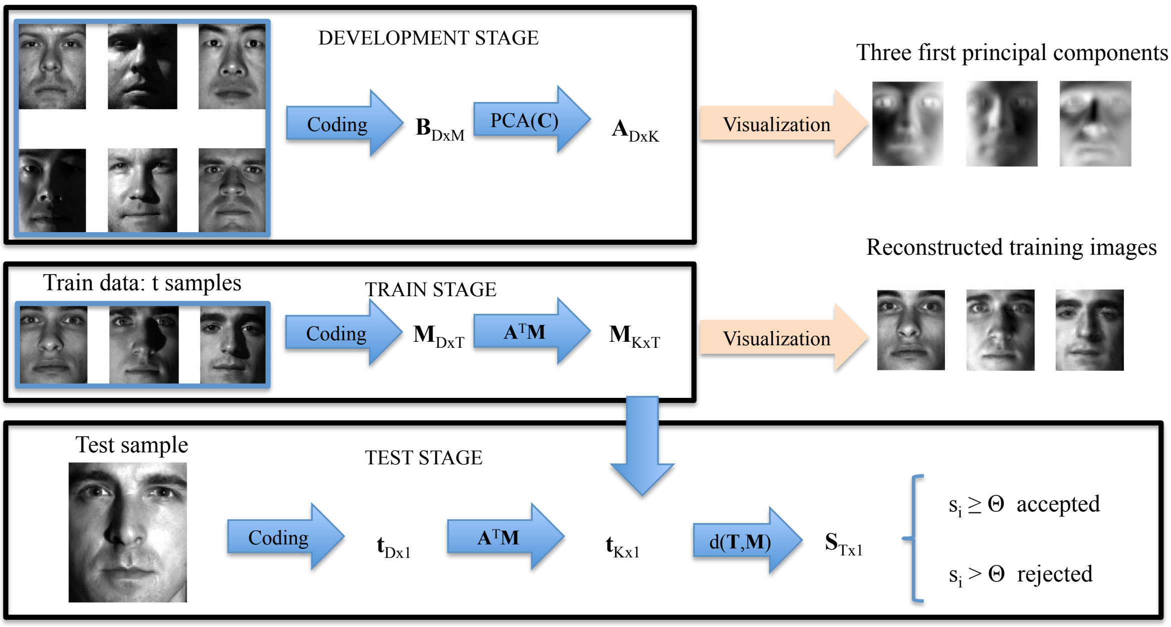

Figure 3.1: *Global scheme of a eigenfaces for classification approach.*

done for visualization in Fig. 3.1, note that if the projection matrix **A** is orthogonal (formed by orthonormal unit vectors), then $\boldsymbol{A}^T = \boldsymbol{A}^{-1}$ and therefore reconstruction can be carried out by a **A** right multiplication.

3.2.2. Eigenvoices

The work peformed by Turk and Pentland on *eigenfaces* was soon mirrored into the speech recognition field under the concept of *eigenvoices*: the directions that best represent the variation among different speakers.

The eigenvoice modelling [Kuhn *et al.*, 2000] was first conceived to cope with the issue of speaker adaptation in speech recognition applications when tiny amount of speaker-specific data is available (e.g digit or letter recognition). So far, the speaker adaptation process to turn a speaker independent speech recognition system into a speaker dependent one had been performed via standard algorithms as maximum a posteriori (MAP, [Gauvain and Lee, 1991]) or maximum likelihood linear regression (MLLR, [Gales and Woodland, 1996]). However, even though these methods achieve reasonable performance and do not require large amount of data, they fail when just very limited data is available.

By confining the models to a low-dimensional subspace obtained previously from a background set of training data the number of degrees of freedom to be estimated is drastically diminished, so that even at the presence of scarce amounts of data it is possible to estimate reasonable models.

Under the same idea, the *eigenvoice* approach was introduced within the field of SV as a replacement of MAP Gaussian Mixture Models adaptation in those cases where a sparse amount of specific speaker training data was available [Thyes *et al.*, 2000]. The classification process is similar to that presented for *eigenfaces* in Figure 3.1, but representing training and testing recordings rather than face images as high dimensional points by means of their speaker means supervectors (Section 2.4.3).

A step further on the use of the eigenvoice modelling was investigated in [Lucey and Chen, 2003] and [Kenny *et al.*, 2005a] where a prior probability distribution for the speaker's supervector was considered within the eigenvoice modelling estimation. Those approaches can be met under the term *eigenvoice MAP* (EV-MAP), since a maximum posteriori estimation is utilised instead of maximum likelihood. The prime advantage of the EV-MAP approach is that it reduces even more than the eigenvoice approach the dependence on the data when training speaker models, as the additional prior constrains/drives the posterior distribution.

It is convenient to highlight at this point that although all those approaches are useful in sparse data scenarios, their success is conditioned to the correct estimation of the speaker subspace. If this does not properly represent the speaker variability, the adaptation process will lead the global system to fail.

3.2.3. Eigenchannels

Although the *eigenvoice* and EV-MAP approaches led to significant improvements in the speaker recognition field in some scenarios - those where scarce specific training data was available -, the great step towards a much more accurate technology took place when this framework was viewed under the perspective of the session variability.

Under the idea of adapting a speaker model to a given *channel* as a speaker-independent model is adapted to a given speaker, the *eigenchannel MAP* (EC-MAP) approach was presented in [Kenny *et al.*, 2003]. The EC-MAP approach shares exactly the same principles that EV-MAP, but whereas the latter needs of a low-dimensional speaker space, the former requires a low-dimensional session subspace.

In the methodology proposed in [Kenny *et al.*, 2003], EC-MAP was designed to deal with the session variability at recognition time. Once the speaker models are adapted from an speaker-independent model (e.g UBM) to the target speaker via classical MAP or EV-MAP, then they are adapted to the specific *session effects* of each target test utterance. Thus, the target model is shifted to the specific channel type of the test utterance, avoiding the possible channel mismatch between training and testing utterances.

The success of this methodology brought forward the convenience of explicitly modelling session variability in a continuous manner, leading to more sophisticated approaches as the Joint Factor Analysis (JFA) presented below, which today conform the state of the art in acoustic SV and SLR systems.

3.2.4. The Joint Factor Analysis model

3.2.4.1. Introduction

The previous studies presented in the above section fed the idea of explicitly modelling both the speaker and session variability in a separate and continuous manner under a dual goal. First, to adequately explain the speaker variability and second, and most important, to deal with the session variability issue. Specifically, the scientific community in the field began to be interested in jointly solving the next two questions [Kenny and Dumouchel, 2004b]:

1. How is it possible to adapt a speaker model to the session effects of the enrolment data without performing speaker adaptation?
2. How is it possible to estimate a speaker model independently on the session effects of the enrolment data?

By means of EC-MAP the former question had been solved but the latter remained unanswered. In this context, Joint Factor Analysis (JFA) [1] emerged in the SV field as a framework or modelling technique to jointly respond those questions by properly combining MAP, EV-MAP and EC-MAP approaches under the following hypothesis:

- A speaker means supervector is formed by two components, one expresses the specific speaker information, the other the session distortion related with the recording/training data
- Much (but not all) of the speaker or session variability can be explained by a small number of hidden variables connected with pre-trained subspaces of the supervector space.

The remainder of this section is devoted to build step by step the Joint Factor Analysis model. The analysis starts building the speaker component to then adding the session component, carefully explaining the involved *elements* of the JFA modelling in each stage. For the sake of conciseness, maths behind the model has been set aside and they are extensively introduced later on in this chapter (Sections 3.3, 3.4, 3.5).

3.2.4.2. Definition

Given a classical GMM system with C Gaussian components and F feature dimensions, where a UBM has been previously trained, it can be seen that, by classical MAP, a speaker-dependent means model supervector $\boldsymbol{\mu}_s$ $(CF \times 1)$ of a new speaker s is derived from the UBM means supervector $\boldsymbol{\mu}$ as

$$\boldsymbol{\mu}_s = \boldsymbol{\mu} + \boldsymbol{D}\boldsymbol{z}_s \tag{3.1}$$

[1] We denote by the term Joint Factor Analysis (JFA) the specific modelling strategy designed for SV and SLR based on Factor Analysis, as it jointly models both speaker and session variability. The term Factor Analysis is used to generically refer to the classical mathematical model in which all the subspaces techniques presented are based.

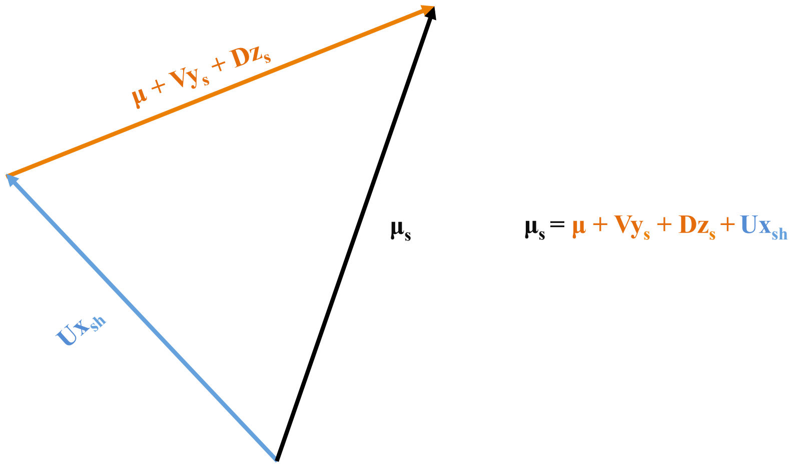

Figure 3.2: *Representation of the speaker supervector decomposed in the speaker and session variability components.*

where the term $\boldsymbol{D}\boldsymbol{z}_s$ represents the *shift/offset* from the mean $\boldsymbol{\mu}$ as a result of the MAP adaptation, and it is formed by the diagonal $CF \times CF$ matrix $\boldsymbol{D}$, and the $CF \times 1$ weights vector $\boldsymbol{z}_s$ which is assumed to be distributed with a standard normal prior (this derivation is detailed in Section 3.4.2, but this result is enough to follow the reasoning).

By the form in equation 2.13 and assuming the prior of $\boldsymbol{z}$ standard normal distributed, it can be inferred that, in MAP, speaker-dependent means supervectors are considered to be normally distributed with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{B} = \boldsymbol{D}^2$, $CF \times CF$. An analogous analysis can be performed with EV-MAP, but considering the variance of the distribution to be confined within a subspace of rank R_s within the supervector space, where $R_s << CF$. Note that the implicit assumption formulated in EV-MAP is then that the eigen-analysis of covariance $\boldsymbol{B}$ results on a few non-zero eigenvalues, exactly R_s. In matrix form

$$\boldsymbol{\mu}_s = \boldsymbol{\mu} + \boldsymbol{V}\boldsymbol{y}_s \tag{3.2}$$

where $\boldsymbol{V}$ is a low-rank matrix ($CDxR_s$) which explains the speaker variance, in this case $\boldsymbol{B} = \boldsymbol{V}\boldsymbol{V}^T$ and $\boldsymbol{y}_s$ the *weights* which represent the speaker s through the speaker variability subspace spanned by $\boldsymbol{V}$. Note, nevertheless, that by varying $\boldsymbol{y}_s$, the model $\boldsymbol{\mu}_s$ varies across the space spanned by $\boldsymbol{V}$; that is within a R_s-dimensional linear manifold of the supervector space.

JFA integrates both modelling ideas in order to derive the speaker-dependent component of a mean speaker supervector model. So that

$$\boldsymbol{\mu}_s = \boldsymbol{\mu} + \boldsymbol{V}\boldsymbol{y}_s + \boldsymbol{D}\boldsymbol{z}_s \tag{3.3}$$

Note that by this form the assumed variance $\boldsymbol{B}$ is now explained by both $\boldsymbol{V}$ and $\boldsymbol{D}$ ($\boldsymbol{B} = \boldsymbol{V}\boldsymbol{V}^T + \boldsymbol{D}^2$), and as such, it combines the advantages of MAP and EV-MAP: first, the variability is supposed to be, to great extent, constrained in the subspace spanned by $\boldsymbol{V}$; and second, other

Term	*Description*	*Dimensionality*
$\boldsymbol{\mu}$	Mean of the new models. Usually, the UBM speaker mean supervector.	$CF \times 1$
$\boldsymbol{V}$	Speaker Variability Subspace. Low-rank matrix.	$CF \times R_s$
$\boldsymbol{D}$	Besides $\boldsymbol{z}_s$, residual speaker term. Full-rank matrix (diagonal)	$CF \times CF$
$\boldsymbol{U}$	Session Variability Subspace. Low-rank matrix.	$CF \times R_c$
$\boldsymbol{y}_s$	Speaker Factors	$R_s \times 1$
$\boldsymbol{z}_s$	Besides D, residual speaker term	$CF \times 1$
$\boldsymbol{x}_h$	Channel Factors	$R_c \times 1$

Table 3.1: *JFA model components description, (equation 3.5).*

speaker variability out of this manifold is also accounted. The vector $\boldsymbol{y}_s$ $(R\times1)$ is usually referred to as ***speaker factors***, since represents the speaker variability within $\boldsymbol{V}$, and mathematically responds to the latent factors within a FA modelling as it will be shown later on in the second part of this Chapter.

Once the speaker-dependent component has been established, the session-dependent component of the means speaker supervector is incorporated. By JFA, it is assumed that every utterance h corresponding to a speaker s produces a *distortion* in its speaker mean supervector and this can be modelled via EC-MAP. The supervector space is then modified by and additional term as

$$\boldsymbol{\mu}_{sh} = \boldsymbol{\mu}_s + \boldsymbol{U}\boldsymbol{x}_{sh} \tag{3.4}$$

where $\boldsymbol{U}$ is a low rank matrix $(CF \times R_c)$ that plays the same role than $\boldsymbol{V}$ in EV-MAP but representing the session variability subspace, and $\boldsymbol{x}_{sh}$ is the analogous term of $\boldsymbol{y}_s$. The components of $\boldsymbol{x}_{sh}$ are usually called ***channel factors*** and unlike the speaker factors, those depend on the utterance h apart from the speaker s.

Summing up, the Joint Factor Analysis, geometrically represented in Figure 3.2, is formulated in matrix terms as

$$\boldsymbol{\mu}_{sh} = \boldsymbol{\mu} + \boldsymbol{V}\boldsymbol{y}_s + \boldsymbol{D}z_s + \boldsymbol{U}\boldsymbol{x}_{sh} \tag{3.5}$$

Table 3.1 describes each component of the model.

Thus, given a recording or training material h belonging to the speaker s, the JFA model is composed by the tuple of speaker-independent *hyperparameters* $\Lambda = \{\boldsymbol{\mu}, \boldsymbol{V}, \boldsymbol{D}, \boldsymbol{U}\}$, the speaker-dependent *factors* $\boldsymbol{y}_s, \boldsymbol{z}_s$ and the speaker- and utterance-dependent $\boldsymbol{x}_{sh}$ factors. As it will be shown later on, the hyperparameters are pre-trained in a development stage, and remain fixed for all speakers and utterances both in training and testing stages. On the other hand, the set of factors are estimated per each utterance given the speaker-specific data and trained hyperparameters.

At this point, it is convenient to highlight some considerations about the JFA model:

- The JFA model generalizes MAP, EV-MAP and EC-MAP. In fact, the three models can be achieved from JFA by suppressing or zeroing terms (e.g EV-MAP can be obtained by setting to zero $\boldsymbol{U}$ and $\boldsymbol{D}$).
- The factors $\boldsymbol{z}_s$, $\boldsymbol{y}_s$ and $\boldsymbol{x}_{sh}$ are considered to be standard normally distributed $N(0, I)$.
- There is not an analogous term to $\boldsymbol{Dz}$ on the session variability component. Note that if there was, the session component would cover the entire supervector space. This would allow to turn a given speaker into any other just by varying session effects.

3.3. Factor Analysis: The Model

The Joint Factor Analysis approach is based on classical Factor Analysis. The following sections, second part of this chapter, offer a mathematical understanding of this model from its original application in a multivariate Gaussian framework to its adaptation to be incorporated into a mixture of multivariate Gaussian densities, and being applied to SV and SLR systems.

3.3.1. Latent variables models

3.3.1.1. A brief historical review

A Latent Variable Model (LVM) is a statistical model that try to explain a high-dimensional process in terms of a low-dimensional set of non-observed variables [Spearman, 1904]. The variables belonging to the original high-dimensional process are called the manifest or *observed* variables, and those which explain the underlying low-dimensional structure of the process are the hidden or *latent* variables.

The LVMs were first introduced in the field of psychometrics by Spearman [1904] with the purpose of discovering/modelling underlying correlations between certain mental conditions or attitudes and the results extracted from several human tests. In this direction, in his work about the *general intelligence* factor (g-factor), Spearman used a LVM to evaluate the correlation between the *mental ability* of children and a set of variables, which were directly extracted from cognitive ability tests. Here, the set of variables derived from the tests played the role of the observed variables, supposed to be somehow connected with the *mental ability*, the latent factor of the model.

The success of those first studies besides the attractive idea of simplifying high-dimensional statistical process by explaining those via low-dimensional structures, went through the scope of psychometrics and led to wider studies in the statistics field, which derived on great advances in the multivariate analysis area. Those advances cover among others, the development of the broadly-used statistical approaches as latent structure analysis [Lazarsfeld and Henry., 1968], Factor Analysis [Bartholomew, 1987; Bartholomew *et al.*, 2011], and also the consolidation of Principal Component Analysis [Pearson, 1901] [Hotelling, 1933], which was not considered traditionally as a LVM.

A first categorization of the LVMs was established by Bartholomew [Bartholomew, 1987], according to the nature of the observed and latent variables (continuous or discrete). The following table sums up this classification:

		Observed variables	
		Continuous	*Discrete*
Latent Variables	*Continuous*	**Factor Analysis**	Latent Trait Analysis
	Discrete	Latent Profile Analysis	Latent Class Analysis

Table 3.2: *Classification of Latent Variable Models according to the nature of the observed and latent variables (continuous or discrete).*

In the following sections and, in general throughout this Dissertation, the focus is placed on the Factor Analysis approach, as both the observed and latent variables (speaker/language supervectors and the speaker/channel factors) are considered to be continuous.

It is outside the scope of this Dissertation the analysis or development of non-linear latent variables models as Generative Topographic Mapping GTM or Independent Component Analysis (ICA) as those are not proved, up to date, to outperform the linear JFA approach to SV and SLR systems. Nevertheless, interested readers can find good references about those non-linear latent variable models in [Bishop, 2007; Bishop *et al.*, 1997; Comon, 1994; Hyvärinen and Oja, 2000].

3.3.1.2. A formal definition

Given an unknown distribution function $p(\boldsymbol{x})$ of D-dimensional variables, $\boldsymbol{x} = (x_1, ..., x_D)$, belonging to the **data or observed space** $\mathcal{X}$, the goal of a LVM is to express $p(\boldsymbol{x})$ in terms of Q-dimensional variables, $\boldsymbol{z} = (z_1, ..., z_Q)$, where $Q < D$. The space spanned by those hidden or latent variables, $\mathcal{Z}$, is called the **latent space**.

The relation between the latent and the observed space is defined by the conditional distribution $p(\boldsymbol{x} \mid \boldsymbol{z})$ as a mapping function $f : \mathcal{Z} \rightarrow \mathcal{X}$, which takes the form:

$$\boldsymbol{x} = f(\boldsymbol{z}) = y(\boldsymbol{z};\ \Phi) + \boldsymbol{e} \tag{3.6}$$

where the function y express a combination of the latent variables $\boldsymbol{z}$ in terms of a set of parameters Φ and $\boldsymbol{e}$ is a D-dimensional noise variable. Note that geometrically, the space spanned by function y, as a combination of Q-dimensional variables, forms a manifold of rank Q into the D-dimensional observed space $\mathcal{X}$, whereas the noise term $\boldsymbol{e}$ allows to escape from this manifold by covering the whole D-space.

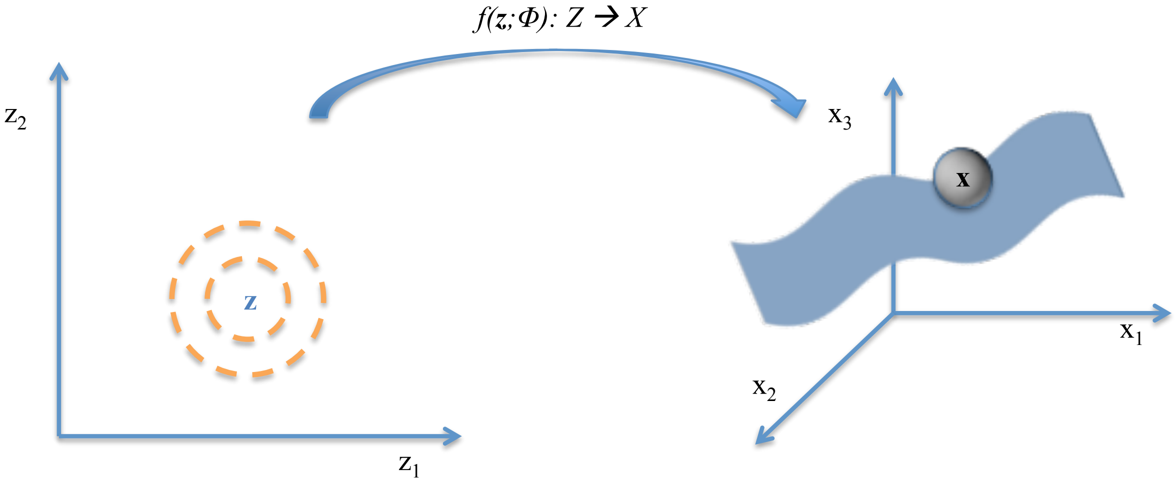

Figure 3.3: *Illustration of a point mapping process from a 2-dimensional latent space to a 3-dimensional observed space. The prior of the latent variables is assumed to be normally distributed $N(O, I)$ (density contours in the left side of the figure). The non-linear mapping function generates a manifold in the observed space where the point is mapped and then modified by a normal distributed noise, also considered $N(O, I)$ (grey sphere).*

In order to complete the model the marginal distribution $p(\boldsymbol{z})$, prior of the latent variables, is also defined. The Figure 3.3 illustrates the whole mapping process from latent to observed variables considering a non-linear mapping function f and normal distributed noise $\boldsymbol{e}$ and prior $p(\boldsymbol{z})$.

From the definition of the above *mapping* scheme, the desired distribution in data space $p(\boldsymbol{x})$ is derived by marginalizing over the latent variables:

$$p(\boldsymbol{x}) = \int_{\mathcal{Z}} p(\boldsymbol{x}, \boldsymbol{z}) d\boldsymbol{z} = \int_{\mathcal{Z}} p(\boldsymbol{x} \mid \boldsymbol{z}) p(\boldsymbol{z}) d\boldsymbol{z} \tag{3.7}$$

This expression is known as the **fundamental equation of latent variables models** [Bartholomew, 1987] and except for specific forms of $p(\boldsymbol{x} \mid \boldsymbol{z})$ and $p(\boldsymbol{z})$ is analytically intractable. To cope with this problem usually normal distributions are considered as they introduce a well-known and friendly framework of practical tractability with respect to the required mathematical manipulation, such as computing equation 3.7 or deriving an EM algorithm to estimate the parameters of the model.

Apart from the issue of the tractability, other reasons can be argued to settle on normal distributions. Among them, if the noise distribution $p(\boldsymbol{e})$ is considered as a sum of a high and unknown number of independent variables with finite variances, the *central limit theorem* endorse also this choice as normal distributed. Also, it has to be taken into account that although the prior in the latent space plays a crucial role in the model, its explicit distribution form does not. In fact, by a simple mapping, it can be easily shown that any prior form could be turned into other before latent variables were translated to the observed space, although the selection of this first mapping could difficult the selection of the main function mapping **f**. In the same line, for practical issues and, regarding continuous domains, the function y is chosen to be *smooth*, that is, it has continuous derivatives up to some desired order over $\mathbb{R}^D$.

3.3.2. Factor analysis

3.3.2.1. Definition

From the above section, it can be readily seen that a LVM is well-defined by three elements:

1. The prior distribution in the latent space $p(\boldsymbol{z})$.
2. The mapping function from the latent to the data or observed space $f : \mathcal{Z} \rightarrow \mathcal{X}$.
3. The noise model in data space $p(\boldsymbol{e})$.

Factor Analysis is differentiated among the other continuous LVMs by supposing:

Factor Analysis: Model definition.

1. **Prior.** The prior in latent space is assumed to be standard **normally distributed**.

$$p(\boldsymbol{z}) \sim N(\mathbf{0}, \boldsymbol{I})$$

2. **Mapping.** The mapping function is considered to be **linear** with form

$$\boldsymbol{x} = f(\boldsymbol{z}) = \boldsymbol{\mu} + \boldsymbol{L}\boldsymbol{z} + \boldsymbol{\epsilon}$$

3. **Noise.** The noise distribution $p(\boldsymbol{e})$ is considered to be also normally distributed with **diagonal** covariance matrix $\boldsymbol{\Psi}$ as

$$\boldsymbol{\epsilon} \sim N(\mathbf{0}, \boldsymbol{\Psi})$$

Table 3.3: *Mathematical model definition of Factor Analysis.*

According to these hypothesis/assumptions, it can be shown (see Appendix A) that both posterior distributions in the observed and latent space are also drawn from normal distributions of the form

$$p(\boldsymbol{x} \mid \boldsymbol{z}) \sim N(\boldsymbol{\mu} + \boldsymbol{L}\boldsymbol{z}, \boldsymbol{\Psi}) \tag{3.8}$$

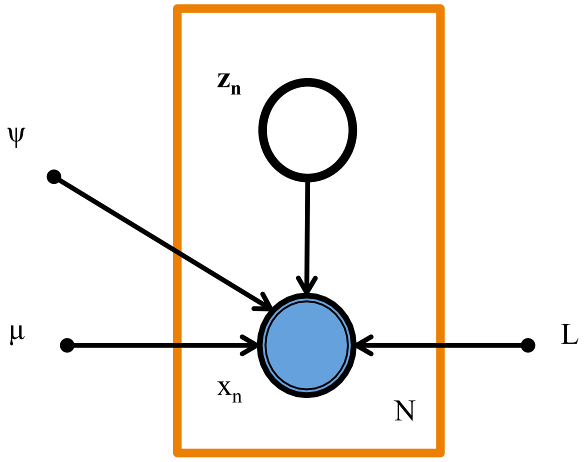

Figure 3.4: *Graphical model representation of Factor Analysis.*

and

$$p(\boldsymbol{z} \mid \boldsymbol{x}) \sim N(A(\boldsymbol{x} - \boldsymbol{\mu}), (\boldsymbol{I} + \boldsymbol{L}^T \boldsymbol{\Psi}^{-1} \boldsymbol{L})^{-1}) \tag{3.9}$$

being A defined as

$$\boldsymbol{A} = \boldsymbol{L}^T (\boldsymbol{L}\boldsymbol{L}^T + \boldsymbol{\Psi})^{-1} = (\boldsymbol{I} + \boldsymbol{L}^T \boldsymbol{\Psi}^{-1} \boldsymbol{L})^{-1} \boldsymbol{L}^T \boldsymbol{\Psi}^{-1} \tag{3.10}$$

where the final algebraic manipulation in equation 3.10 pursues to express A in terms of an inverse matrix in the latent factors domain rather than the observed space for ease of calculation (see Appendix A).

Thus, by analytically solving equation 3.7, the marginal distribution in the observed space is also normal (see Appendix A)

$$p(\boldsymbol{x}) \sim N(\boldsymbol{\mu}, \boldsymbol{L}\boldsymbol{L}^T + \boldsymbol{\Psi}) \tag{3.11}$$

The graphical model representation of FA is depicted in Figure 3.4.

3.3.2.2. Parameter estimation

Apart from the derivations carried out in the above subsection, the only evidence in form of tangible data at our disposal is given by the set of N observed variables $\boldsymbol{X} = \boldsymbol{x}_1, ..., \boldsymbol{x}_N$, that we assume independent and identically distributed. Now, once the model has been properly defined, those data come on the scene to estimate the hyperparameters that define the model $\Theta = \{\boldsymbol{L}, \boldsymbol{\Psi}\}$ [1].

Note that if we knew the latent factors values $\boldsymbol{z}_i$ associated to each observed point $\boldsymbol{x}_i$, the problem of estimating $\boldsymbol{L}$ and then $\boldsymbol{\Psi}$, would turn into a straightforward problem, which might be solved from the defined mapping equation through classical least squares techniques. However,

[1] Note that the parameter $\boldsymbol{\mu}$ is assumed to be set to zero, without loss of generality.

the latent factors are still *hidden* and this fact forces us to develop an estimation procedure able to manage this uncertainty. To this aim, an EM algorithm [Dempster *et al.*, 1977], where the latent factors play the role of the *missing* values, is used.

The EM exploits the fact that to maximize the likelihood of the marginal distribution $p(\boldsymbol{x})$ given the observed data $\boldsymbol{X}$ is equivalent to maximize the expectation, with respect to the posterior distribution $p(\boldsymbol{z} \mid \boldsymbol{x})$, of the joint distribution $p(\boldsymbol{x}, \boldsymbol{z})$ likelihood, whose form is known. In terms of traditional EM as usually stated, the auxiliar function $Q(\Theta, \Theta^{(t)})$ to maximize is then given by

$$Q(\Theta, \Theta^{(t)}) \doteq \sum_{n=1}^{N} \mathbb{E}_{p(\boldsymbol{z}_n|\boldsymbol{x}_n,\Theta^{(t)})}\left[p(\boldsymbol{x}_n, \boldsymbol{z}_n \mid \Theta)\right] \tag{3.12}$$

which is iteratively maximized in function of the current estimate of the hyperparameters $\Theta^{(t)}$

$$\Theta^{(t+1)} = \underset{\Theta}{\operatorname{argmax}}(Q(\Theta, \Theta^{(t)})) \tag{3.13}$$

Specifically, given the set of observed variables $\boldsymbol{X}$ and hyperparameters $\Theta = \{\boldsymbol{L}, \boldsymbol{\Psi}\}$ the joint distribution $p(\boldsymbol{x}, \boldsymbol{z})$, also called the *complete-data likelihood*, $\mathcal{L}c_\Theta$, is given by the expression

$$\mathcal{L}c_\Theta = \prod_n p(\boldsymbol{x}_n, \boldsymbol{z}_n \mid \boldsymbol{L}, \boldsymbol{\Psi}) = \prod_n p(\boldsymbol{x}_n \mid \boldsymbol{z}_n, \boldsymbol{L}, \boldsymbol{\Psi}) p(\boldsymbol{z}_n \mid \boldsymbol{L}, \boldsymbol{\Psi}) \tag{3.14}$$

By taking natural logs - note that the goal is to maximize - this simplifies to:

$$log\mathcal{L}c_\Theta = log\Big\{\prod_n p(\boldsymbol{x}_n, \boldsymbol{z}_n \mid \boldsymbol{L}, \boldsymbol{\Psi})\Big\} = \sum_n log\Big\{p(\boldsymbol{x}_n \mid \boldsymbol{z}_n, \boldsymbol{L}, \boldsymbol{\Psi})\Big\} + \sum_n log\Big\{p(\boldsymbol{z}_n \mid \boldsymbol{L}, \boldsymbol{\Psi})\Big\} \tag{3.15}$$

but since the distribution of the latent variables $\boldsymbol{z}_n$ does not depend on $\boldsymbol{L}$ or $\boldsymbol{\Psi}$, the second term can be discarded for the maximization purpose. Thus, the problem reduces to dealing with the first term. This being

$$\underset{\Theta}{\operatorname{argmax}}(log\mathcal{L}c_\Theta) \equiv \underset{\Theta}{\operatorname{argmax}} \sum_n log\Big\{p(\boldsymbol{x}_n \mid \boldsymbol{z}_n, \boldsymbol{L}, \boldsymbol{\Psi})\Big\} \tag{3.16}$$

Let $\mathcal{L}$ be this simplified likelihood function. The goal is to maximize, as a function of the hyperparameters, its expectation with respect the posterior $p(\boldsymbol{z} \mid \boldsymbol{z})$, $\mathbb{E}_{p(\boldsymbol{z}|\boldsymbol{x})}[\mathcal{L}]$, which by substituting from the conditional distribution $p(\boldsymbol{x} \mid \boldsymbol{z})$ expression, equation 3.8, and some algebra manipulation (see Appendix A) it can be seen that it takes the following form

$$\mathbb{E}_{p(\boldsymbol{z}|\boldsymbol{x})}[\mathcal{L}] = C - \frac{N}{2} ln \mid \boldsymbol{\Psi} \mid - \frac{1}{2} \sum_{i=1}^{N} \{\boldsymbol{x}_i^T \boldsymbol{\Psi}^{-1} \boldsymbol{x}_i - 2\boldsymbol{x}_i^T \boldsymbol{\Psi}^{-1} \boldsymbol{L} \mathbb{E}[\boldsymbol{z}_i \mid \boldsymbol{x}_i] + tr[\boldsymbol{L}^T \boldsymbol{\Psi}^{-1} \boldsymbol{L} \mathbb{E}[\boldsymbol{z}_i \boldsymbol{z}_i^T \mid \boldsymbol{x}_i]]\} \tag{3.17}$$

At this point we are ready to properly use the EM algorithm via its Expectation and Maximization step

- **E-step.** Given current estimation of $\boldsymbol{L}$ and $\boldsymbol{\Psi}$, estimate

$$\mathbb{E}[\boldsymbol{z} \mid \boldsymbol{x}] = \boldsymbol{A}\boldsymbol{z} \tag{3.18}$$

$$\mathbb{E}[\boldsymbol{z}\boldsymbol{z}^T \mid \boldsymbol{x}] = \boldsymbol{I} - \boldsymbol{A}\boldsymbol{L} + \boldsymbol{A}\boldsymbol{x}\boldsymbol{x}^T\boldsymbol{A}^T \tag{3.19}$$

where $\boldsymbol{A} = (\boldsymbol{I} + \boldsymbol{L}^T\boldsymbol{\Psi}^{-1}\boldsymbol{L})^{-1}\boldsymbol{L}^T\boldsymbol{\Psi}^{-1}$ as defined in 3.10

- **M-step.** Estimate new $\boldsymbol{L}$ and $\boldsymbol{\Psi}$ via the following update equations (Appendix A)

$$\frac{\partial \mathbb{E}_{p(\boldsymbol{z}|\boldsymbol{x})}[\mathcal{L}]}{\partial \boldsymbol{L}} = 0 \Rightarrow \boldsymbol{L}^* = \left(\sum_i^N \boldsymbol{x}_i \mathbb{E}[\boldsymbol{z}_i \mid \boldsymbol{x}_i]^T\right)\left(\sum_i^N \mathbb{E}[\boldsymbol{z}_i\boldsymbol{z}_i^T \mid \boldsymbol{x}_i]\right)^{-1} \tag{3.20}$$

$$\frac{\partial \mathbb{E}_{p(\boldsymbol{z}|\boldsymbol{x})}[\mathcal{L}]}{\partial \boldsymbol{\Psi}} = 0 \Rightarrow \boldsymbol{\Psi}^* = \frac{1}{N} diag\left[\sum_i^N \boldsymbol{x}_i\boldsymbol{x}_i^T - \left(\sum_i^N \boldsymbol{x}_i \mathbb{E}[\boldsymbol{z}_i \mid \boldsymbol{x}_i]^T\right)\boldsymbol{L}^T\right] \tag{3.21}$$

These steps are repeated iteratively until convergence, which will depend on several factors, such as the number of latent factors chosen, the quantity of available observed data or the proper initialization of $\boldsymbol{L}$ and $\boldsymbol{\Psi}$. Regarding the latter point, a well-known appropriate initialization for the hyperparametes is to set $\boldsymbol{\Psi} = I$ and $\boldsymbol{L}$ as the result of performing PCA over the dataset $\boldsymbol{X}$ (in order to establish a Q-dimensional, latent factor space, just Q eigenvectors associated to the biggest eigenvalues should be taken into account). This issue will be addressed when applying FA for the SV and SLR purposes.

3.4. Factor Analysis on Gaussian Mixture Models

The FA model described in the above sections refers to a single multivariate Gaussian distribution. This section extends the model to mixtures of multivariate Gaussian models (GMMs), since as it was shown in Chapter 2 (Section 2.4.1), they constitute the base of the state-of-the-art acoustic systems in both SV and SLR acoustic systems.

This section begins defining the sufficient statistics associated to the generation of the observed data points and a GMM model. Then, the MAP adaptation is analysed from a matrix perspective to give some insight about its links with the FA framework applied to SV and SLR. Finally, the focus is placed on how the latent factors of a FA model are estimated within a GMMs framework.

3.4.1. Sufficient statistics

Besides the uncertainty caused by the hidden latent factors, the application of the FA model on GMMs brings also other uncertainty: the *Gaussian occupation alignment*, that is, the mapping between Gaussian mixtures and observed variables: given a feature vector o_t[1], there is not a deterministic way to establish from which Gaussian was generated.

To this aim, the 0th and 1st-order Baum-Welch statistics, hereafter sufficient statistics of the data respect the GMM model (λ_{GMM}) are considered. Those equations although introduced in Section 2.4.1.2 are re-written here for the sake of clarity

$$0th \longrightarrow n_k = \sum_t P_{kt} \tag{3.22}$$

$$1st \longrightarrow \boldsymbol{f}_k = \sum_t P_{kt}\boldsymbol{o}_t \tag{3.23}$$

where P_{kt} is the Gaussian Occupation Probability defined for Gaussian k and feature in time t as

$$P_{kt} = \frac{w_k p_k(\boldsymbol{o}_t)}{\sum\limits_{k=1}^{K} w_k p_k(\boldsymbol{o}_t)} \tag{3.24}$$

being

$$p_k(\boldsymbol{o}_t) = N(\boldsymbol{o}_t \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = (2\pi)^{-\frac{d}{2}} \mid \boldsymbol{\Sigma}_k^{-\frac{1}{2}} \mid \exp(-\frac{1}{2}(\boldsymbol{o}_t - \boldsymbol{\mu}_k)'\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{o}_t - \boldsymbol{\mu}_k)) \tag{3.25}$$

3.4.2. MAP revisited

In order to highlight the similarities versus the classical MAP adaptation procedure introduced in Chapter 2 (Section 2.4.1.2) and FA modelling, it is convenient to derive and re-write at this point the MAP means adaptation equation (equation 2.13) in matrix form, in terms of the UBM mean supervector, $\boldsymbol{\mu}$, as

$$\boldsymbol{\mu}_s = \boldsymbol{\mu} + \boldsymbol{D}\boldsymbol{z} \tag{3.26}$$

where the transformation matrix D is a full rank $CF \times CF$ diagonal matrix defined as

$$\boldsymbol{I} = \tau \boldsymbol{D}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{D} \tag{3.27}$$

being $\boldsymbol{I}$ the $CF \times CF$ identity matrix and $\boldsymbol{\Sigma}$ a $CF \times CF$ diagonal matrix, whose diagonal is formed by the supervector of covariances, that is, by stacking the K diagonal covariances of the UBM model.

[1] Note that in both SV and SLR, the feature vectors or observations $\boldsymbol{o_t}$ play the role of the observed variables x_i used in the formal definition of FA.

It can be readily seen [Gauvain and Lee, 1994], that given $\boldsymbol{\mu}$ and $\boldsymbol{D}$, the MAP criterion, that is, maximizing equation 3.26 in function of $\boldsymbol{z}$, is reduced to solve the system of linear equations $\boldsymbol{A}\boldsymbol{z} = \boldsymbol{b}$, where $\boldsymbol{A}$ and $\boldsymbol{b}$ can be expressed in matrix form as

$$\boldsymbol{A} = \boldsymbol{I} + \boldsymbol{D}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{N} \boldsymbol{D} = \boldsymbol{D}^T \boldsymbol{\Sigma}^{-1} (\tau I + \boldsymbol{N}) \boldsymbol{D} \tag{3.28}$$

$$\boldsymbol{b} = \boldsymbol{N}^T \boldsymbol{\Sigma}^{-1} \overline{\boldsymbol{f}} \tag{3.29}$$

where N $(CF \times CF)$ is a diagonal matrix built as C blocks defined as $\boldsymbol{N}_k = n_k \boldsymbol{I}$ being $\boldsymbol{I}$ the $F \times F$ identity matrix; and $\overline{\boldsymbol{f}}$ the first order statistic supervector built as the concatenation of all $\boldsymbol{f_k}$ centralized by the UBM mean supervector

$$\overline{\boldsymbol{f_k}} = \sum_t P_{kt} (\boldsymbol{o}_t - \boldsymbol{\mu}_k) \tag{3.30}$$

Rewriting now $\boldsymbol{A}\boldsymbol{z} = \boldsymbol{b}$ as:

$$\boldsymbol{D}^T \boldsymbol{\Sigma}^{-1} (\tau \boldsymbol{I} + \boldsymbol{N}) \boldsymbol{D}\boldsymbol{z} = \boldsymbol{D}^T \boldsymbol{\Sigma}^{-1} \overline{\boldsymbol{f}} \tag{3.31}$$

and removing both sides term $\boldsymbol{D}^T \boldsymbol{\Sigma}^{-1}$ this simplifies to

$$\boldsymbol{D}\boldsymbol{z} = (\tau \boldsymbol{I} + \boldsymbol{N})^{-1} \overline{\boldsymbol{f}} \tag{3.32}$$

being $\boldsymbol{D}\boldsymbol{z}$ the *offset* term in equation 3.26
Again, due to the uncertainty of the Gaussian alignment this process can be carried out via an EM algorithm where the sufficient statistics are updated in the E-step and equation 3.32 is applied in the M-step.

The MAP updates equation keep certain similarities respect to the FA modelling, in the sense that some hidden factors $\boldsymbol{z}$ and a transformation matrix $\boldsymbol{D}$ define the model. However in MAP adaptation the transformation matrix $\boldsymbol{D}$ is full rank (diagonal) in the observed space and fixed during the maximization process. Thus $\boldsymbol{D}$ merely acts as a scaling factor of the terms in $\boldsymbol{z}$. Note also, that, as the term $\boldsymbol{D}\boldsymbol{z}$ covers the whole observed space there is no need to include an error/noise term as in the FA model. But, the most important difference lies on the fact that MAP adaptation, unlike FA where subspaces drive the final definition of the new models, does not make use of strong priors (apart from the the UBM model) about the location of speaker or session variability within the supervector space.

3.4.3. Latent factors and hyperparameters estimation

The formulation to estimate the hyperparameters and latent factors that define the FA model using GMMs, follows a similar procedure that this presented in section 3.3.2.2. In fact, from the EM algorithm, just the E-step must be slightly modified by adequately including the sufficient statistics, as defined in above sections, instead of directly include observed data.

The new equations are now:

- **E-Step:**

$$\mathbb{E}[\boldsymbol{z} \mid \boldsymbol{x}] = \boldsymbol{\Omega}\overline{\boldsymbol{f}} \tag{3.33}$$

$$\mathbb{E}[\boldsymbol{z}\boldsymbol{z}^T \mid \boldsymbol{x}] = \boldsymbol{I} - \boldsymbol{\Omega}\boldsymbol{L} + \boldsymbol{\Omega}\overline{\boldsymbol{f}}\,\overline{\boldsymbol{f}}^T\boldsymbol{\Omega}^T \tag{3.34}$$

where, $\boldsymbol{\Omega} = (\boldsymbol{I} + \boldsymbol{L}^T\boldsymbol{N}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Psi}^{-1}\boldsymbol{L})^{-1}\boldsymbol{L}^T\boldsymbol{\Psi}^{-1}$

- **M-step:**

$$\frac{\partial\mathbb{E}_{p(\boldsymbol{z}|\boldsymbol{x})}[\mathcal{L}]}{\partial\boldsymbol{L}} = 0 \Rightarrow \boldsymbol{L}^* = \left(\sum_i^N \boldsymbol{x}_i\mathbb{E}[\boldsymbol{z}_i \mid \boldsymbol{x}_i]^T\right)\left(\sum_i^N \mathbb{E}[\boldsymbol{z}_i\boldsymbol{z}_i^T \mid \boldsymbol{x}_i]\right)^{-1} \tag{3.35}$$

$$\frac{\partial\mathbb{E}_{p(\boldsymbol{z}|\boldsymbol{x})}[\mathcal{L}]}{\partial\boldsymbol{\Psi}} = 0 \Rightarrow \boldsymbol{\Psi}^* = \frac{1}{N}diag\left[\sum_i^N \boldsymbol{x}_i\boldsymbol{x}_i^T - \left(\sum_i^N \boldsymbol{x}_i\mathbb{E}[\boldsymbol{z}_i \mid \boldsymbol{x}_i]^T\right)\boldsymbol{L}^T\right] \tag{3.36}$$

Setting aside the issue of hyperparameters initialization that will be addressed in the following section, and apart from some simplifications, modifications or alternatives presented in following chapters, it is worth at this point to identify the SV JFA parameters with their values and common estimation procedures. This information is shown in Table 3.4.

$$\boldsymbol{\mu}_{sh} = \boldsymbol{\mu} + \boldsymbol{V}\boldsymbol{y}_s + \boldsymbol{D}\boldsymbol{z}_s + \boldsymbol{U}\boldsymbol{x}_{sh}.$$

Term	*Set/Estimated to/as.*
$\boldsymbol{\mu}$	Set to UBM speaker means supervector.
$\boldsymbol{V}$	EM, updated via equation 3.35., with $\boldsymbol{y}_s$ as latent factors
$\boldsymbol{U}$	EM, updated via equation 3.35., with $\boldsymbol{x}_{sh}$ as latent factors
$\boldsymbol{y}_s$	Point estimated via equation 3.33., considering $\boldsymbol{V}$ as subspace
$\boldsymbol{x}_{sh}$	Point estimated via equation 3.33., considering $\boldsymbol{U}$ as subspace
$\boldsymbol{D}\boldsymbol{z}_s$	Estimated via equation 3.32.

Table 3.4: *Summary of JFA model parameters estimation.*

Note that the terms $\boldsymbol{D}$ and $\boldsymbol{z}_s$ has been considered grouped to be classically assigned to the offset derived of a MAP estimation. In the following chapter, it will be evaluated the importance of doing a separate estimation as it is done for the other paired terms.

Note also that the covariance $\boldsymbol{\Psi}$ and its update equation 3.36 are not considered in Table 3.4. This responds to the fact that, for the sake of simplicity, the covariance of the JFA model is considered to be fixed and equal to the UBM covariance.

3.5. Factor Analysis on Speaker/Language Recognition

3.5.1. Joint versus disjoint estimation

Before starting the hyperparameters estimation procedure, a preliminary decision has to be taken concerning the order in which they are generated. Whether estimating the speaker $\boldsymbol{V}$ or session variability $\boldsymbol{U}$ subspaces before the other or at the same time must be carefully decided.

When JFA was introduced a simultaneous optimization of both subspaces was proposed [Kenny *et al.*, 2005a; Kenny and Dumouchel, 2004b]. This can be achieved via an EM algorithm, which alternatively maximizes the model respect one of the subspaces while keeping fixed the other. However this procedure has a major drawback: there is no way to explicitly constraint the session variability subspace to capture only relevant session information or the speaker variability to capture just speaker information. In that sense, the EM algorithm is blind and it will fit the data as best as it can. Further, if for instance, the speaker variability subspace is considered to have a greater number of degrees of freedom than the session variability subspace, after some iterations this will likely dominate the maximization procedure to the detriment of the session variability subspace; as a result the procedure will end with *contaminated* subspaces.

In order to avoid this fact as well as simplifying the estimation procedure, several variants of the estimation procedure can be accomplished. Some of this variants are analysed in [Vogt *et al.*, 2008b], where a hybrid approach between the simultaneous and isolated (training both separately) estimation of the subspaces is proposed as good trade-off between performance and computational requirements. In this hybrid approach $\boldsymbol{V}$ is trained separately but considering a pre-trained $\boldsymbol{U}$, taking therefore into account the session variability during its estimation.

3.5.2. Initialization of variability subspaces

In order to complete the description of the Joint Factor Analysis approach applied to SV and SLR systems, it is convenient to give the corresponding details of the initialization of the variability subspaces. Even though, in theory the ML procedure could account much part of the work and save the need of a *smart* initialization, its convergence could be strongly affected by a dummy election, greatly slowing the estimation process.

As in the case of the eigenfaces/eigenvoices approaches, an analysis of the variation among means supervectors belonging from different speakers provides a good starting point for the ML procedure. In order to do this PCA fits with the problem. The type of variance analysed will mark the difference between each subspace.

3.5.2.1. Session variability subspace

When creating the session variability subspace, the main interest is to retain differences between utterances belonging or not to same speakers but avoiding, as much as possible, the components produced by the speaker information.

Let $\boldsymbol{X} = \boldsymbol{\mu}_1, ...\boldsymbol{\mu}_N$ a set of mean supervectors belonging to C different speakers ($C < N$), a good estimation of the session variability variance is given by the within-class scatter matrix of $\boldsymbol{X}$, where the mean of every speaker is subtracted from its corresponding utterances:

$$\boldsymbol{S_w} = \sum_c \sum_{i \in c} (\boldsymbol{\mu}_i - \boldsymbol{\mu}_c)(\boldsymbol{\mu}_i - \boldsymbol{\mu}_c)^T \tag{3.37}$$

Therefore, the principal component analysis of $\boldsymbol{S_w}$ will serve us a good starting point as the session variability subspace estimation.

Nevertheless, two practical issues should take into account when performing the principal component analysis of $\boldsymbol{S_w}$:

1. How many directions (principal components) should be kept?.
2. How can we avoid a computationally prohibitive process?.

How big is the subspace that should be estimated? Or how many principal components should be kept?, are indeed the same question formulated under a different perspective. In principle, when estimating the subspace we do not know its size. A possible solution would be to apply a Bayesian approach [Bishop, 2007] into the estimation procedure with respect to the size of the subspace. Thus, the size will be treated as other unknown parameter and it could be estimated besides the other hyperparameters. Other alternative, which can be easily embedded into the ML presented framework, is to simply inspect the values of the eigenvalues associated to the principal components, discarding those which do not accumulate variance, that is, those whose associated eigenvalues are zero or nearby zero.

The computational issue involves some algebra manipulation. Handling a within-class scatter matrix of a large dimension can be computationally prohibitive. For instance, in a typical GMM framework for SV, at least 1024 Gaussian and 38 dimensions are managed. The corresponding within-scatter matrix to this system is therefore an enormous matrix of 38912×38912 dimensions ($\backsim$ 5.6 GB in float/single precision). Performing and eigen-analysis of this large matrix can turn out in a never-ending task.

Fortunately, the number of supervectors N of the set $\boldsymbol{X}$ that we start considering, uses to be much smaller than the supervector dimensionality CF. This fact encourages the use of the following theorem to reduce the size of the problem:

Given a $N \times M$ matrix $\boldsymbol{A}$ that can be decomposed in the form $\boldsymbol{A} = \boldsymbol{\Phi\Phi^T}$, then [Fukunaga, 1990]:

$$eig(\boldsymbol{A}) = eig(\boldsymbol{\Phi\Phi}^T) = \boldsymbol{\Phi} eig(\boldsymbol{\Phi}^T\boldsymbol{\Phi}) \tag{3.38}$$

where the $eig(\boldsymbol{.})$ operator represents the eigen-analysis function [1]. Given that $\boldsymbol{S_w}$ can be easily decomposed in the form $\boldsymbol{S}_w = \boldsymbol{\Phi\Phi}^T$ where $\boldsymbol{\Phi}$ is defined as $\boldsymbol{\Phi} = (\boldsymbol{\mu}_1 - \overline{\boldsymbol{\mu}}, ..., \boldsymbol{\mu}_n - \overline{\boldsymbol{\mu}})$, the problem is reduced to perform the eigen-analysis of a $M \times M$ matrix rather than a $CF \times CF$ one.

[1] The eigenvalues must be rescaled to fit the equality, see [Fukunaga, 1990].

To give some insight about how this reduction is possible without loss of generality note that the rank of $\boldsymbol{S_w}$ is at most M as follows of the axioms:

- if $\boldsymbol{A}$ is an $N \times M$ matrix $\Rightarrow rank(\boldsymbol{A}) \leqslant min(N, M)$
- $rank(\boldsymbol{AB}) \leqslant min(rank\boldsymbol{A}, rank\boldsymbol{B})$

However, although the problem has been significantly reduced, that might not be enough, as typically the number of samples could be also high ($\backsim$ 10k), so that handling a very large matrix is still needed. To overcome this problem, it is convenient to realize that a complete eigen-decomposition of the $\boldsymbol{S_w}$ matrix is not necessary. Indeed, the final goal is to yield a subspace, that is, a low-rank matrix formed by a few columns or principal directions. From the 10k possible eigenvectors that can be extracted just a few tens/hundreds are enough, as they accumulate most of the $\boldsymbol{S_w}$ variance. This can be solved by iterative eigen-decomposition algorithms [Arnoldi, 1951; Lanczos, 1950], where just the required eigenvectors are iteratively approximated without the need of computing all of them [1]. In the exprimental part of this Dissertation it will be shown how these methods significantly diminished the computational constraints of the real SV and SLR systems.

3.5.2.2. Speaker variability subspace

When creating the speaker variability subspace, the main interest is to model the differences between utterances of a wide range of speakers rather than avoiding the distortions produced by the session variability. In that case, the variance expressed by the between-class scatter matrix $\boldsymbol{S_b}$ is a good estimation of the speaker variability, as each speaker is represented by the average of all its utterances. Thus, defining:

$$\boldsymbol{S_b} = \sum_c (\boldsymbol{\mu}_c - \overline{\boldsymbol{\mu}})(\boldsymbol{\mu}_c - \overline{\boldsymbol{\mu}})^T \tag{3.39}$$

Then, the procedure to obtain a first estimation of the speaker variability subspace will be the same as that described above for the session variability subspace but replacing $\boldsymbol{S_w}$ by $\boldsymbol{S_b}$. The same algebra derivations can be used to achieve good computational performance.

Note that, however, in this case, the eigen-decomposition is not as costly as that described for $\boldsymbol{S_w}$, as the utterances are grouped by speakers and therefore the dimensions of the resulted $\boldsymbol{\Phi}^T\boldsymbol{\Phi}$ matrix $S \times S$, where S is the number of speakers, is much smaller.

3.6. Summary

In this chapter a review of the Joint Factor Analysis grounds has been performed, giving a timeline covering the most important milestones from the use of subspaces to model variability

[1] An implementation of the Arnoldi's routines can be found through the ARPACK library, both in Fortran and C++ [Lehoucq *et al.*, 1997; Sorensen and Gomes, 1997]. Also, the MATLAB function *eigs* includes the Arnoldi algorithm by linking with ARPACK.

in related fields to its apparition as a new modelling manner of handling variability. The basis of Joint Factor Analysis, to consider the variability as a continuous source and to make use of priors in form of variability subspaces has been also exposed.

Further, the theory behind the mathematical model Factor Analysis has been extensively documented, with details of the training of both hyperparameters and latent factors associated. This theory has been extended to its use in mixture of Gaussian densities to fit with the inclusion of Joint Factor Analysis into the well known GMM-UBM framework for SV or SLR purposes.

Chapter 4

Factor Analysis applied to SV and SLR systems: PART I (algorithmics)

THIS CHAPTER PRESENTS *where* and *how* JFA is integrated into SV and SLR, analysing and discussing its multiple forms to yield robust and efficient acoustic systems.

4.1. Introduction

Systems using FA gained prevalence due to their enhanced ability to deal with complex sources of speaker/language inter-session variation, being nowadays present in the most successful text-independent SV and SLR systems.

The explosion in its use from the beginnings of 2004 to date has derived in a high degree of variants in its forms of application. Those variants arose to meet different needs, which could be categorized in the following four groups:

1. **Integrating Factor Analysis into the diverse existing state-of-the-art systems**. The success of FA approaches soon demanded its incorporation to the different existing state-of-the-art systems. This chapter explores and details this integration into the SV and SLR focused on the acoustic GMM and SVM systems.

2. **Integrating Factor Analysis into different levels/domains of the recognition scheme** (feature, model and statistic domain). As other inter-session variability approaches, the global framework of FA has been adapted to be applied at different levels of the recognition process. Specifically, regarding acoustic systems, FA has been applied in three different levels, namely, the model, feature and statistics domain, understood the latter as an intermediate level between feature and model domain, where measurements are the sufficient statistics extracted from the recordings and a reference model.

3. **Natural evolution of Factor Analysis approaches to better fit with the speaker and spoken language recognition tasks.** From the beginnings of its application,

different improved versions of the global FA scheme has been developed to better fit with the specific problems of speaker or spoken language recognition. In this chapter, those steps given towards an evolved FA version are described.

4. **Achieving a proper trade-off between recognition and computational efficiency.** A major reason stymieing the deployment of a fully-based FA model system is that when dealing with large size problems, its implementation tends to be prohibitive. To counteract this deficiency, several simplifications to the JFA procedure have been developed to relax the computational constraints with low cost in terms of verification rates.

The remainder of this chapter is organized as follows. First, a detailed review of the integration of FA in both GMM and SVM acoustic systems at the aforementioned three levels/domains (model, feature and statistics) is presented. Then, the linear scoring approach as an efficient alternative to classical scoring is described. Finally, all the pieces are assembled to design and present efficient forms to build JFA acoustic systems for both SV and SLR.

Original contributions of this chapter includes the adaptation of FA in the statistics domain of a SVM system for SLR as well as the development of competitive and efficient systems in both SV and SLR presented at NIST speaker and language recognition evaluations (SRE and LRE) from 2006 to 2010 (SRE06, SRE08, SRE10 and LRE07, LRE09).

4.2. FA: Where and How

4.2.1. FA in the model domain

JFA was initially conceived to be integrated within the well-known GMM-UBM framework for speaker verification [Reynolds *et al.*, 2000]. The proposed scheme included a FA modelling of the enrolment target models rather than use MAP adaptation, acting therefore in the model domain.

4.2.1.1. FA in GMM

A. The Original Recipe

Initial works conducted by Patrick Kenny, [Kenny and Dumouchel, 2004a,b], proposed a general JFA *recipe* to yield an integration into a GMM system. That *recipe* can be synthesized in five steps as detailed in pseudo-code in Table 4.1.

Note that in this original *recipe* the hyperparameters $\boldsymbol{V}$, $\boldsymbol{D}$ are considered speaker-dependent. As it will be shown later on during the course of this chapter, this assumption, although well justified from a theoretical point of view, was soon relaxed in more efficient versions of JFA to finally remain independent of the speaker models. A major reason for this fact is that in most of situations the training material for a specific speaker is not enough to actually introduce con-

Original JFA integration within classical GMM-UBM framework

```
I   Train an Universal Background Model, $\lambda_{ubm} = \{\omega_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^{K}$

    1. $\boldsymbol{O}_{dev} := observations(devData);$
    2. $\lambda_{ubm} := clustering(\boldsymbol{O}_{dev});$            % K-Means or Binary Splitting.
    3. $\lambda^*_{ubm} := EM_{ML}(\boldsymbol{O}_{dev});$            % Maximum Likelihood via EM iterations.

II  Initialization of Hyperparameters, $\Lambda = \{\boldsymbol{\mu}, \boldsymbol{V}, \boldsymbol{U}\}$ (section 3.5.1)

    1. $\boldsymbol{\mu} := \boldsymbol{\mu}_{ubm}$
    2. for each utterance i $\boldsymbol{O}_{dev_i}$ in dev set $\boldsymbol{O}_{dev}$:
           $\lambda_i := EM_{MAP}(\boldsymbol{O}_{dev_i}, \lambda^*_{ubm});$    % training model via MAP adaptation
           $\boldsymbol{X}(:,i) := \boldsymbol{\mu}^i;$                   % stacking mean supervector in column form
       end
    3. $\boldsymbol{S}_b := betweenScatterMatrix(\boldsymbol{X});$ $\boldsymbol{S}_{\boldsymbol{w}} := withinScatterMatrix(\boldsymbol{X});$
    4. $\boldsymbol{V} := PCA(\boldsymbol{S}_b);$
    5. $\boldsymbol{U} := PCA(\boldsymbol{S}_w);$

III Hyperparameters Refinement, $\Lambda = \{\boldsymbol{V}, \boldsymbol{U}, \boldsymbol{D}\}$ (section 3.4.3)

    1. $\boldsymbol{O}_{dev2} := observations(devData2);$
    2. $\Lambda^* := maximize_{\Lambda}(\boldsymbol{O}_{dev2}, \Lambda);$

IV  Train target models, $\Lambda_s = \{\boldsymbol{V}_s, \boldsymbol{D}_s, \boldsymbol{y}_s, \boldsymbol{z}_s\}$ (section 3.4.3)

    1. $\boldsymbol{O}_{train} := observations(trainData);$
    2. for each speaker s $\boldsymbol{O}_{train_s}$ in train set $\boldsymbol{O}_{train}$:
           $\Lambda_s := maximize_{\Lambda}(\boldsymbol{O}_{train_s}, \Lambda^*);$
       end

V   Testing

    1. $\boldsymbol{O}_{test} := observations(testData);$
    2. for each speaker j with observations $\boldsymbol{O}_{test_j}$ in train set $\boldsymbol{O}_{test}$:
           for each model $\lambda_{sh}$ defined by $\Lambda_s$:
               $score_{\lambda_{sh},j} := \frac{l(\Lambda_s, \boldsymbol{O}_{test_j})}{l(\Lambda, \boldsymbol{O}_{test_j})};$
           end
       end
```

Table 4.1: *Original Joint Factor Analysis integration within classical GMM-UBM framework.*

siderable modifications in them, being preferable to account all the discriminating information into their latent factors associated ($\boldsymbol{y}_s$ and $\boldsymbol{z}_s$).

Other interesting point to highlight concerns the scoring approach, which is proposed as a classical log-likelihood ratio between the target model and the Universal Background Model - UBM -. The difference with respect to the classical scoring lies in the form of the likelihood function, where channel factors of the testing utterance are considered and integrated out to account all their possible values. The likelihood function for a test observations $\boldsymbol{O}$ of a recording h faced to a model of speaker s with hyperparameters Λ_s is expressed as

$$P(\boldsymbol{O} \mid \Lambda_s) = \int P(\boldsymbol{O} \mid \Lambda_s, \boldsymbol{x}_{sh}) N(\boldsymbol{x}_{sh} \mid \boldsymbol{0}, \boldsymbol{I}) d\boldsymbol{x} \tag{4.1}$$

Hence, the session variability encountered in the testing recording h is taken into account. To evaluate the final score the EM auxiliary function in equation 3.17, as a lower bound of the expression 4.1, is used taking into account that some of the terms are cancelled as they are identical for the target model and the UBM model. A detailed derivation of this scoring approach can be found in [Kenny *et al.*, 2007].

B. Simplifications

The proposed scheme, detailed in the above section, translates the complete mathematical background described in the previous Chapter 3 into a SV GMM based system. However, despite of its promising gains supported by a strong theoretical framework, this approach was not immediately adopted by the scientific community. A major reason for this was the lack of a large corpora able to adequately exploit the advantages of FA including strong priors of speaker and session variability. Also, implementing the *recipe* step by step demanded a high cost in terms of computational resources, so the balance between the computational cost and the system performance was not at that moment as attractive as it is nowadays.

To counteract these drawbacks, soon, several modifications/simplifications were proposed to cope with these two main difficulties. Those simplifications were first focused on simplifying the model and second on finding shortcuts in its development, which speed up the process under an acceptably low loss of performance.

Among the several proposed simplifications found in the literature, it is convenient to rescue by its posterior relevance, the following ones:

- **Compute subspaces $\boldsymbol{U}$ and $\boldsymbol{V}$ in a disjoint manner rather than simultaneously** [Kenny *et al.*, 2005b]. As it was previously stated, this simplifications was one of the first performed. Training $\boldsymbol{U}$ and $\boldsymbol{V}$ subspaces separately allows the use of similar and more simplified procedures to train them up. Further, session variability information can be modelled and suppressed, as it will be shown, before training the speaker variability subspace.

- **Consider the hyperparameters $\boldsymbol{V}$ and $\boldsymbol{D}$ independent of the target speaker model** [Vogt and Sridharan, 2008]. Although, theoretically, $\boldsymbol{V}$ and $\boldsymbol{D}$ should depend on the target speaker, in practical situations there is not enough training data for the speaker to get a significant improvement by doing this adaptation. Keeping fixed $\boldsymbol{V}$ and $\boldsymbol{D}$ led to significant improvements in computational terms allowing the use of a fixed set of hyperparameters irrespective of the speaker treated.

- **Consider the channel factors to be independent of the target speaker model** [Vair *et al.*, 2006]. By loosening the speaker-dependence constraint the channel factors can be computed through the sufficient statistics associated to the UBM model and the session variability subspace. This fact greatly simplify the test verification process, as for each test utterance just a single estimation of the channel factors is required rather than one for each target model.

All these fundamental optimizations in conjunction with the advances carried out in the verification stage, treated in section 4.2, made JFA a viable tool to be efficiently integrated into SV and SLR systems. Further, as it will be shown in the experimental part of this Dissertation, thanks to the laudable efforts conducted by different institutions as NIST (National Institute of Standards and Technology) to acquire larger and more complete databases [NIST, 2010], the power of FA to deal with speaker and session variability has been largely proved.

C. The Speaker Variability and The Residual Term Dz

For the sake of simplicity and due to the great results achieved just by modelling the session variability via FA, the speaker variability term, as conceived in the original model, was often left aside and was, in several works, set as the offset derived of a MAP adaptation. That is, the JFA original speaker component offset represented by terms $\boldsymbol{Vy}+\boldsymbol{Dz}$ was synthesized into the $\boldsymbol{Dz}$ term. Hence, the speaker subspace disappeared and with it, many of the computational resources requirements.

However, from a theoretical point of view, it remained clear that an adequate inclusion of a speaker variability prior could lead to further improvements since the profits of both classical MAP and eigenvoice MAP could be jointly accounted as it was discussed in the previous chapter. However, first experiments in that sense did not achieve significant improvements by including both terms, $\boldsymbol{Vy}$ and $\boldsymbol{Dz}$ [Kenny *et al.*, 2008a].

Fortunately, the work conducted by Kenny *et al.* [2008b] detected the cause of this conflict between theory and practice, as a non-proper estimation of the speaker component. As stated in [Kenny *et al.*, 2008b] one of the prime reasons that led to this non-proper estimation of the hyperparameters $\boldsymbol{D}$ and $\boldsymbol{V}$ in previous studies was the fact that both had been trained in a joint manner via a maximum likelihood (ML) procedure. Considering $\boldsymbol{D}$ diagonal and $\boldsymbol{V}$ composed by say 300 eigenvoices, the number of free parameters to estimate in $\boldsymbol{V}$ is 300 times the ones in $\boldsymbol{D}$. In this context, it is not surprising that the ML procedure devoted more of its attention to estimate $\boldsymbol{V}$ at the expense of $\boldsymbol{D}$, which elements results to be close to zero.

To palliate this undesired behaviour, in [Kenny *et al.*, 2008b], a disjoint training procedure was proposed with the aim of properly estimate both terms. In that sense, once initial eigenvoice models are computed, the term $\boldsymbol{\mu}+\boldsymbol{V}\boldsymbol{y}$ can be used to centralize the statistics used to train $\boldsymbol{D}$. Thus, the ML procedure aims to modelling the speaker variability not present in the subspace $\boldsymbol{V}$ resulting in a major benefit in terms of discrimination and therefore revealing the importance of modelling the speaker variability component $\boldsymbol{D}\boldsymbol{z}$ under the FA theoretical framework.

4.2.1.2. FA with SVM

There are several ways to incorporate FA into the model domain of a SVM system. One of the most obvious could be just compensating GMMs means supervector by using the standard FA framework, and then make use of those compensated models to feed a SVM-SV system as described in Section 2.4.3. This solution, although easy to perform, has the main drawback that both training and test utterances must be modelled to get the mean supervector.

On the other hand, in a parallel way to the development of Factor Analysis, a new session variability technique, coined Nuisance Attribute Projection (NAP) [Solomonoff *et al.*, 2004], was designed to be integrated into a SVM system. NAP is based on projecting away the non-desired (session variability) directions/components by including a projection operator $\boldsymbol{P}$ into the kernel operation as follows

$$\begin{aligned} k(\boldsymbol{x},\boldsymbol{y}) = \phi(\boldsymbol{x}) \centerdot \phi(\boldsymbol{y}) &= [\boldsymbol{P}\phi(\boldsymbol{x})]^T[\boldsymbol{P}\phi(\boldsymbol{y})] \\ &= \phi(\boldsymbol{x})^T\boldsymbol{P}\phi(\boldsymbol{y}) \end{aligned} \tag{4.2}$$

where $\phi(\cdot)$ is an expansion function from the feature space to the high-dimensionality space, and P is the projection operator defined as $\boldsymbol{P}=\boldsymbol{I}-\boldsymbol{Z}\boldsymbol{Z}^T$, being $\boldsymbol{Z}$ estimated as:

$$\underset{Z}{\text{argmin}} = \sum_{i,j} \boldsymbol{W}_{i,j} \parallel \boldsymbol{P}\phi(\boldsymbol{x}_i) - \boldsymbol{P}\phi(\boldsymbol{x}_j) \parallel_2^2 \tag{4.3}$$

being $\boldsymbol{W}_{i,j}$ the labels matrix, so that $\boldsymbol{W}_{i,j}=1$ if $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ belong to the same speaker and $\boldsymbol{W}_{i,j}=0$ otherwise. The $\boldsymbol{Z}$ matrix is found to be orthonormal ($\boldsymbol{Z}^T\boldsymbol{Z}=\boldsymbol{I}$) such as $\boldsymbol{P}$ can be defined formally as a projection ($\boldsymbol{P}=\boldsymbol{P}^2$).

The NAP approach bears many similarities to FA, in the sense that both attempts to compensate session variability by establishing a strong prior of the variability represented by a low-dimensional subspace. In fact, as demonstrated in [Campbell *et al.*, 2006c], by using a linear kernel where the expansion function transforms each recording (defined by the observations $\boldsymbol{O}$) to its means supervector $\Phi(\boldsymbol{O})=\boldsymbol{\mu}$, the session variability subspace computed via NAP is identical to that computed via FA ($\boldsymbol{Z}=\boldsymbol{U}$).

The main differences with the complete FA formulation are that in NAP i) a removing process rather than modelling process is carried out regarding the session variability; and ii) there is not an analogous term for modelling speaker variability by including a prior via a low-dimensional subspace as in FA.

4.2.2. FA in the feature domain

The specific integration of FA within the model domain suffer from two shortcomings. First, the need to formulate specific forms to take into account the session variability present on the test utterance, as it is the case in the GMM approach; and second and most important the lack of flexibility to directly extend the model to other modelling approaches or tasks. These reasons motivated the search of new approaches to integrate FA within the feature domain. The underlying idea is clear, once the feature vectors are *clean* of session variability, whatever kind of modelling should be benefited without performing additional modifications.

4.2.2.1. FA in GMM

In this direction and applied into a classical GMM-UBM framework, the work conducted by Vair *et al.* [2006] presented and elegant form to perform session variability compensation within the feature domain. The technique strongly inspired in the Feature Mapping approach [Reynolds, 2003], presented in Chapter 2 (Section 2.6.2.4), proposed a frame-by-frame feature compensation in the following form

$$\hat{\boldsymbol{o}}_t{}^{(h)} = \boldsymbol{o}_t^{(h)} - \sum_k P_{kt} \boldsymbol{U}_k \boldsymbol{x}_h \tag{4.4}$$

where $\boldsymbol{o}_t^{(h)}$ is the t frame of a utterance h, P_{kt} is the *Gaussian occupation probability* for Gaussian k and frame t, defined as in equation 3.24 and $\boldsymbol{U}_k$ is the submatrix of the session variability subspace corresponding to Gaussian k (that is, rows from $(k-1)*F+1$ to kF, being F the feature vector dimension).

Thus, the corresponding session variability to each frame is directly subtracted frame by frame supported by the prior subspace $\boldsymbol{U}$ being the channel factors $\boldsymbol{x}$ estimated for the utterance h. To alleviate a bit the costly operation, the sum in k use to be constrained to the five most likely Gaussian for the frame t (top-5 Gaussian for frame t).

Note, that this compensation it is possible since the channel factors are considered to be only dependent of the utterance rather than the utterance and the speaker model, otherwise, the reference model would remain tied to the compensated utterance. In order to avoid this issue, the UBM is considered as the reference model, to compute sufficient statistics, channel factors, and also Gaussian occupations probabilities.

4.2.2.2. FA in SVM

As an inherent property/advantage of the feature domain compensation, there is nothing additionally to do to extend the compensation to other modelling approach. The compensated features can now fed a SVM system without the need to be modified.

A very similar approach can be also derived from NAP to be applied at feature domain, by means of the so-called *feature NAP* (fNAP) [Campbell *et al.*, 2008], which can be derived by rewriting equation 4.5 as:

$$\hat{\boldsymbol{o}_t}^{(h)} = \boldsymbol{o}_t^{(h)} - \sum_k P_{kt} \boldsymbol{n}_{hk} \tag{4.5}$$

where here, $\boldsymbol{n}_h$ is the nuisance means supervector derived from the NAP projection of the mean supervector associated to the utterance h; and index k refers to the sub-vector corresponding to Gaussian k.

4.2.3. FA in the statistics domain

A hybrid version between performing the compensation in model and feature domain, is doing it in the statistics domain, where the term *statistic* shortly refers to the sufficient statistics extracted from the data material and a reference model (normally, the UBM).

This approach addresses much of the advantages presented by the model and feature domain compensation, whilst avoiding their principal drawbacks. As it will be shown, i) it eliminates the costly frame-by-frame compensation and ii) it allows an easy integration within a non-modified SVM system. Further, it beautifully links with the linear scoring technique presented later on in the Section 4.2, leading to accurate and efficient SV and SLR systems.

4.2.3.1. FA in GMM

As it was stated before, it is desirable to apply the compensation in a stage before the model domain, as this would allow applying the compensation directly to the test features extracted from data without the need to create a model or finding specific forms to it. The work conducted by Brümmer *et al.* [2009], accomplishes the compensation at the statistics domain inspired in the above presented scheme where session variability compensation was performed in the feature domain.

This feature domain compensation idea can be reused in the statistics domain in order to get a session-variability-compensated first-order statistic $\overline{\boldsymbol{f}}_c$, following the next form

$$\overline{\boldsymbol{f}_c} = \overline{\boldsymbol{f}} - \boldsymbol{N}\boldsymbol{U}\boldsymbol{x} \tag{4.6}$$

where $\boldsymbol{N}$ and $\overline{\boldsymbol{f}}$ are the zero order and first order centralized statistics in matrix form defined as in Section 3.4.2 and $\boldsymbol{x}$ the corresponding channel factors estimated for the given statistics and session variability subspace $\boldsymbol{U}$.

This approach has the desirable property of avoiding the need of a computationally expensive *frame by frame* compensation whilst allowing an easy integration into a SVM system as it presented in the following section.

		Advantages	Disadvantages
Compensation Domain	*Feature*	1. Extensible to others modelling approaches 2. Symmetric application on both training and testing utterances	1. Frame-by-Frame Compensation 2. Does not include $\boldsymbol{Vy}$ and $\boldsymbol{Dz}$ terms
	Statistics	1. Extensible between SVM and GMM models 2. Symmetric application on both training and testing utterances 3. Avoids frame-by-frame compensation	1. Non-extensible to all others modelling approaches
	Model	1. Includes the complete JFA model 2. Avoids frame-by-frame compensation	1. Dependence on the modelling approach 2. Non-symmetric compensation on models/testing utterances

Table 4.2: *General main advantages/disadvantages of applying FA in the different domains of an acoustic SV or SLR system.*

4.2.3.2. FA in SVM

A modification to the work in [Campbell *et al.*, 2006b] was introduced in [Gonzalez-Dominguez *et al.*, 2010d] by employing a session variability compensation scheme within the statistics domain, which utilise the channel compensated first-order statistics derived from a GMM system. Thus, a single MAP adaptation is needed in order to obtain compensated GMM supervectors.

The proposed scheme has fundamental advantages over the past described methods. On the one hand, although session variability compensation techniques applied to the feature domain such as feature Nuissance Attribute Projection (fNAP) [Campbell *et al.*, 2008] or feature Latent Factor Analysis (fLFA)[Castaldo *et al.*, 2007][Campbell *et al.*, 2008] have the prime advantage of allowing any type of posterior modeling, its application implies a frame-by-frame compensation over the set of features rather than a single compensation in model or statistics domain. This becomes a major drawback when large amounts of data must be processed, as in language recognition. On the other hand, once first-order statistics are channel compensated, no other FA techniques applied at model domain such as [Matrouf *et al.*, 2007] or NAP [Solomonoff *et al.*, 2005] are necessary. This turned out in a major saving of computational time in acoustic systems as well as significant benefits in terms of verification rates.

4.3. The Linear Scoring Approach

There are several scoring techniques associated to FA [Glembek *et al.*, 2009]. Among them, one which deserves special attention is the *linear scoring approach* [Brümmer *et al.*, 2009], which by means of an elegant derivation turns the costly scoring stage into a single dot product without significant loss of classification performance. The remainder of this section is devoted to derive the linear scoring approach from the classical GMM scoring method presented in Section 2.4.1.3.

Classic scoring is presented as a ratio between the likelihood of the dataset $\boldsymbol{O}$ of the target model for the speaker s, λ_s, and the UBM model, λ_{ubm}, as

$$score_{\boldsymbol{O},\boldsymbol{\lambda_s}} = \frac{P(\boldsymbol{O} \mid \lambda)}{P(\boldsymbol{O} \mid \lambda_{ubm})} \tag{4.7}$$

Taking logarithms for practical issues, the score simplifies to

$$score_{\boldsymbol{O},\lambda_s} = log(P(\boldsymbol{O} \mid \lambda_s)) - log(P(\boldsymbol{O} \mid \lambda_{ubm})) \tag{4.8}$$

Linear scoring proposes a linear approach of this scoring function based on the first order Taylor's series expansion of the first term, $log(P(\boldsymbol{O} \mid \lambda_s))$ evaluated at the UBM mean supervector point, as follows (see Appendix B)

$$log(P(\boldsymbol{O} \mid \lambda_s)) \simeq log(P(\boldsymbol{O} \mid \lambda_{ubm})) + \bigtriangledown P(\boldsymbol{O} \mid \lambda_{ubm})^T[\boldsymbol{\mu}](\boldsymbol{\mu}_s - \boldsymbol{\mu}) \tag{4.9}$$

being $\boldsymbol{\mu}_s$ and $\boldsymbol{\mu}$ the mean supervectors of λ_s and λ_{ubm} respectively and and $(\boldsymbol{\mu}_s - \boldsymbol{\mu})$ the difference of target model λ_s and UBM mean supervectors. It can be see also (see Appendix B) that

$$\bigtriangledown P(\boldsymbol{O} \mid \lambda_{ubm})_k[\boldsymbol{\mu}] = \sum_t \boldsymbol{\Sigma}_k^{-1} P_{kt}(\boldsymbol{o_t} - \boldsymbol{\mu}_k) \tag{4.10}$$

The linear approach carries itself a number of advantages with respect to the classic scoring method. First of all, the need of computing the term $log(P(\boldsymbol{O} \mid \lambda_{ubm}))$ for every utterance with data $\boldsymbol{O}$ is removed since is cancelled. To see that it suffices to substitute equation 4.9 in equation 4.8

$$\begin{aligned} S_{\boldsymbol{O},\boldsymbol{\lambda}_s} &= log(P(\boldsymbol{O} \mid \lambda_{ubm})) + \bigtriangledown log(P(\boldsymbol{O} \mid \lambda_{ubm})^T)[\boldsymbol{\mu}](\boldsymbol{\mu}_s - \boldsymbol{\mu}) - log(P(\boldsymbol{O} \mid \lambda_{ubm})) \\ &= \bigtriangledown log(P(\boldsymbol{O} \mid \lambda_{ubm})^T)[\boldsymbol{\mu}](\boldsymbol{\mu}_s - \boldsymbol{\mu}) \end{aligned} \tag{4.11}$$

Further, the term $(\boldsymbol{\mu}_s - \boldsymbol{\mu})$ is just the offset in a classical MAP adaptation in which only one EM iteration is done. Taking advantage of this fact, target speaker models can be expressed in GMM linear scoring as the *offsets* in MAP adaptation and therefore avoiding the dependence of the UBM from this step on. Moreover, it can be shown that the term $\bigtriangledown log(P(\boldsymbol{O} \mid \lambda_{ubm})^T)$, evaluated at the UBM mean supervector point, corresponds to the first order statistics of the data $\boldsymbol{O}$ with respect to the UBM, normalized by the covariance matrix $\boldsymbol{\Sigma}$ (see Appendix B). Therefore, the scoring function is reduced to a dot product between the MAP *offset* model and the first order statistic vector calculated from $\boldsymbol{O}$ with respect to the UBM.

Summarizing the above described, to obtain the score given a dataset of frames $\boldsymbol{O}$, a target model and a UBM is simplified to the next steps:

1. Compute 0rd and 1st normalized order stats from $\boldsymbol{O}$ (train and test) with respect to the UBM model:

$$0th \longrightarrow n_k = \sum_t P_{kt} \tag{4.12}$$

$$1st_{norm} \longrightarrow \boldsymbol{f}_k = \sum_t \boldsymbol{\Sigma}_k^{-1} P_{kt}(\boldsymbol{o}_t - \boldsymbol{\mu}_k) \tag{4.13}$$

2. Compute the target model of the speaker s as the offset in MAP adaptation from the training sufficient statistics:

$$\boldsymbol{\mu}_s = (\tau \boldsymbol{I} + \boldsymbol{N})^{-1} \overline{\boldsymbol{f}} \tag{4.14}$$

3. Compute the score as the dot product of the testing first stats and the target model:

$$S_{\boldsymbol{O},\lambda_s} = \boldsymbol{\mu}_s^T \overline{\boldsymbol{f}} \tag{4.15}$$

The step 2. can be easily substituted by the offset $\boldsymbol{V}\boldsymbol{y} + \boldsymbol{D}\boldsymbol{z}$ instead of the MAP offset adaptation integrating thus linear scoring within a FA framework. Also the first-order statistics belonging to the test utterance can be compensated by means of equation 4.6 to take into account the session variability of the test utterance.

The figure 4.1 illustrates the linear scoring approximation by representing the likelihood function and its linear approximation as a scoring function in the GMM means space. The actual likelihood function is represented by the curve and it produces a score, $S_{\boldsymbol{O},\boldsymbol{\mu}_s}$, for each par of observations and model with supervector means $\boldsymbol{\mu}$. The line, tangent to the likelihood in point $\boldsymbol{\mu}$, represents the approximation of the likelihood via linear scoring. Note that as the target model has been derived by MAP from the UBM or in a similar procedure, both models should be close into the GMM means space. This fact guarantees that the produced score is a good estimation of the actual score.

4.4. Toward Efficient and Robust Text-Independent Speaker and Language Recognition Acoustic Systems

It has been a prime goal in the deployment of this Thesis to yield robust acoustic but at same time efficient, in computational terms, SV and SLR systems [Gonzalez-Dominguez *et al.*, 2010b,d, 2009]. The goa of this work has been largely corroborated through different international evaluations such as those promoted by the National Institute of Standards and

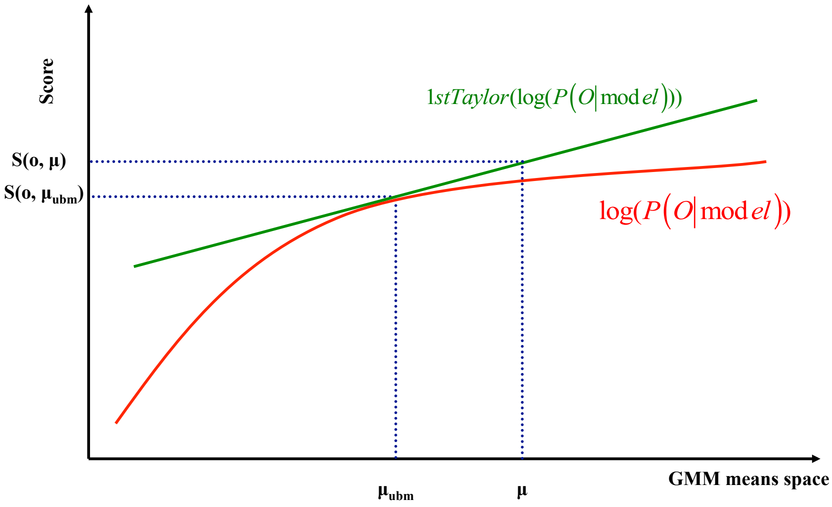

Figure 4.1: Linear Scoring Representation. The scoring is computed as an approximation of the likelihood function over the point defined by the UBM mean supervector.

Technology (NIST) in both speaker and language recognition evaluations (SRE, LRE) since the year 2006 (SRE06, SRE08, SRE10, LRE07 and LRE09).

In this section, two efficient and robust GMM-UBM systems based on FA for SV and SLR are presented. Those compress the state-of-the-art acoustic approach for both disciplines and are the current base for further research.

4.4.1. An efficient JFA based GMM-UBM systems for SV

The algorithm presented in Table 4.4 summarizes an efficient version of JFA integrated into a GMM-UBM classical framework, which includes much of the simplifications described in section 4.2. The compensation is carried out in the statistics domain and the scoring is performed via linear scoring.

Note that the session compensation at the training stage is accounted by compensating the first-order statistics before computing the speaker component by using equation 4.6, and a similar scheme is done at the testing stage. For the sake of simplicity a single recording is considered to be available for every speaker, otherwise a channel factors vector x_h for each utterance and first h should be considered, to compensate first-order statistics vectors for each pair utterance h and speaker s. Those, once compensated, could be accumulated remaining the rest of the process the same.

Table 4.3 analyses the computational time of the efficient JFA system presented besides an analogous system which use SVM with session variability compensated first-order statistics. As it can be seen, using similar schemes of session variability compensation as well as to incorporate linear scoring produce great improvements in terms of computational time.

4.4.2. An efficient JFA based GMM-UBM systems for SLR

A very similar approach to that described above for SV is presented in 4.5 for SLR purposes, with several modifications. These being:

1. The within scatter matrix refers languages as classes rather than speakers.
2. The speaker variability subspace disappears, as it is considered session variability.
3. Training languages models are computed as average of individual models for utterances belonging to same language.

An analogous term to the speaker variability subspace, the language variability subspace, could be also considered into this scheme, however due to often the number of languages to recognise is much smaller than the number of speakers, the latent factors associated to this subspace degenerate to vectors of a few dimensions that do not significant contribute to the recognition performance. Nevertheless, as it was shown in [Castaldo *et al.*, 2009], those vectors contain discriminant information and could be used for instance as the training features in a GMM-SVM framework and as long as the task covers more languages, they could gain more and more importance.

4.5. Summary

This chapter has extensively covered the integration of the theoretical FA framework into the well-known GMM and SVM systems for SV and SLR. The different forms of FA to fit with those classification schemes at three different levels, namely, the feature, the model and the statistics domain have been exposed and detailed. The advantages and disadvantages of each of those approaches has been individually examined and are summed up in Table 4.2.

A novel contribution based on the integration of FA into a SVM system in the statistics domain has been also presented. This approach inherits the benefits of the compensation in the statistics domain, avoiding the costly frame-by-frame compensation process and fully treatment of all the FA components.

The final part of this chapter was concentrated into developing efficient SV and SLR systems. In this direction, complete *recipes* to achieve efficient FA based systems integrated to SV and SLR have been exposed. Those algorithms are supported by several and novel publications conducted during the research, which has originated this Dissertation.

Step	**System**	
	JFA	**SVM-SV**
Development		
UBM training (2M feature vectors, gender dependent)	4h	(4h)
Training Variability Subspace $\boldsymbol{U}/\boldsymbol{V}$	1h/1h	1h
Feature extraction (per $\sim$ 265s file)		
MFCC	2s	(2s)
Training (per $\sim$ 265s file)		
GMM-train	8s	(8s)
FA point estimate	0.1s	(0.1s)
SVM-train	-	120s
Total (train)	10.1s	130.1s
xRT train (CPU/speech)	0.04RT	0.50RT
Testing (per $\sim$ 265s file)		
SV-train	-	8s
FA point estimate	0.01	(0.01)
Scoring (frame by frame/ linear scoring)	0.2s/1×10^{-4}	3.2s
t-norm (100 models)	20s/1×10^{-2}	320s
Total (test)	22.2s/2.02s	331.2s
xRT test (CPU/speech)	0.08RT/7.5×10^{-3}RT	1.24RT

Table 4.3: *Execution times for acoustic JFA and SVM supervector with session variability compensation systems. Numbers in brackets means already compute through the other system.*

An efficient version of JFA integrated into classical GMM-UBM for SV

```
I   Train an Universal Background Model, $\lambda_{ubm} = \{\omega_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\}_{i=1}^{N}$

    1. $\boldsymbol{O}_{dev} := observations(devData)$;
    2. $\lambda_{ubm} := clustering(\boldsymbol{O}_{dev})$;          % K-Means or Binary Splitting.
    3. $\lambda^*_{ubm} := EM_{ML}(\boldsymbol{O}_{dev})$;            % Maximum Likelihood via EM iterations.

II  Initialization of Hyperparameters, $\Lambda = \{\boldsymbol{\mu}, \boldsymbol{V}, \boldsymbol{U}\}$

    1. $\boldsymbol{\mu} := \boldsymbol{\mu}_{ubm}$
    2. for each utterance i $\boldsymbol{O}_{dev_i}$ in dev set $\boldsymbol{O}_{dev}$:
           $\lambda_i := EM_{MAP}(\boldsymbol{O}_{dev_i}, \lambda^*_{ubm})$;     % training model via MAP adaptation
           $\boldsymbol{X}(:, i) := \boldsymbol{\mu}^i$;                      % stacking mean supervector in column form
       end
    3. $\boldsymbol{S}_b := betweenScatterMatrix(\boldsymbol{X})$; $\boldsymbol{S}_w := withinScatterMatrix(\boldsymbol{X})$;
    4. $\boldsymbol{V} := PCA(\boldsymbol{S}_b)$;
    5. $\boldsymbol{U} := PCA(\boldsymbol{S}_w)$;

III Hyperparameters Refinement, $\Lambda = \{\boldsymbol{\mu}, \boldsymbol{V}, \boldsymbol{U}, \boldsymbol{D}\}$

    1. $\boldsymbol{O}_{dev2} := observations(devData2)$;
    2. $\Lambda^* := maximize_{\Lambda}(\boldsymbol{O}_{dev2}, \Lambda)$;

IV  Train speaker target models

    1. $\boldsymbol{O}_{train} := observations(trainData)$;
    2. for each speaker s with train material $\boldsymbol{O}_{train_s}$ in train set $\boldsymbol{O}_{train}$:
           $[\boldsymbol{n_s}, \boldsymbol{f_s}] := sufficientStatistics(\boldsymbol{O}_{train_s}, \boldsymbol{\mu}_{ubm})$
           $\boldsymbol{x} := pointEstimate(\boldsymbol{n}_s, \boldsymbol{f}_s, \boldsymbol{U}^*)$;
           $\boldsymbol{f}^*_s := compensate(\boldsymbol{f}_s, \boldsymbol{U}^*, \boldsymbol{x})$
           $\boldsymbol{y}_s := pointEstimate(\boldsymbol{n}_s, \boldsymbol{f}^*_s, \boldsymbol{V}^*)$;
           $\boldsymbol{z}_s := pointEstimate(\boldsymbol{n}_s, \boldsymbol{f}^*_s, \boldsymbol{D}^*)$;
           $\boldsymbol{\mu}_s := \boldsymbol{V}\boldsymbol{y} + \boldsymbol{D}\boldsymbol{z}$;
       end

V   Testing

    1. $\boldsymbol{O}_{test} := observations(testData)$;
    2. for each utterance j $\boldsymbol{O}_{test_j}$ in train set $\boldsymbol{O}_{test}$:
       $[\boldsymbol{n_s}, \boldsymbol{f}_s] := sufficientStatistics(\boldsymbol{O}_{test_s}, \boldsymbol{\mu})$
       $\boldsymbol{x} := pointEstimate(\boldsymbol{n}_s, \boldsymbol{f}_s, \boldsymbol{U}^*)$;
       $f^*_s := compensate(\boldsymbol{f}_s, \boldsymbol{U}^*, \boldsymbol{x})$
           for each model $\lambda_s$:
               $S_{\mathbf{0},\lambda_s} := \boldsymbol{\mu}_s \cdot \boldsymbol{f}^h_s$;
           end
       end
```

Table 4.4: A robust and efficient acoustic GMM system for speaker verification.

An efficient version of JFA integrated into classical GMM-UBM for SLR

```
I   Train an Universal Background Model, $\lambda_{ubm} = \{\omega_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\}_{i=1}^{N}$

    1. $\boldsymbol{O}_{dev} := observations(devData)$;
    2. $\lambda_{ubm} := clustering(\boldsymbol{O}_{dev})$;        % K-Means or Binary Splitting.
    3. $\lambda^*_{ubm} := EM_{ML}(\boldsymbol{O}_{dev})$;          % Maximum Likelihood via EM iterations.

II  Initialization of Hyperparameters, $\Lambda = \{m, \boldsymbol{U}\}$

    1. $\boldsymbol{\mu} := \boldsymbol{\mu}_{ubm}$
    2. for each utterance i $\boldsymbol{O}_{dev_i}$ in dev set $\boldsymbol{O}_{dev}$:
           $\lambda_i := EM_{MAP}(\boldsymbol{O}_{dev_i}, \lambda^*_{ubm})$;     % training model via MAP adaptation
           $\boldsymbol{X}(:, i) := \boldsymbol{\mu}^i$;                   % stacking mean supervector in column form
       end
    3. $\boldsymbol{S}_w := withinScatterMatrix(\boldsymbol{X})$;
    4. $\boldsymbol{U} := PCA(\boldsymbol{S}_w)$;

III Hyperparameters Refinement, $\Lambda = \{\boldsymbol{\mu}, \boldsymbol{U}\}$

    1. $\boldsymbol{O}_{dev2} := observations(devData2)$;
    2. $\Lambda^* := maximize_{\Lambda}(\boldsymbol{O}_{dev2}, \Lambda)$;

IV  Train language target models

    1. $\boldsymbol{O}_{train} := observations(trainData)$;
    2. for each language l
           for each utterance h $\boldsymbol{O}_{train_h}$ in train set $\boldsymbol{O}_{train_l}$:
               $[\boldsymbol{n}_h, \boldsymbol{f}_h] := sufficientStatistics(\boldsymbol{O}_{train_h}, \boldsymbol{\mu}_{ubm})$
               $\boldsymbol{x}_h := pointEstimate(\boldsymbol{n}_h, \boldsymbol{f}_h, \boldsymbol{U}^*)$;
               $\boldsymbol{f}^*_h := compensate(\boldsymbol{f}_h, \boldsymbol{U}^*, \boldsymbol{x})$
               $\boldsymbol{z} := pointEstimate(\boldsymbol{n}_h, \boldsymbol{f}^*_h, D^*)$;
               $\boldsymbol{\mu}_h := \boldsymbol{D}\boldsymbol{z}$;
           end
           $\lambda_l := average(\lambda_l, \lambda_h)$;
       end

V   Testing

    1. $\boldsymbol{O}_{test} := observations(testData)$;
    2. for each utterance h $\boldsymbol{O}_{test_h}$ in train set $\boldsymbol{O}_{test}$:
           $[\boldsymbol{n}_h, \boldsymbol{f}_h] := sufficientStatistics(\boldsymbol{O}_{test_h}, \boldsymbol{\mu})$
           $\boldsymbol{x}_h := pointEstimate(\boldsymbol{n}_h, \boldsymbol{f}_h, \boldsymbol{U}^*)$;
           $\boldsymbol{f}^*_h := compensate(\boldsymbol{f}_h, \boldsymbol{U}^*, \boldsymbol{x}_h)$
           for each model l:
               $S_{\lambda_l,h} := \boldsymbol{\mu}_l \cdot \boldsymbol{f}^*_h$;
           end
       end
```

Table 4.5: *A robust and efficient acoustic GMM system for spoken language recognition.*

Chapter 5

Experimental Framework

This chapter describes the adopted experimental framework to assess and present the set of experiments/results contained within this Dissertation.

5.1. Introduction

As in other young fields of scientific research, a common practice at the beginnings of the SV or SLR research works in the 1970's decade was to report experimental results using data expressly captured for the specific set of experiments conducted, being this collection process, most of the times, carried out at hand [Atal, 1976; Markel and Davis, 1979].

The increasing interest in the area linked to the apparition of several research groups interested in facing similar or same problems, soon demanded common experimental frameworks (i.e databases and protocols) which allow establishing fair comparison among different groups technology as well as fostering collaborative research and stimulating intellectual discussions in the area.

In that sense, a crucial milestone in the development of SV and SLR technologies was the foundation and organization by the American National Institute of Standards and Technology (NIST) of the international evaluations series in the area, the speaker and spoken language recognition evaluation series, NIST SRE's and NIST LRE's respectively. The SRE's and LRE's series, starting at 1996 have meant a beneficial feedback cycle which extends to present and where new challenges supported by new databases are dealt with in a common forum to mayor benefit of the SV and SLR technology. In particular, NIST evaluations are designed to foster research progress to ([Doddington *et al.*, 2000]):

1. Exploring promising new ideas in speaker recognition.

2. Developing advanced technology incorporating these ideas.

3. Measuring the performance of this technology.

In this chapter the databases besides the adopted evaluation protocols for both SV and SLR used to assess the experiments conducted in this dissertation are described. As well, the baseline systems used in order to compare described techniques and new algorithms are detailed.

5.2. Databases

This section describes most widely used databases by the scientific community to assess SV and SLR technology.

5.2.1. Automatic speaker recognition databases

- **Switchboard 1** [Godfrey and Holliman, 1993; Godfrey *et al.*, 1992]. SWB1 consisted of conversational speech recorder over landline telephone from both carbon button and electret telephone handsets. The recordings are around 2.5 minutes, containing american-english speech of a 543 U.S participants. SWBI was released in 1997

- **Switchboard 2** [Graff *et al.*, 1998, 2002, 1999]. SWB2 was acquired in three phases according to the three different areas of U.S.A were it was collected, namely, Mid-Atlantic, Midwest and southern regions. It is based on landline conversational speech as SWB1, but a higher degree of variability was captured as participants were encouraged to use a variety of handsets. The three phases of SWB2 was released in 1998, 1999 and 2002 with about 657, 679 and 640 different speakers respectively.

- **Switchboard Cellular** [Graff *et al.*, 2001]. SWBCELL contains conversational speech in American-english recorded over cellular networks, mostly consisted of GSM and CDMA. SWBCELL was released in two parts in 2001 and 2004 respectively with a total of 254 and 419 speakers respectively. SWBCELL was released in 2001.

- **Ahumada** [Ortega-Garcia *et al.*, 2000]. The Ahumada database was recorded by the ATVS biometric recognition group. It contains speech in Spanish over telephone and two types of microphones under controlled conditions. Ahumada was included into the NIST SRE 2001 evaluation, providing multi-language variation (english, spanish).

- **Mixer.** [Cieri *et al.*, 2006, 2007]

  The increasing need of counting with more appropriate data to cope with the new challenges emerged in SV, required to develop a more ambitious mechanism to collect speech data. In that context, the Mixer collection database arose to satisfy this demand, with a main goal of building a very challenging database which includes variability across different aspects such as languages, handsets/channel (different microphones and handsets), age variation, gender and speech style (i.e conversational telephone and interview speech).

  The Mixer database development goes hand in hand with the NIST SRE's providing the needed data to evaluate new challenges proposed in those.

I **Mixer 1, 2, 3** [Cieri *et al.*, 2006]. The three first phases of Mixer contains the conversational telephone speech kernel of the global Mixer database. They contain more than 1867 speakers, where multi-language data captured in a wide number of handsets is considered. Also, they are balanced in gender and the age variations is broad (16 - >50).

II **Mixer 4** [Cieri *et al.*, 2007]. Mixer 4 was focused to cope with multi-channel data. Up to 14 different microphone were set to simultaneously record incoming calls, bringing about a broad microphone variability. Mixer 4, with more than 400 participants, has been used in past SRE's evaluations since 2005, being one of the main focus in 2006 and 2008.

III **Mixer 5, 6** [Cieri *et al.*, 2007]. Mixer 5 and Mixer 6 followed a dual goal. First to capture interview data besides conversational telephone speech, and second to collect speech data where particular low or high vocal effort is done by the participant. Mixer 5, with more than 200 participants, has been used in past 2008 and 2010 evaluations being one of the main focus in 2010.

- **Ahumada III** [Ramos *et al.*, 2008]. Ahumada III is a forensic speech database in Spanish collected from *real forensic cases*. In its current release, the database presents 61 male speakers recorded using the systems and procedures followed by Spanish Guardia Civil police force. As a forensic, Ahumada III contains a huge variety conditions in terms of number of available calls and amount of data.

5.2.2. Automatic language recognition databases

The databases used in this Dissertation concerning SLR are governed by the NIST LRE's series. Since 1996 NIST LRE's series have included data belonging to different languages, mostly collected by the Linguistic Data Consortium (LDC) [1]. Table 5.3 shows all the languages that have been labelled as *target* in any of the LRE's. Among the databases included in this data, it is worth to highlight the following projects.

- **CallFriend** [Graff, 1996]. The CallFriend project includes a wide variety of different language databases acquired following a identical protocol by the LDC primarily in support of the project on Language Identification (LID), sponsored by the U.S. Department of Defense.

- **CallHome** [Graff, 1996]. As CallFriend, the CallHome project cover a wide variety of languages recorded over telephone speech. It was collected by the LDC primarily in support of the project on Large Vocabulary Conversational Speech Recognition (LVCSR), sponsored by the U.S. Department of Defense.

[1] Linguistic Data Consortium http://www.ldc.upenn.edu/

- **Voice of America** [Graff, 2009]. The VOA project arose as a collaboration between the LDC, the Speech@FIT group (Brno, Czech Republic) and the official external radio and television broadcasting service of the U.S Government (Voice of America broadcast news) and therefore it contains broadcast speech. VOA recordings are publicly available on VOA's website [1], they contains more than 50 languages, and they were included in the 2009 LRE.

Table 5.1 collects all the languages that have been target languages in one or more NIST LRE evaluations besides their coverage in terms of millions of native speakers and other side information, such as the regions were those are spoken.

5.3. Evaluation of Performance

In a speaker, language and, in general, in a verification biometric systems two types of error can occur, namely false rejection (FR) and false acceptance (FA). The former is produced when a *true* identity is rejected by the system whilst the latter happens when a non-valid identity claim is accepted. Both types of errors depends on the threshold defined in the system, as its value will mark when an individual will be accepted or rejected. In that sense, the higher threshold the higher false rejection and lower false acceptance errors, as the system will be more strict when accepting users (desirable for instance in security systems, bank, personal data, etc.). By other hand, the lower threshold the lower false rejection and the higher false alarm.

The pair of errors (FA, FR) define the *operating-point* of the system, that is the FA and FR errors given the fixed threshold, being the matter of fixing the threshold a trade-off between the two types of errors. In practice, in order to measure the FA and FR of a system a large test corpus is used and counts of the number of errors of each type are used. The Figure 5.1 represents the FA and FR error of a determined system with non-target and target distributions and a given threshold.

In function of this two types of errors, two measures adopted by NIST in the language and speaker recognition evaluations will be adopted to measure the systems performance.

5.3.1. Detection Error Trade-off curve

The Detection Error Trade-off (DET) curves are a well-known visual form to represent the systems performance of biometric systems and in general binary classifications systems. Basically, a DET curve plots the FR error versus FA error, and it can be seen as non-linear scaled-axes version of ROC curves. This scaled has the main objective of obtaining more linear systems error curves, allowing a better visual comparison of the performance systems. An example of a DET curve is depicted in Figure 5.2.

[1] http://www.voanews.com/english/news/

Language/Dialect	**Family**	**Official in / region for dialects**	**∼ native speakers (mil.)**
Amharic	Afro-Asiatic	Ethiopia	32
Arabic	Afro-Asiatic	26 states north-Africa and Middle East	280
Bengali	Indo-European	Bangladesh, India, Sierra Leone	230
Bosnian	Indo-European	Bosnian and Herzegovina, Montenegro	4
Chinese(Cantonese)	Sino-Tibetan	China, Taiwan, Singapore	70
Chinese(Mandarin)	Sino-Tibetan	China, Taiwan, Singapore	1365
Chinese(Min)	Sino-Tibetan	China, Taiwan, Singapore	50
Chinese(Wu)	Sino-Tibetan	China, Taiwan, Singapore	90
Creole(Haitian)	Creole	Haiti	12
Croatian	Indo-European	Croatia, Bosnian and Herzegovina	5,5
Dari	Indo-European	Afghanistan	30
English(American)	Indo-European	U.S.A	309
English(Indian)	Indo-European	India	90
Farsi	Indo-European	Iran, Afghanistan, Tajikistan	70
French	Indo-European	20 countries [France, Canada ...]	110
Georgian	Kartvelian	Georgia	7
German	Indo-European	7 countries centre-Europe [German, Austria ...]	97
Hausa	Afro-Asiatic	9 countries Africa [Nigeria, Cameroon ...]	25
Hindustani(Hindi)	Indo-European	India	180
Hindustani(Urdu)	Indo-European	Pakistan, India	60
Japanese	Japonic	Japan	130
Korean	Korean	North Korea, South Korea, Yanbian (china)	78
Pashto	Indo-European	Afghanistan, Pakistan	60
Portuguese	Indo-European	Brazil, Angola, Mozambique, Portugal	236
Russian	Indo-European	8 countries east-europe [Rusia, Kazakhstan ...]	175
Spanish(Caribbean)	Indo-European	Dominican Republic, Cuba, Puerto Rico	25
Spanish(non-Caribbean)	Indo-European	21 countries [Mexico, Spain ...]	500
Tamil	Dravidian	India, Sri Lanka, Singapore	66
Thai	Tai-Kadai	Thailand, Northern Malasya	60
Turkish	Altaic	Turkey, Ciprus	83
Ukrainian	Indo-European	Ukraine	47
Vietnamese	Austro-Asiatic	Vietnam	73

Table 5.1: *Information about languages/dialect involved as target languages in LRE series.*

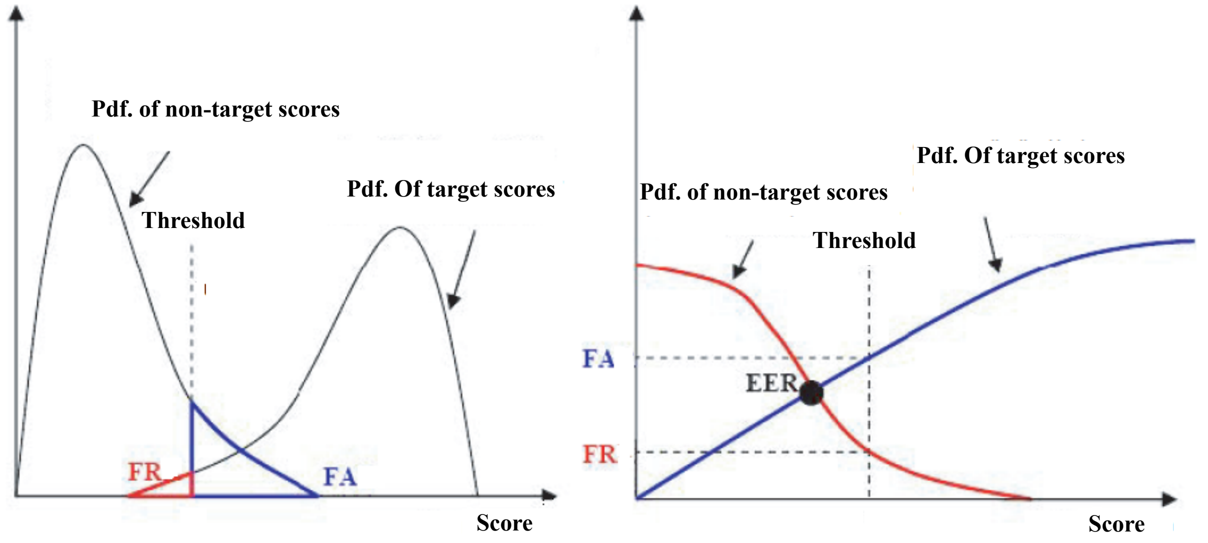

Figure 5.1: Representation of False Reject (FR) and False Acceptance (FA) errors in a biometric recognition system.

5.3.2. Detection cost

Apart from the DET curve and its inherent EER associated, NIST provides and additional cost function which measure the system performance establishing a fixed cost to FA and FR errors as well as a priori probability for target and non-target individuals. This cost is defined for speaker verification as:

$$C_{Det} = C_{FR} \cdot P_{FR|S_T} \cdot P_T + C_{FA} \cdot P_{FA|S_{NT}} \cdot P_{S_{NT}} \qquad (5.1)$$

where C_{FR} and C_{FA} are the associated costs to FR and FA errors respectively; $P_{FR|S_T}$ (the probability of false reject given a target speaker) measures the system FR; $P_{FA|S_{NT}}$ (the probability of false acceptance given a non-target speaker) measures the system FA; and finally, P_T and $P_{NT} = 1 - P_T$ the prior target and non-target probability.

In NIST speaker evaluations and by extension, in this work, costs and target probability will be set as follows:

- $C_{FA} = C_{FR} = 1$
- $P_T = 0.001$

Regarding SLR an average cost which accumulates all the possible errors considering the

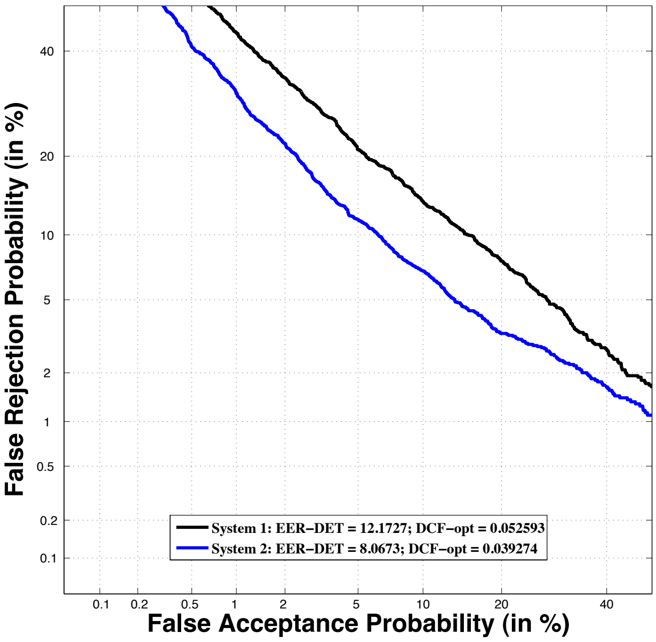

Figure 5.2: *Example of DET curve. System 1 and 2 are compared in terms of FR and FA errors.*

involved languages is used. This cost, the C_{avg}, is defined as:

$$C_{avg} = \frac{1}{N_L}\sum_{L_T}\left\{C_{FR}\cdot P_{L_T}\cdot P_{FR|L_T} + \sum_{L_{NT}} C_{FA}\cdot P_{L_{NT}}\cdot P_{FA|L_T,L_{NT}} + C_{FA}\cdot P_{oos}\cdot P_{FA|L_T,L_O}\right\} \quad (5.2)$$

where

N_{L_T} is the number of target languages, $N_{L_{NT}}$ the number of languages non-target, L_O represents the "out-of-set" languages, and P_{oss} the prior probability that a language be a out-of-set language, defined as:

$$P_{oss} = \begin{cases} 0.0 & for\ the\ closed-set\ condition \\ 0.2 & for\ the\ open-set\ condition \end{cases}$$

5.4. Protocol and Tasks Definition

In this Dissertation the protocols defined by NIST SRE (2006, 2008) and LRE 2009 have been used to evaluate the SV and SLR systems respectively.

5.4.1. Automatic speaker recognition task definition

As a SV task, the essence of NIST SRE is to determine whether a specified speaker represented by a determined amount of training speech, is speaking during a given test segment of conversational speech. Those *trials* are conditioned by the nature of the training and test recordings (duration, speech style) defining so different task conditions.

Task	# Models	#Tests	#trials
tel-tel SRE'08	1788	2573	57050
tel-tel SRE'08-male	648	895	12922
tel-tel SRE'08-female	1140	1678	24128
10s-10s SRE'08	1789	1526	21951
10s-10s SRE'08-male	648	545	7799
10s-10s SRE'08-female	1141	981	14152

Table 5.2: *Composition of development datasets.*

5.4.1.1. Task conditions

In order to evaluate the presented systems and algorithms through next chapters, the following task conditions has been used:

A *tel-tel* SRE'08.

The *tel-tel* SRE'08 condition includes the telephone part of the core condition of NIST SRE08 evaluation. It consists of trials were telephone conversational recordings of approximately five minutes total duration ($\sim$ 2.5 minutes of effective speech) are involved in both training and testing. Table 5.2 details the number of models, test recordings and trials considered in this task.

B *10s-10s* SRE'08.

It consists of trials where telephone conversational recordings of approximately 10s are involved in both training and testing. Details can be found in Table 5.2.

C A simulated challenging "real-world" scenario (SRE'05, SRE'06).

We simulate by means of this condition an adverse scenario where problems as treated in chapter 7 (database mismatch and short durations) are simulated. To this aim, data from the 2005 and 2006 NIST Speaker Recognition Evaluations (NIST SRE) was used to develop an experimental framework. These datasets were chosen as they cover a wide range of acoustic (telephone and microphone) and environmental scenarios, allowing for vigorous testing under mismatched conditions.

Two development datasets, namely $dTel$ and $dMic$, were differentiated. The $dTel$ consists of SRE'04 and SRE'05 telephone data supplemented with data belonging to SWBII phase I and phase II databases. This collection was chosen to provide a broad coverage of telephone conditions, whilst also providing a high number of different speakers. The $dMic$ dataset was obtained from the microphone subset of the MIXER corpus and SRE'05 data.

In order to simulate the data scarcity problem, the $dMic$ set was divided into sets with differing amounts of data, obtaining different degrees of data scarcity. Specifically, three

	Databases	# Speakers	# Utterances
	SWB-II	325	1300
$dTel$	MIXER(SRE'04)	150	994
	MIXER(SRE'05-tel)	40	297
$dMic$	MIXER(SRE'05-mic)	45	1260
$dMic_{10}$	MIXER(SRE'05-mic)	45	450
$dMic_5$	MIXER(SRE'05-mic)	45	225
$dMic_3$	MIXER(SRE'05-mic)	45	135

Table 5.3: *Composition of the development dataset C ("real-world" scenario).*

restricted sets were built: $dMic_{10}$, $dMic_5$ and $dMic_3$. These were formed with only 10, 5 and 3 utterances per speaker present in $dMic$. Table 5.3 shows a breakdown development dataset compositions.

The SRE'06 data was utilised as the test dataset. Testing was performed using the test conditions specified in the SRE'06 [NIST, 2006] protocol, and using additional conditions specified and distributed by participating sites during the SRE'08 [1]. The test conditions examined were as follows: *1conv4w-1conv4w*, *1conv4w-1mic*, *1mic-1conv4w* and *1mic-1mic*.

5.4.2. Automatic spoken language recognition task definition

LRE'09 evaluation included, for the first time, data coming from two very different audio sources. Besides Conversational Telephone Speech, hereafter CTS, used in past evaluations, telephone speech belonging to broadcast news was used for both train and test purposes. Broadcast data was obtained via an automatic acquisition system from Voice of America news (VOA) where telephone and non-telephone speech is mixed. Up to 2 terabytes of speech, automatically labelled in language and type, were distributed to participants. Further, around 80 audited segments for each target language (of approximately 30 seconds duration each) was provided too for development purposes.

Both closed and open-set modes were defined as tasks in this evaluation each one tested with duration segments of 3, 10 and 30 seconds. We refer to closed-set as the task when only target languages are included in the test trials set, and to open-set when other non-target languages (unknown to participants) are also included. In this evaluation, 23 target languages were involved in closed-set as it is showed in Table 5.4 and 40 in open-set. More detailed information can be found in the LRE'09 evaluation plan [NIST, 2009].

[1] Additional conditions for auxiliary microphone training and testing were distributed on the SRE'08 Google Group list. Thanks to Doug Reynolds, David van Leeuwen, Albert Strasheim and Nicholas Scheffer for preparing and scrutinising these lists. Further details on these conditions can be obtained from the author or at http://groups.google.com/group/sre2008

Language	**Abbreviation**	**Data Type (VOA/CTS)**
Amharic	*amha*	$VOA/-$
Arabic	*arab*	$-/CTS$
Bengali	*beng*	$-/CTS$
Bosnian	*bosn*	$VOA/-$
Chinese (Cantonese)	*cant*	$VOA/-$
Chinese (Mandarin)	*mand*	VOA/CTS
Creole	*creo*	$VOA/-$
Croatian	*croa*	$VOA/-$
Dari	*dari*	$VOA/-$
English (Indian)	*inen*	$-/-$
English (American)	*usen*	VOA/CTS
Farsi	*fars*	VOA/CTS
French	*fren*	$VOA/-$
Georgian	*geor*	$VOA/-$
German	*germ*	$-/CTS$
Hausa	*haus*	$VOA/-$
Hindi	*hind*	VOA/CTS
Japanese	*japa*	$-/CTS$
Korean	*kore*	VOA/CTS
Pashto	*pash*	$VOA/-$
Portuguese	*port*	$VOA/-$
Russian	*russ*	VOA/CTS
Spanish	*span*	VOA/CTS
Tamil	*tami*	$-/CTS$
Thai	*thai*	$-/CTS$
Turkish	*turk*	$VOA/-$
Ukranian	*ukra*	$VOA/-$
Urdu	*urdu*	$VOA/-$
Vietnamese	*viet*	VOA/CTS

Table 5.4: *Alphabetical list of languages used as development for LRE'09 evaluation. In bold, LRE'09 target languages.*

In order to face this new challenge, where database mismatch play and important role [Ramos *et al.*, 2008], an ATVS development dataset was set up, ATVS-Dev09 onwards. This dataset was built to reproduce in the most accurately possible way, blind evaluation conditions by using different sets of CTS and VOA data provided by NIST. ATVS-Dev09 covered all target evaluation languages and test evaluation duration segments (3, 10 and 30 seconds). Table 5.3 shows the 23 evaluation target languages along with ATVS available data type per language.

Specifically, the CTS training material (ATVS-DevTrain09) consisted of the Callfriend database, the full-conversations of LRE'05 and development data of LRE'07. For Russian data we used also RuSTeN [1]. Telephone broadcast data was obtained from speech segments (minimum length 30s.) extracted from VOA long files using telephone labels provided by NIST.

The test material (ATVS-DevTest09) was obtained from the test part of LRE'07 (for target languages in both LRE'07 and LRE'09), and from manually labelled data from VOA provided by NIST. Finally, about 15,000 segments, balanced in segments of 3, 10 and 30 seconds, while LRE'09 evaluation included about 15.000 segments per duration (∼45,000 segments) and therefore about 1 million trials since every segment is tested against every target language.

5.5. Summary

In this chapter the experimental protocol used for the experiments presented in this Dissertation has been detailed. The experimental protocol adopted as well as the databases used are those well-known proposed by NIST in their two past language and speaker recognition evaluations (LRE 2009 and SRE2010). This fact favours the replication or comparison of all the experiments conducted in this Dissertation by other researchers.

[1] LDC 2006S34 ISBN 1-58563-388-7, www.ldc.upenn.edu

Chapter 6

Factor Analysis applied to SV and SLR systems: Part II (experimental)

This chapter presents and analyses the experimental results obtained by applying JFA in both SV and SLR systems.

6.1. Introduction

In previous Chapters 3 and 4, the theoretical framework of JFA, as well as how this is integrated within SV and SLR systems was addressed. It is now the purpose of this chapter to empirically evaluate the performance of JFA when dealing with large tasks of SV and SLR such as the challenging NIST speaker and language evaluations. Particularly, the speaker recognition evaluation NIST SRE 2008 (SRE'08) and the language recognition evaluation NIST LRE 2009 (LRE'09), which databases and protocols are defined in previous Chapter 5, have been used to this aim.

This chapter is clearly differenced in two parts. In the first part, the performance of JFA in SV is assessed and analysed in the telephone part of SRE'08, by comparing a step-by-step built JFA system versus a classical GMM-UBM framework. Second part is then devoted to evaluate the performance of JFA in SLR, in the context of the LRE'09, but also its fusion potential with other state-of-the-art techniques. Both parts present an exhaustive and detailed analysis supported by a wide set of experiments, which will lead us to a deep evaluation of the JFA performance as well as to empirically support one of the main attainments and goals of this Dissertation; to get robust, accurate and efficient SV and SLR systems. Further, other contributions of this Thesis, as the use of SVM through FA session variability compensated statistics for SLR as well as the use of anchor models as a back-end of SLR are evaluated [Gonzalez-Dominguez *et al.*, 2010b,d].

6.2. Joint Factor Analysis applied in SV Systems

In this section, the results obtained by incorporating JFA modelling within a classical GMM-UBM acoustic system are presented. Those results, conducted on the male/female telephone conditions of SRE'08 (Section 5.4.1), will provide us a clear and quantitative idea of the benefits of using JFA in order to palliate the session variability problem as well as the strengths of using a more proper modelling scheme for speaker variability as it is proposed in JFA.

Rather than presenting a direct comparison with/without using JFA, the objective of this section is to perform a step-by-step analysis, where different elements of the JFA model are sequentially included. Thus, a proper analysis of the importance of each element within the global JFA modelling scheme is evaluated. To this aim, through this section speakers models will be conducted from the classical MAP adaptation to the full JFA modelling by enabling/disabling elements of the JFA modelling equation; $\boldsymbol{\mu}_{hs} = \boldsymbol{\mu} + \boldsymbol{V}\boldsymbol{y}_s + \boldsymbol{D}\boldsymbol{z}_s + \boldsymbol{U}\boldsymbol{x}_{hs}$.

1. **MAP adaptation** ($\boldsymbol{V} = \boldsymbol{U} = 0$). The classical GMM-UBM framework where MAP adaptation is used to derive speaker models from a UBM, is set as the baseline system of this analysis. Here, the speaker component of new speaker models is defined by classical MAP adaptation terms, $\boldsymbol{\mu} + \boldsymbol{D}\boldsymbol{z}_s$, whilst no special care is taken regarding the session component.

2. **MAP adaptation with session variability compensation** ($V = 0$). MAP adaptation is, in this case, still adopted to derive speaker models, but session variability compensation is applied in the statitics domain for both training and test recordings. This compensation is accomplished previously to perform MAP adaptation.

3. **Eigenvoice adaptation with session variability compensation** ($D = 0$). We evaluate through this system the inclusion of eigenvoice adaptation rather than MAP adaptation once training and testing recordings have been session compensated.

4. **JFA modelling**. The full JFA model, which combine eigenvoice and MAP adaptation is then evaluated to represent speaker models is evaluated in this step. Again, session variability compensation is previously performed in the statistic domain.

5. **JFA modelling with $\boldsymbol{D}$ trained on data**. Finally, the JFA modelling where *residual* matrix $\boldsymbol{D}$ is estimated on training data rather explicitly derived from MAP adaptation is evaluated.

Whenever necessary, we will shortly refers to the above systems as: MAP, MAP-SVC, EV-SVC, JFA, and JFA-D respectively.

Property	**Value**
#Gaussian	**1024**
Features	**38 MFCC (19 + Δ)**
UBM training	**KMeans + 5 ML iterations**
MAP relevance factor	**16.0**
Scoring	**Linear Scoring**
Scoring Normalization	**t-norm, z-norm, zt-norm**

Table 6.1: *UBM data distribution and main properties used in the GMM-UBM baseline system configuration.*

6.2.1. GMM-UBM with standard MAP adaptation [MAP]

The system used as baseline is a GMM-UBM system with linear scoring as that explained in Section 4.3. 1024 multivariate Gaussian of 38 dimensions were used to model MFCC features (19 coefficients + Δ) extracted by using a sliding Hamming window of 20ms and a 50% of overlapping. MEL filters were scaled between 300 and 3000Hz to focus as much as possible to speech voice.

Two gender dependent UBM models were trained via 5 iterations of ML preceded by a K-Means clustering stage, using a total of 6 millions vectors (per gender) extracted from the different databases described in Section 5.2.1 up to Mixer 5 [1].

Regarding session variability compensation, blind classical techniques, CMN, RASTA and Feature Warping were sequentially applied, being the sliding warping window set to 3s. T-norm, z-norm and zt-norm (Section 2.3.3) were applied in order to produce normalized scores being both the t-norm and z-norm cohorts composed over about 250 recordings extracted from Mixer 4 database (SRE'06 evaluation data). Table 6.1 collects the main configuration properties of this system.

6.2.2. MAP adaptation with session variability compensation [MAP-SVC]

As an initial step to evaluate the behaviour of FA dealing with the session variability problem, a first set of experiments was conducted by compensating first order statistics of both training and test recordings. This compensation was carried out by suppressing from the first order statistics the session variability component, $\boldsymbol{Ux}$, estimated via Factor Analysis as is detailed in Section 4.2.3.1.

To accomplish such compensation, two gender-dependent session variability subspaces of 50 eigenchannels trained via PCA namely, U_PCA_50-female and U_PCA_50-male, were used as a starting point (the analysis of the optimum number of eigenchannels and the benefits of using a ML procedure to refine the initial subspaces is carried out later on in this section). Regarding

[1] In terms of NIST evaluations, the background dataset was composed by data belonging up to SRE'06 (included); SRE'08 data was used as evaluation data.

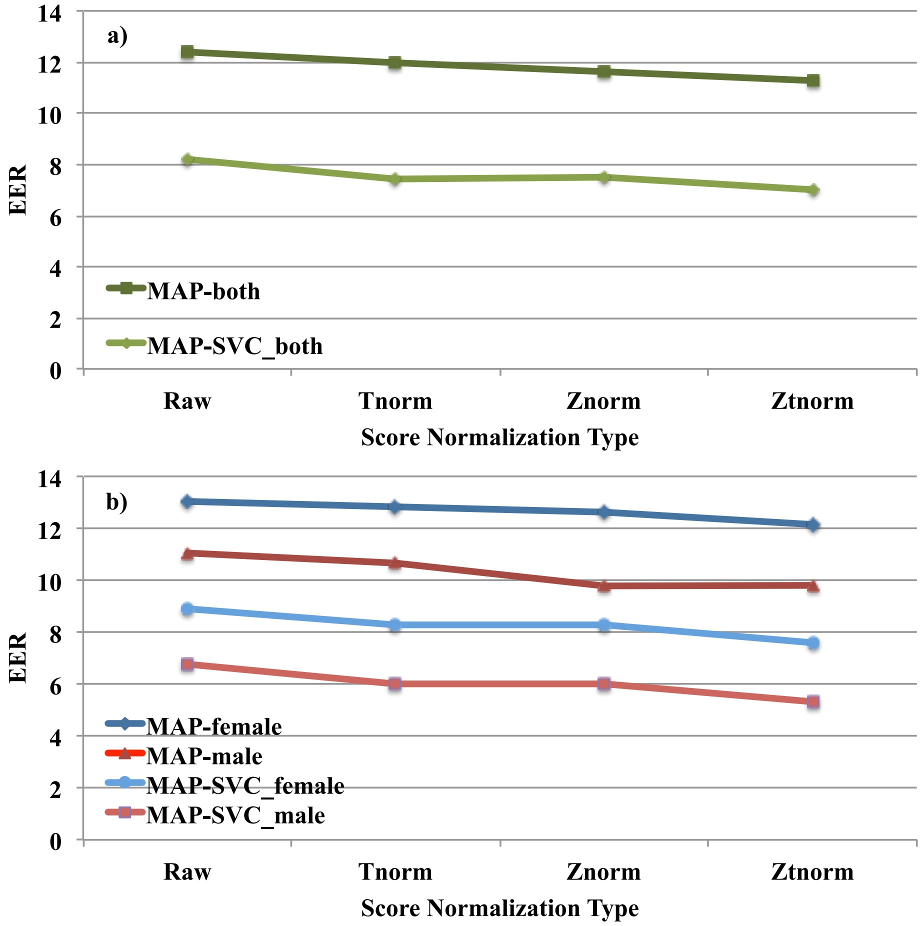

Figure 6.1: *Results on SRE'08 tel-tel conditions for a GMM-UBM system with/without session variability compensation applied to first order statistics, in both training and test recordings. a) pooled scores. b) separate male and female scores.*

the training data composition used to estimate those initial session variability subspaces, a total number of 553 female and 468 male speakers respectively, with an average of 8 recordings per each were used. As a constraint, those speakers were selected to have a minimum of 3 different recordings in a bid of actually capture as much session variability as possible. The common speakers from SRE'06 and SRE'08 evaluations were carefully excluded in order to avoid overfitting scenarios.

The success of this first approach to palliate the session variability issue can be observed in Figure 6.1.a, where obtained results are directly compared to those achieved by the baseline system. A global improvement of 38% is obtained after zt-normalization is applied (35% over raw scores). Separating by gender gains of around 38% and 45% for male and female respectively are achieved as depicted in Figure 6.1.b. Table 6.2 collects EERs and costs for this first set of experiments.

	Equal Error Rate (EER in %)			
System	*Raw*	*Tnorm*	*Znorm*	*Ztnorm*
MAP-both	12.44/0.059	12.03/0.052	11.66/0.052	11.29/0.049
MAP-SVC-both	8.18/0.042	7.44/0.038	7.49/0.041	**7.00/0.037**
MAP-female	13.09/0.064	12.86/0.056	12.65/0.056	12.17/0.053
MAP-SVC-female	8.93/0.044	8.30/0.041	8.29/0.043	**7.59/0.041**
MAP-male	11.07/0.050	10.68/0.042	9.80/0.044	9.82/0.039
MAP-SVC-male	6.79/0.038	6.01/0.030	5.99/0.031	**5.33/0.027**

Table 6.2: *Results on SRE'08 tel-tel conditions for a GMM-UBM system with/without session variability compensation.*

6.2.2.1. Effect of varying the number of eigenchannels.

Experiments above were conducted with an initial session variability subspace of 50 eigenchannels. We explore here the effect of varying the eigenchannels considered in both male and female conditions.

Figure 6.2.a shows the EER evolution from 0 eigenchannels (non-compensated system) to 300 eigenchannels for both genders. As it can be seen a minimum can be established in both genders at around 50 eigenchannels; point from which the EER tends to slightly increase. A major reason for this behaviour can be attributed to the fact that once main session-variability directions have been captured, additional directions considered could account some speaker information. The important point is then to find out this inflection point where the change in tendency occurs. At the eigen-analysis stage this fact can be detected by inspecting the eigenvalues associated to each eigenchannel/direction. Usually, those should present a scenario as depicted in Figure 6.2.b, where the elbow in the curve gives an insight of the appropriate number of eigenchannels. The higher associated eigenvalue the higher confidence to that directions represents session variability (eigenchannels with zero or nearby zero associated eigenvalue should be discarded).

6.2.2.2. Effect of ML refinement in training the session variability subspace.

The effect of using a ML procedure via a EM algorithm to refine the initial session variability subspaces is evaluated in this section. Experiments ranging from 0 (PCA) to 10 EM iterations are showed for both gender in Figure 6.3.a, where scores are zt-normalized.

Even though a slight improvement with respect to the PCA initialization is achieved in both cases after the first iteration, further EM iterations do not yield higher performance. In order to analyse in depth this behaviour, we conducted a similar experiment on the female part, but adding SRE'08 data (different from the evaluation data) within the session variability subspace estimation. A comparison between those experiments are depicted in Figure 6.3.b.

In this case the ML refinement shows to be quite more effective (10% of improvement from 1 to 10 EM iterations versus a 3% without using SRE'08 data). This fact leads to a dual

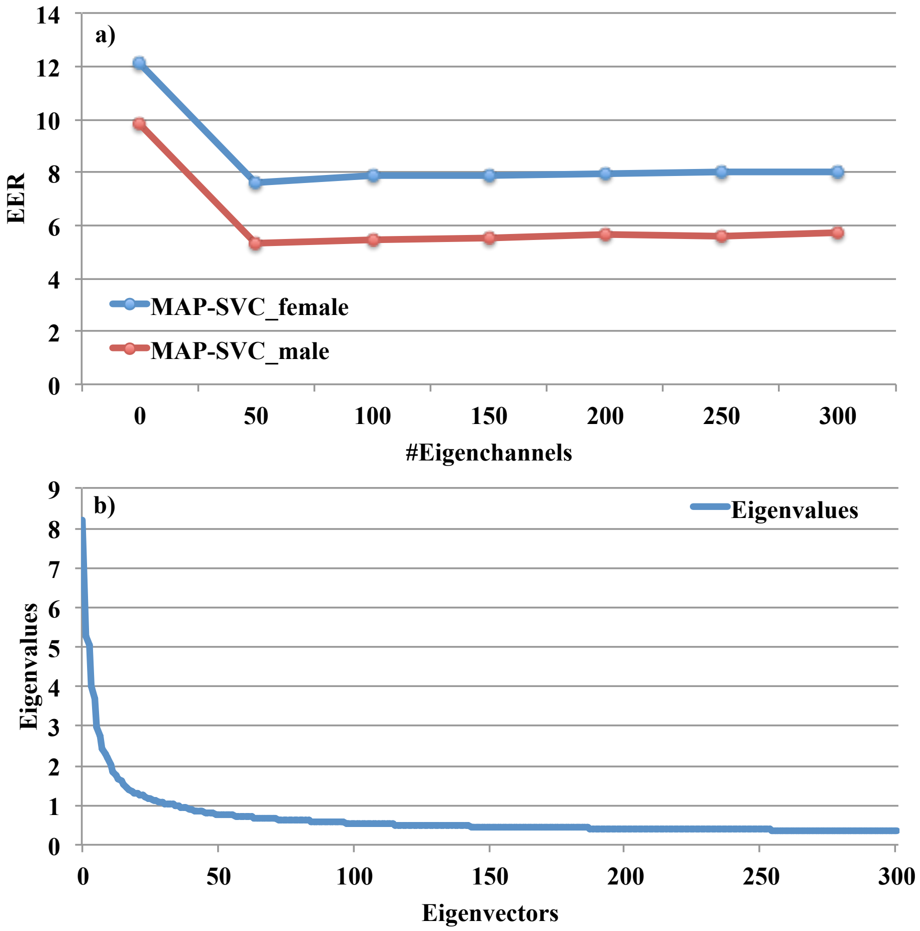

Figure 6.2: a) Evolution of the Equal Error Rate for a session variability compensated system in function of the eigenchannels considered. b) Eigenvalues associated to the eigenchannels estimated for a session variability subspace sorted in descending order.

interpretation. First it warns about the need of having at disposal data as similar as possible to the test/target data in order to maximize the FA performance (this point will be largely treated in Chapter 7); but, second, it is also a call to caution, as a fine data adjustment supported by lower EERs in the ML procedure could lead to develop over-fitting systems, which likely fail in other target scenarios/conditions.

6.2.3. Eigenvoice adaptation with session variability compensation [EV-SVC]

We modify in this section the speaker variability component by substituting the MAP offset, $\boldsymbol{Dz}$, by the eigenvoice adaptation term, $\boldsymbol{Vy}$, in order to evaluate the eigenvoice approach after session variability compensation has been carried out.

As the same manner that performed for the session variability subspace, two initial gender-dependent speaker variability subspaces trained via PCA were obtained. In this case a total

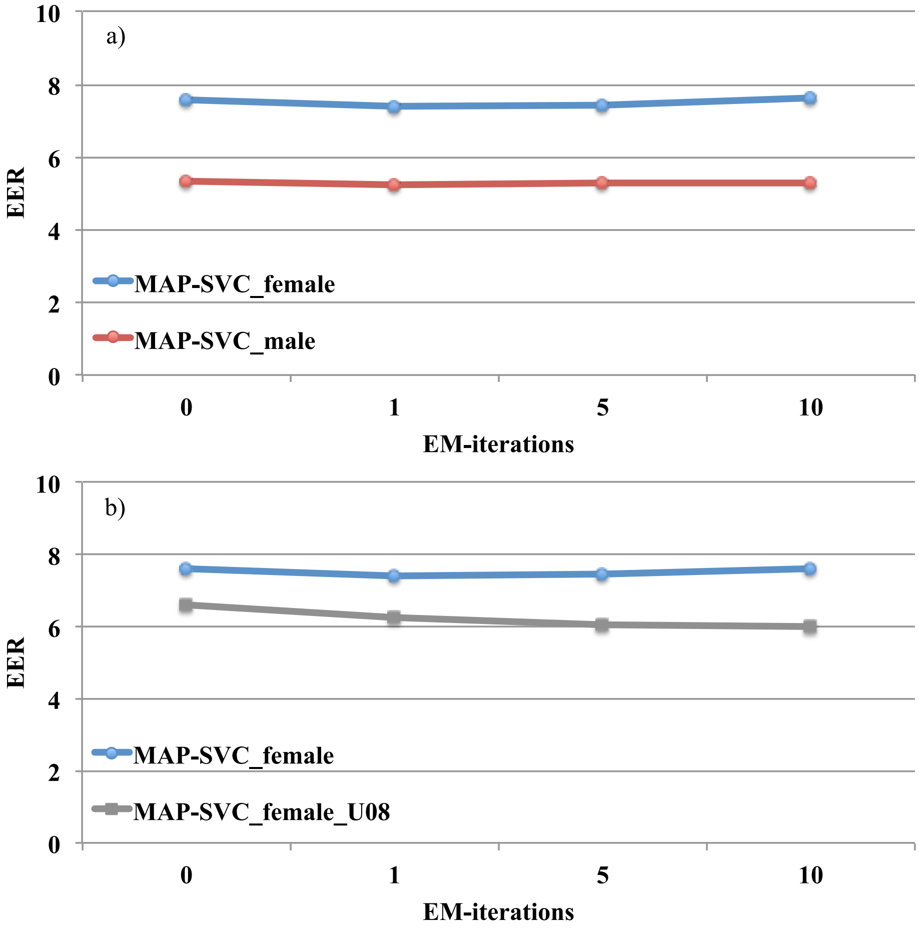

Figure 6.3: *Effect of EM iterations on the ML refinement of the session variability subspace. a) comparison of male and female results in funcion of the EM iterations. b) comparison of the effect of the EM iterations in the female part by using or not data very similar to test data.*

number of 611 female and 580 male speakers with an average of 8 recordings per speaker, and a minimum of 2, were included to train both subspaces respectively; all those recordings were previously session-variability compensated using the initial session-variability subspaces U_PCA_50-female and U_PCA_50-male respectively.

A comparison of results obtained on both female and male conditions for this system considering from 50 to 300 eigenvoices, those obtained by the baseline system and the baseline system with session variability compensation are collected in Tables 6.3 and 6.4 respectively.

We analyse those results below by analysing the following important elements which modify the global behaviour of this system.

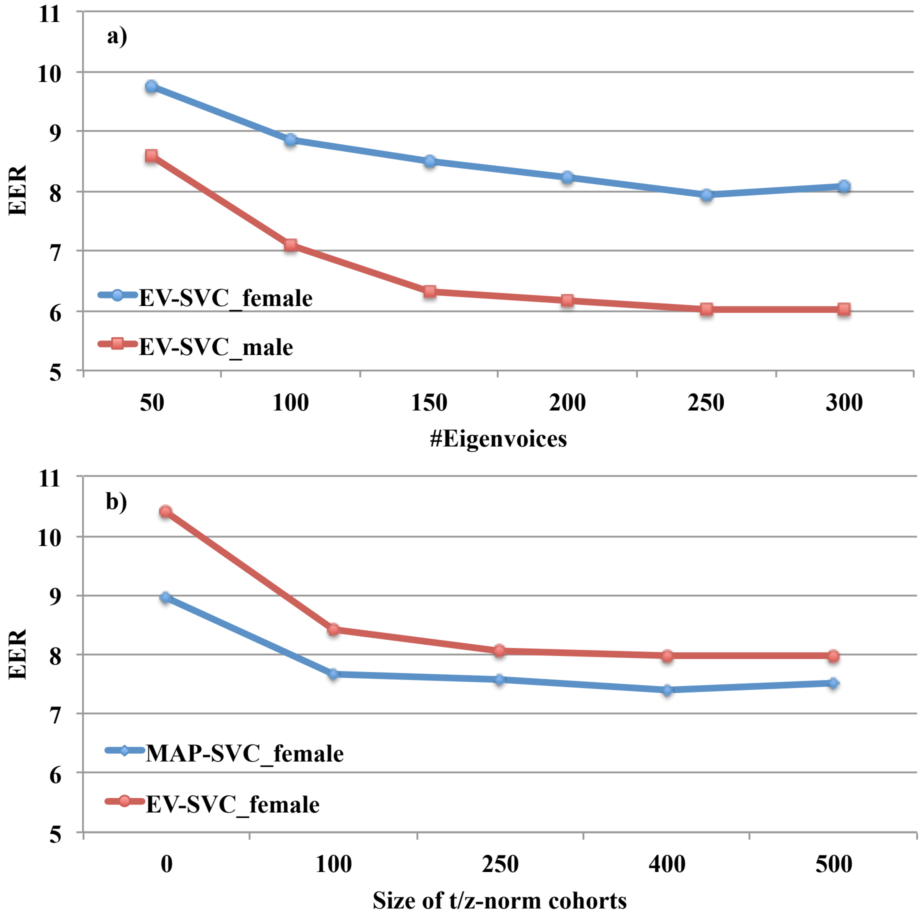

Figure 6.4: *a) Effect of varying the number of eigenvoices in the EV-SVC system. b) Effect of the size of scoring normalization cohort in both the MAP-SVC and the EV-SVC system.*

6.2.3.1. Effect of varying the number of eigenvoices.

As it can be inferred from Tables 6.3 and 6.4, the number of considered eigenvoices largely varies the obtained results, yielding a convergence at around 300 eigenvectors. Improvements of 17.5% and 30% from 50 to 300 eigenvectors for female and male conditions are achieved respectively. This effect, better visualized in Figure 6.4.a where the EER evolution is depicted in function of the eigenvoices number, it is consistent with the fact that, by means of this approach, all the speaker variability is considered to be confined in the speaker variability subspace. Thus, the smaller subspace considered, the larger speaker variability is susceptible to be outside of the subspace and therefore neglected.

A more interesting point is that notwithstanding results converge at some number of eigenvectors (300), achieving an acceptable performance, those do not yield the performance obtained by the baseline system with session variability compensation where classical MAP was used to represent speaker models. This fact highlights that even taking into account a considerable number of eigenvoices, there is some residual speaker information that we are not able to capture as it is not confined in the estimated subspace. On contrary, it is also worth noting that by a moderate loss of performance the speaker variability associated of a new speaker model can be represented by a 300-vector, $\boldsymbol{y}$ instead of a 38912-dimensional one, $\boldsymbol{z}$, as it is the case in MAP adaptation.

Regarding the difference between the number of eigenvoices needed to reach convergence respect to the number of eigenchannels (300 versus 50), it seems clear that we are able to better capture speaker variability rather than session variability. However, this fact should not lead us to conclude that, at general, there exits more variability associated to the speaker than that associated to the session variations, since here the training data used plays an important role. In this case where just telephone data is considered, and being this recorded under same conditions and acquisition protocol, session variation could not be so high as in other different scenarios, such as those usually present for instance in forensic speaker recognition (this point is discussed in Chapter 7).

6.2.3.2. Effect of Scoring Normalization.

As it can be appreciated also in presented results, the effect of scoring normalization has a larger impact in the eigenvoice approach than that produced when compensating session variability. A major reason for this fact lies on that whereas session variability compensation is identically applied in training and test utterances, here just models are shifted by the $\boldsymbol{V}\boldsymbol{y}$ term, resulting on a irregular misalignment which depends on the data used for training the model.

Fortunately, this misalignment can be largely diminished by an appropriate combination of z- and t-norm scoring normalization. Figure 6.4.b shows the EER evolution of the female condition for the eigenvoice-based and MAP-based session variability compensated systems, with respect the size of the t/z-norm cohorts. Results shows that while a relative improvement of 16% is reached by using cohorts of 500 elements in the MAP scheme, same cohorts get a 24% of improvement when eigenvoice adaptation is utilised.

6.2.4. JFA modelling

Once, the session and speaker component has been separately evaluated, we compose the global JFA model in this section. Here, the speaker variability is jointly modelled by the classical MAP adaptation and the component provides by the eigenvoice approach $\boldsymbol{\mu}+\boldsymbol{D}\boldsymbol{z_s}+\boldsymbol{V}\boldsymbol{y_s}$. As in the other approaches evaluated, the session variability compensation is applied in both training and test recordings in the statistics domain.

Tables 6.5 and 6.6 collects the results obtained with the JFA system besides above detailed systems for male and female conditions. Expectedly, the combination of elements outperforms the best results achieved so far in above sections. The fact of considering the prior represented by the speaker variability subspace but also allowing the speaker variability to lie outside of it, gets to join the advantages provided for both approaches. Improvements of 4% and 4.5% respect

	Equal Error Rate (EER in %)			
System (female)	*Raw*	*Tnorm*	*Znorm*	*Ztnorm*
MAP	13.09/0.064	12.86/0.056	12.65/0.056	12.17/0.053
MAP-SVC	8.93/0.044	8.30/0.041	8.29/0.043	**7.59/0.041**
EV-SVC_50	14.27/0.069	12.15/0.059	12.35/0.060	9.77/0.046
EV-SVC_100	11.84/0.060	10.35/0.050	10.37/0.050	8.86/0.041
EV-SVC_150	11.05/0.056	9.88/0.046	9.81/0.047	8.50/0.040
EV-SVC_200	10.82/0.054	9.52/0.045	9.36/0.047	8.22/0.040
EV-SVC_250	10.66/0.053	9.29/0.045	9.25/0.046	7.94/0.040
EV-SVC_300	10.41/0.052	9.24/0.044	9.21/0.045	**8.07/0.039**

Table 6.3: *Results on SRE'08 female tel-tel condition for MAP, MAP-SVC and EV-SVC systems. The number of eigenvoices considered in the EV-SVC system ranges from 50 to 300. A number of 50 eigenchannels was considered regarding the session variability subspace.*

	Equal Error Rate (EER in %)			
System (male)	*Raw*	*Tnorm*	*Znorm*	*Ztnorm*
MAP	11.07/0.050	10.68/0.042	9.80/0.044	9.82/0.039
MAP-SVC	6.79/0.038	6.01/0.030	5.99/0.031	**5.33/0.027**
EV-SVC_50	13.56/0.071	10.76/0.052	11.07/0.053	8.58/0.040
EV-SVC_100	10.76/0.060	8.89/0.043	8.87/0.045	7.10/0.032
EV-SVC_150	9.66/0.056	8.11/0.038	8.19/0.041	6.32/0.030
EV-SVC_200	9.35/0.053	7.67/0.036	7.88/0.038	6.17/0.030
EV-SVC_250	8.96/0.051	7.48/0.036	7.57/0.038	6.01/0.029
EV-SVC_300	8.81/0.049	7.33/0.034	7.48/0.037	**6.01/0.028**

Table 6.4: *Results on SRE'08 male tel-tel condition for MAP, MAP-SVC and EV-SVC systems. The number of eigenvoices considered in the EV-SVC system ranges from 50 to 300. A number of 50 eigenchannels was considered regarding the session variability subspace.*

to the MAP-SVC system; and a 40%, 48% respect to the MAP systems, are achieved for the female and male conditions respectively.

6.2.5. JFA modelling with D trained on data

Finally, in order to complete the analysis, a final step consisting of training on data the residual matrix $\boldsymbol{D}$ as described in Section 4.2.1.1 was performed. To this aim a separate set of 105 and 91 speakers involving a total of 325 and 273 recordings from Mixer 4, were used to estimate diagonal $\boldsymbol{D}$ female and male matrices respectively.

As it can be seen in Figure 6.5 the effect of training $\boldsymbol{D}$ matrix on data slightly improve the results obtained by JFA, although there is not a significant difference over using the term $\boldsymbol{Dz}$

	Equal Error Rate (EER in %)			
System (female)	*Raw*	*Tnorm*	*Znorm*	*Ztnorm*
MAP	13.09/0.064	12.86/0.056	12.65/0.056	12.17/0.053
MAP-SVC	8.93/0.044	8.30/0.041	8.29/0.043	7.59/0.041
EV-SVC	10.41/0.052	9.24/0.044	9.21/0.045	8.07/0.039
JFA	9.03/0.044	8.22/0.040	8.18/0.041	**7.29/0.039**

Table 6.5: *Results on SRE'08 tel-tel (female) condition for the four evaluated systems, MAP, MAP-SVC, EV-SVC and JFA.*

	Equal Error Rate (EER in %)			
System (male)	*Raw*	*Tnorm*	*Znorm*	*Ztnorm*
Baseline-male	11.07/0.050	10.68/0.042	9.80/0.044	9.82/0.039
MAP-SVC	6.79/0.038	6.01/0.030	5.99/0.031	5.33/0.027
EV-SVC	8.81/0.049	7.33/0.034	7.48/0.037	6.01/0.028
JFA	7.25/0.041	6.26/0.030	6.01/0.031	**5.09/0.025**

Table 6.6: *Results on SRE'08 tel-tel (male) condition for the four evaluated systems, MAP, MAP-SVC, EV-SVC and JFA.*

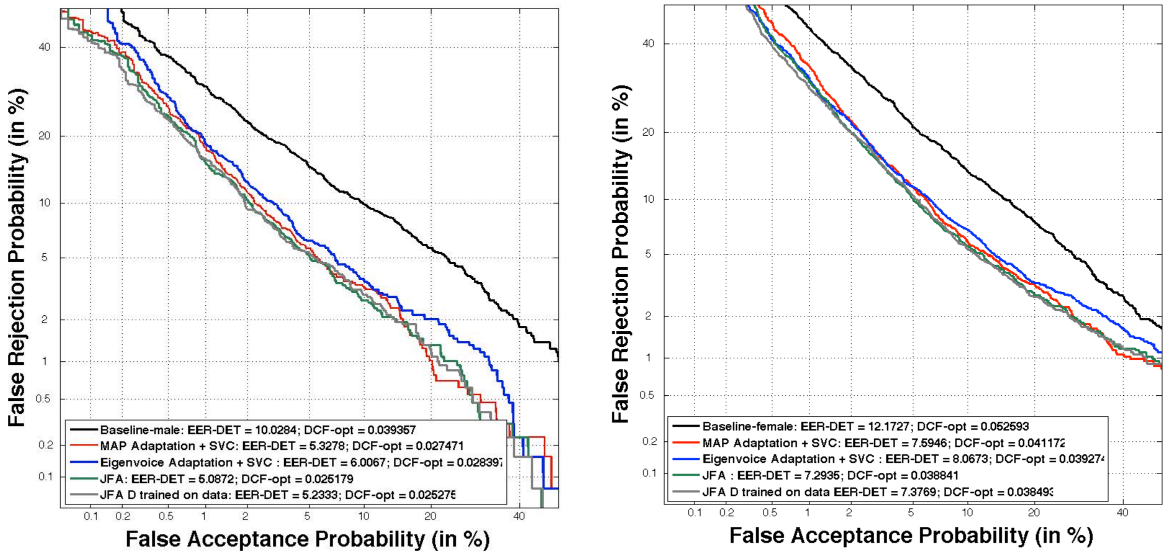

Figure 6.5: *Results on SRE'08 tel-tel conditions for all the systems considered namely, MAP, MAP-SVC, EV-MAP, JFA and JFA with D matrix trained on data. a) male condition, b) female condition*

derived from MAP. Note in that sense, that as explained in Section 4.2.1.1, the actual key of using to get a well estimated $\boldsymbol{D}$ lies on training this in a separate way of $\boldsymbol{V}$, as in a ML process the higher number of free parameters of $\boldsymbol{V}$ could overshadow $\boldsymbol{D}$ as pointed out in Section 4.2.1.1.

6.3. Joint Factor Analysis applied in SLR Systems

Once proved its outstanding performance when dealing with session variability in SV, JFA is evaluated, in this section, in the context of SLR. However, unlike the above section, this analysis will not stop at the evaluation of FA in the proposed acoustic systems, but it is going beyond to show how well FA can be incorporated as a part of a global SLR system where multiple and very different systems are combined.

To this aim, the ATVS SLR system presented at the last NIST Language Recognition evaluation (NIST LRE 2009) [Gonzalez-Dominguez *et al.*, 2010d] will serve as an excellent example in order to i) evaluate the performance and potential of fusion of JFA-based systems, and ii) establish a fair comparison of JFA acoustic systems and high-level systems in challenging conditions of variability and duration.

6.3.1. ATVS SLR submitted to NIST LRE 2009

The ATVS SLR includes most of the development and contributions in the field of SLR collected in this Dissertation and it achieved an excellent 2^{nd} rank position in the challenging open-set 30s condition (core condition) of the NIST LRE 2009 evaluation. It consisted of four different combinations of acoustic and phonotactic subsystems. Those being:

- ***ATVS4*** is a phonotactic-only system, fusion of the 10 PhoneSVM systems (Section 2.5.2).
- ***ATVS3*** is a fast and reliable GMM system with linear scoring and session variability compensation applied in the statistic domain as that denoted in the SV section as MAP-SVC (Section 6.2). We refer here this system as Factor Analysis GMM Linear Scoring (FA-GMM-LS) to become evident the type of modelling (GMM) and the use of FA. This system is designed to optimize the computational time but with a high level of recognition performance.
- ***ATVS2*** consisted of a fusion via an anchor-model back-end [Gonzalez-Dominguez *et al.*, 2010d] of all the ATVS's acoustic (FA-GMM and SVM-FA-SV) and phonotactic (PhoneSVM) systems, as shown in Figure 6.6.
- ***ATVS1*** (primary) is a fusion of ATVS2 with primary system from other participant (TNO, leaded by prof. David Van Leeuwen), where the latter consisted of a fusion of six acoustic systems: three GMM-SVM and three FA-GMM-LS.

6.3.2. Configuration of spectral systems

A parameterization consisting of 7 MFCCC with CMN-Rasta-Warping concatenated to 7-1-3-7 SDC-MFCCs was used for spectral systems.

According to the data type, two UBMs namely UBM_{CTS} and UBM_{VOA} with 1024 Gaussian were trained. Data from CallFriend, LRE'05 and train part of LRE'07 was used for training

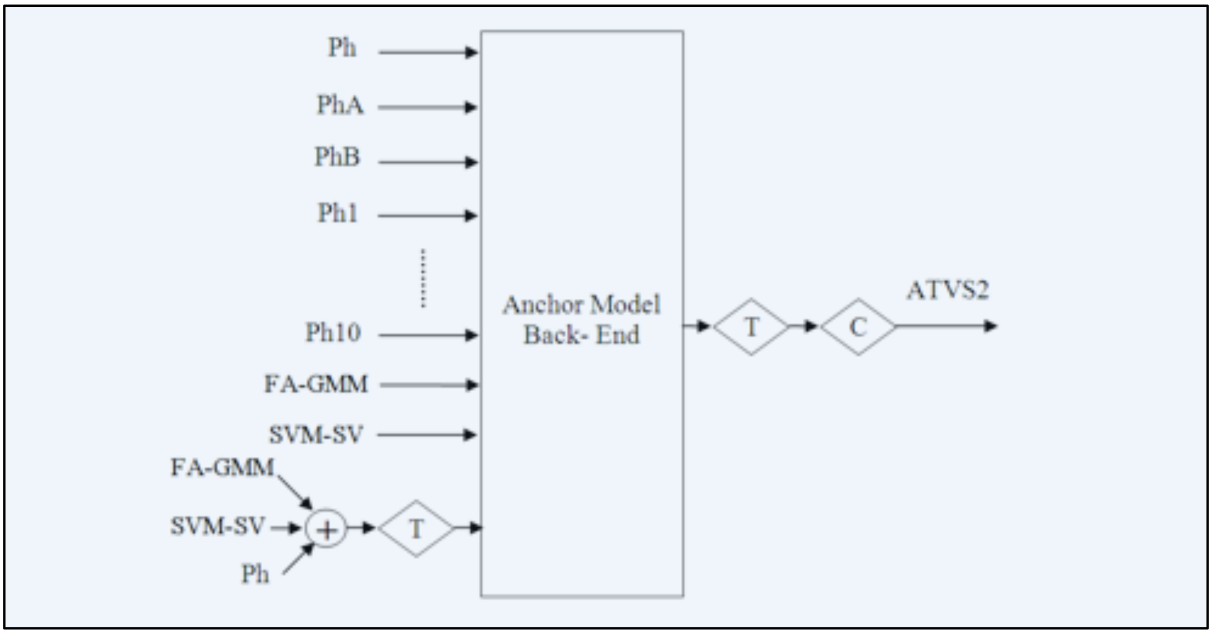

Figure 6.6: *Fusion scheme for ATVS2 submitted system.*

UBM_{CTS}, while the training of UBM_{VOA} was composed by VOA development data provided by NIST. Distribution per hours of this training is as follows. A total of 38.5 hours was used in UBM_{CTS} training, including about 2.75 hours per 14 available languages. For UBM_{VOA} a total number of 31.2 hours balanced on 1.42 hour per 22 languages was used (IndianEnglish was not included due to data scarcity for this language).

Further, two different FA-GMM-LS systems were developed by using above UBMs. Two session variability subspaces matrices were trained from CTS and VOA data respectively, U_{CTS} and U_{VOA}. We found this approach to outperform the approach where mixed data (CTS,VOA) is processed to train a unique session variability subspace. In this work, session variability subspaces were trained via EM algorithm after a PCA initialization as described in Chapter 3 and only top-50 eigenchannels were taken into account turns out in a $CF \times 50$ (C components and F dimensions) dimension matrix. In order to train the session variability subspaces, a large amount of data was used. U_{CTS} was trained with a total number of 350 hours by using 600 segments of about 150 seconds per the 14 languages available; while U_{VOA} was trained with 550 hours, using 600 segments of about 150 seconds as well but of the 22 languages available. Data distribution for training UBMs and session variability subspaces is summarized in Table 6.7.

Compensated statistics via Factor Analysis by using U_{CTS} and U_{VOA} as described in 4.2.3.1 and 4.2.3.2 were used on the SVM-SV system.

6.3.3. Configuration of high-level systems

The phonotactic ATVS system was a fusion of 10 different Phone-SVM subsystems (Ph1 to Ph10) as described in Section 2.5.2. Ph1 to Ph7 use phonetic tokenizers developed by ATVS and

Prior model	Databases	#Languages	#Hours/language	Total
UBM_{CTS}	$CallFriend, LRE05, TrainLRE07$	14	2.75	38.5
U_{CTS}	$CallFriend, LRE05, TrainLRE07$	14	25	350
UBM_{VOA}	VOA	22	1.42	31.2
U_{VOA}	VOA	14	25	550

Table 6.7: *Distribution of data used for training Universal Background Models and Session Variability Subspaces.*

Ph8 to Ph10 use phonetic tokenizers trained with Hungarian, Czech and Russian data respectively [1]. The ATVS phonetic tokenizers are based on Hidden Markov Models (HMMs), trained with HTK [Young *et al.*, 2006] and later transformed to be used by the SPHINX [Lee *et al.*, 1990] speech recognition engine for faster recognition. The phonetic HMMs are three-state left-to-right models with no skips, and the output pdf of each state is modeled as a weighted mixture of 20 Gaussians. The acoustic processing is based on 13 Mel Frequency Cepstral Coefficients (MFCCs) (including $C0$) and velocities and accelerations for a total of 39 components, computing a feature vector each 10 ms and performing Cepstral Mean Normalization (CMN). The languages of the phonetic decoders from Ph1 to Ph6 and the corresponding corpora used for training are English (with the corpus with ELDA catalogue number S0011), German (S0051), French (S0185), Arabic (S0183 + S0184), Basque (S0152) and Russian (S0099)[2]. Ph7 uses a phonetic decoder in Spanish trained on Albayzin spanish speech database [Moreno *et al.*, 1993] downsampled to 8 kHz, which contains about 4 hours of high-quality phonetically labelled speech. Once the speech segment has been transformed into a sequence of recognized phonetic tokens (with any of the phonetic decoders), this sequence is used to estimate count-based 1-grams, 2-grams and 3-grams, pruned with a probability threshold, resulting in about 40,000 n-grams. These are rearranged as a feature vector, which is taken as the input of an SVM that classifies the test segment as corresponding (or not) to one language. PhoneSVMs are combined in different ways to obtain different front-end systems. Each PhX system consisted of 22 VOA and 14 CTS models trained separately. Channel dependent t-norm is the last stage of those phonotactic front-ends.

6.3.4. Fusion and calibration

Input vectors to the fusion systems anchor model based back-end have dimension 216 (36 ATVS models -14CTS+22VOA- x 6 component systems) while primary is 438 as scores from the TNO site are added. Back-end t-norm was design as channel-independent (VOA+CTS), while calibration was duration-dependent. Anchor model training was 90/10 bootstrapped while calibration training was bootstrapped with 80/20 using available training data. A channel independent t-norm (models from VOA and CTS) stage was applied for scoring normalization.

[1] These have been developed and made available for research purposes by the Speech Processing Group at Faculty of Information Technology, Brno University of Technology.

[2] www.elda.org.

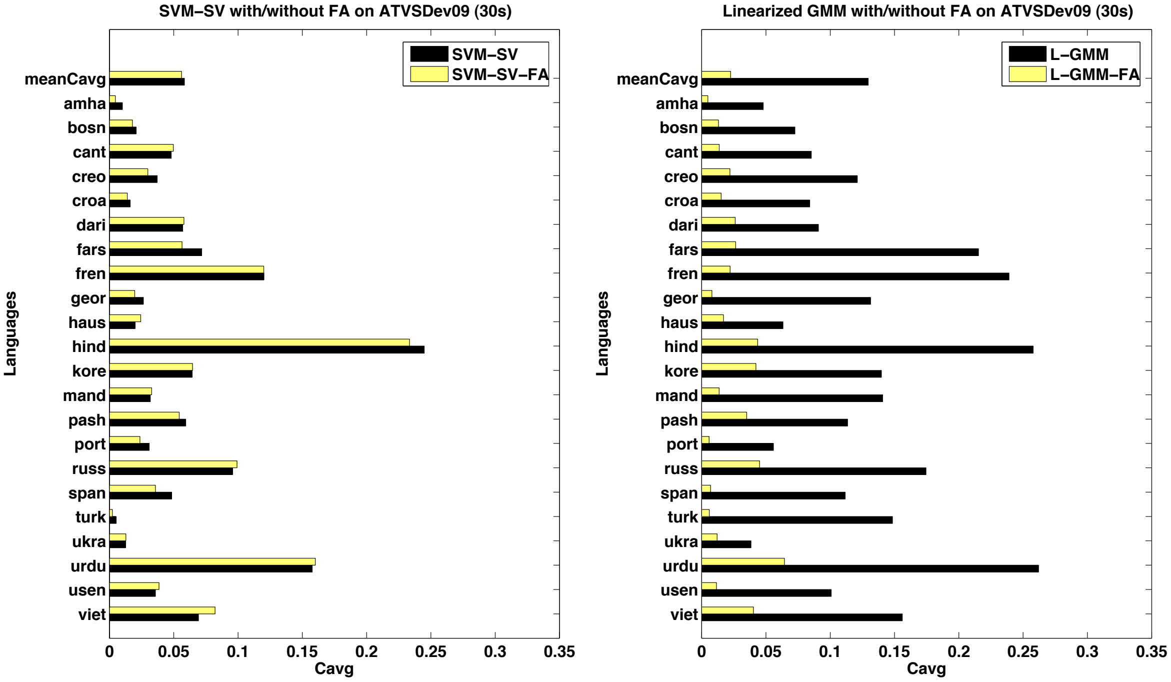

Figure 6.7: *Effect of session variability compensation on SVM-SV and FA-GMM-LS systems. Results on ATVS-Dev09 using VOA models and* U_{VOA}.

	Equal Error Rate (EER in %)					
	ATVS-Dev09			LRE'09		
	03s	10s	30s	03s	10s	30s
$ATVS1$	16.50	**6.48**	**1.56**	17.97	**7.87**	**3.71**
$ATVS2$	**16.17**	7.25	2.02	**17.92**	8.39	4.26
$ATVS3$	20.37	10.30	3.25	21.93	10.65	5.67
$ATVS4$	18.80	9.41	3.73	20.87	10.81	6.55

Table 6.8: *ATVS submitted systems performance (meanCavg x 100) on development and evaluation datasets.*

LRE'09 considered three different nominal durations for the test segments: 3, 10 and 30 seconds of speech. The same individual subsystems were used to perform language recognition tests for the different durations. However, calibration was trained specifically for the estimated different durations. As the calibration was applied after the back-end, a single score for each test segment was used, and scores from all the speech types (VOA, CTS) were pooled for training. Thus, all the available scores for each duration from each target language were used to train logistic regression, and the linear transformation obtained was used to calibrate the scores from testing data.

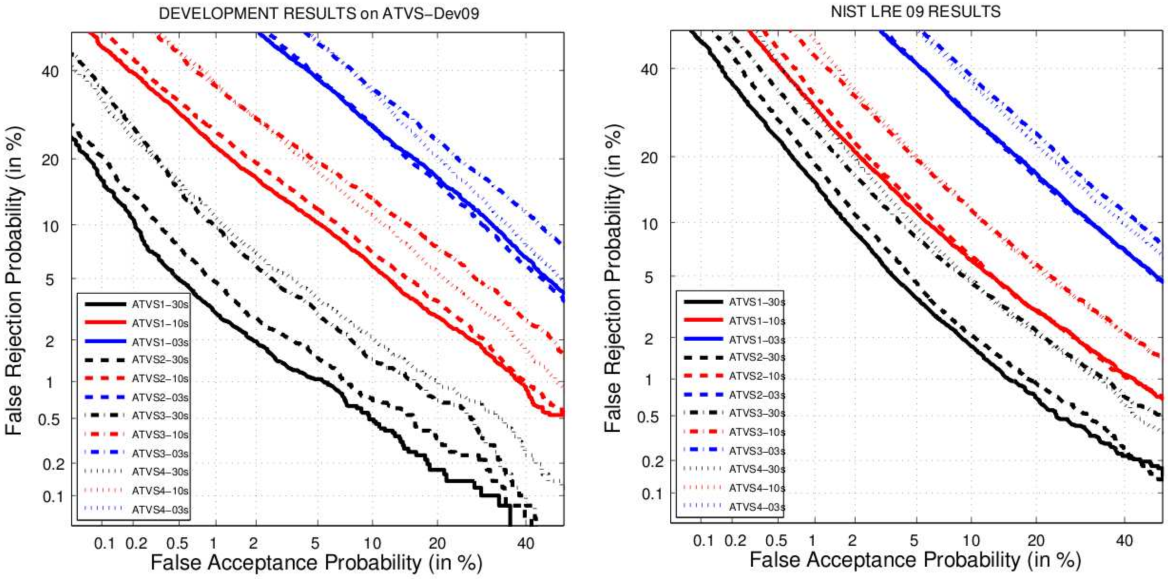

Figure 6.8: *Pooled DETs per ATVS submitted systems on development (ATVS-Dev09) and evaluation (LRE'09) per all target test segment durations (3, 10 and 30 seconds).*

6.3.5. Performance of JFA-based spectral systems

The need of proper session variability compensation is showed in Figure 6.7 where both spectral systems, FA-GMM-LS and SVM-SV are assesed with and without compensation via factor analysis on ATVSDev09. Results shows that channel compensation via FA is crucial in GMM modelling performance, getting an improvement of about 82% in $meanC_{avg}$ terms. Also, system SVM-SV take advantage of this compensation but to a lesser extent (4%). This effect appears due to differences in SVM and GMM modelling. In GMM, target languages models, trained with huge amount of data, are far shifted with respect UBM reference model after even a single MAP adaptation. This mean shifting includes not only information belonging to the language but session variability found in the training database which it is mainly independent of the languages. This leads to models that are growing strongly affected by session variability effects. On the contrary, the SVM exhibits a higher robustness to this problem due to its ability to estimate an hyperplane separating target single utterances models against all non-target ones. However, once session variability compensation is applied, both GMM and SVM-SV system, as well as the fusion of both clearly outperforms the performance achieved without session variability compensation via FA.

6.3.6. Performance of global system

The performance of ATVS submitted systems is summarized in Figure 6.8 for development (ATVSDev09) and evaluation (LRE'09) tests. Here, the discrimination per each system (ATVS1-4) and test segment duration (3, 10 and 30 seconds) is showed in a pooled DET curve. Several global observations can be immediately extracted. Firstly, the good behaviour of the anchor models fusion scheme introduced is justified as being ATVS1 (fusion of systems) the system with lower error rates. The effect of test segment duration in system performance is also highlighted and it affects in a similar manner to both, acoustic and high level systems. Further, a slight degradation in the evaluation results with respect to development ones is showed. This degradation performance, common to all participants, is usually due to the database mismatch (this problem is discussed in Chapter 7) among the development and testing databases, and is a common effect in LRE's. Table 6.8 summarizes this information in terms of *meanCavg* (mean of Cavg per language) per system, evaluation dataset and test segment durations. It is also worth pointing out that acoustic systems outperform phonotactic ones except for short durations, and this with a much smaller computational complexity, but fusion of both kind of systems improve results, which encourages the use of multilevel approaches for language recognition.

In more detail, Figure 6.9 compares systems performance per target language. Again, results are presented on both, development and evaluation, but only for 30s test segment duration. Analysis shows the varying degrees of recognition difficulty among the different target languages. In the same way, Figure 6.10 presents in detail the effect of test segment duration per language for our primary system (ATVS1).

6.4. Summary

This chapter empirically supports FA as an effective and efficient tool to deal with the session variability problem. In the first part, a wide set of experiments conducted on the telephone part of the challenging NIST SRE'08 evaluation have largely proven that its application to explicitly modelling both speaker and session variability lead to a major benefit of systems perform. Specifically, an outstanding global improvement of 40% and 48% for female and male conditions is achieved respect a non-compensated classical GMM-UBM system.

Those results have been affirmed in the second part of this chapter, where FA has been used to deal with the session variability problem in the context of the NIST LRE'09. In this case, FA has been proven to be a critical part in the development of robust and accurate SLR systems; getting improvements up to 82% over a baseline GMM system without session variability compensation, as well as enhancing the acoustic SVM-SV system via the original contribution presented in this Dissertation. Further, it has been also demonstrated that the use of FA does not hinder the additional gains obtained by fusing very different systems such as the acoustic and high-level systems presented; showing an excellent behaviour in the fusion strategy.

Equal error rates (EERs) and associated costs (DCF) to all the experiments presented through this chapter are included in Appendix C. Also, results on SLR detailed per language are included in that appendix.

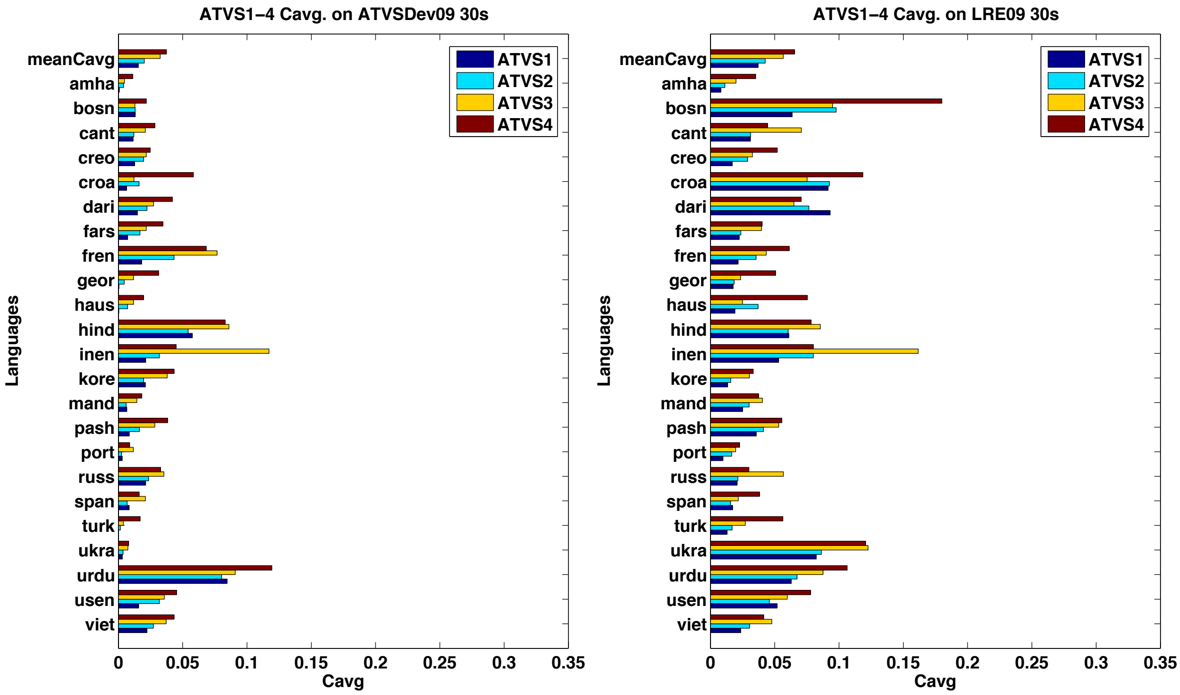

Figure 6.9: Comparision of ATVS submitted systems on both, development (ATVS-Dev09) and evaluation (LRE'09) datasets for 30 seconds test duration segments.

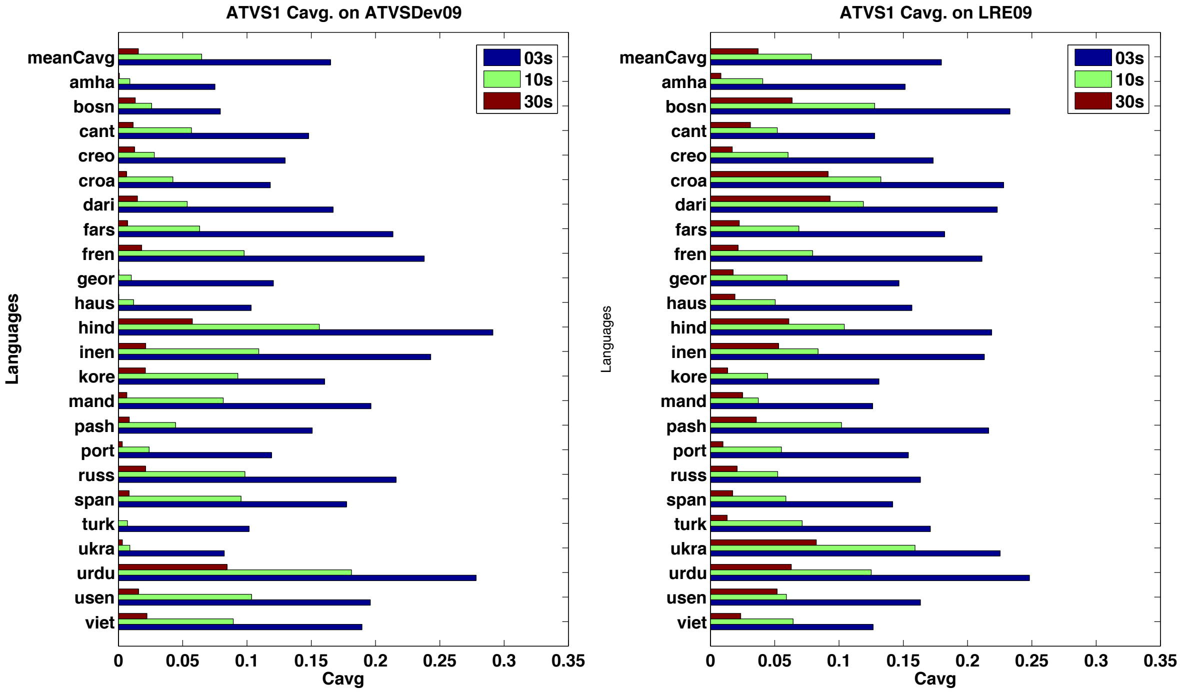

Figure 6.10: ATVS primary system performance on both, development (ATVS-Dev09) and evaluation (LRE'09) datasets (3, 10 and 30s).

Chapter 7

Factor Analysis in challenging SV and SLR scenarios

This chapter explores the use of FA approaches applied to palliate major challenges in the deployment of real SV and SLR systems in the framework of forensic speaker recognition.

7.1. Introduction

Apart from the session variability problem, two major issues can be identified to significantly degrade SV and SLR systems hindering their deployment in real applications. These are, i) *the short durations*, that is, to have at disposal small amounts of speech in either the training or the test phase and ii) *the database mismatch*, understood this as the variation in the conditions between the dataset used for training and fitting a system (referred to as background or development database) and the data used in real-world operational conditions (known as evaluation or operational database). The latter might be considered as a session variability problem taken to the extreme. But, the fact that each database is usually subject to very different types of session variations, turns the database mismatch problem into an enormous complication for the FA techniques. A major reason that lies on the variability subspaces could not faithfully represent the real session variability encountered in the test/operational data if the training material is far from this in terms of session variations.

These two problems frequently occur in *forensic speaker recognition* [Gonzalez-Rodriguez *et al.*, 2007b; Ramos, 2007], mainly because the limitation in the availability of real-casework databases for system tuning, and also because the conditions of the speech in real-world forensic recording are extremely variable.

The purpose of this chapter is to explore several forms based on Factor Analysis intended to palliate as much as possible these two major issues framed into adverse scenarios as those encountered in forensics tasks. On the one hand, it will be emphasized that FA may be a double-edge sword if a depth understanding of the faced problem and the FA theory is neglected. For instance, a non adequate estimation of the variability subspaces could lead the system to fail.

Original contributions of this chapter includes advances in the following lines:

1. Collecting public real-casework databases [Ramos *et al.*, 2008]

2. Exploring new ways to deal with the database mismatch problem via Factor Analysis [Gonzalez-Dominguez *et al.*, 2010a]

7.2. Facing the database mismatch problem via FA

From a statistical point of view, in FA, the variability subspaces ($\boldsymbol{U}$, $\boldsymbol{V}$) act as a strong prior, since the target data variability, both session and speaker, is supposed to be mostly constrained within them. As a consequence, an important issue in the successful application of the FA model is appropriate training of the subspace transform matrices. Ideally, these matrices should accurately represent the types of inter- and intra-speaker variations expected within and between recording sessions. For this purpose, a suitable dataset that accurately represents the conditions of the target domain is essential.

Unfortunately, this requirement for suitable data cannot be satisfied in all situations. Forensic speaker recognition is an area that gives us a wide range of examples of this situation where database mismatch problem is regularly present. This fact is mainly due to two factors. Firstly, despite the efforts made to collect new databases [Ramos *et al.*, 2008], the available data is still very limited. Secondly, real world forensic recordings tend to be extremely variable, making a case-by-case treatment necessary in most situations. In those cases where only a limited amount of data is available, the estimation procedure described above leads to poorly estimated variability subspaces since the real variability in target domain is not sufficiently represented.

The work in [Gonzalez-Dominguez *et al.*, 2010a] considers the problem of data availability for training the FA subspaces, and the appropriate estimation of these subspaces, under the idea of dealing with the limited data problem by exploiting data from a data-rich domain in the session subspace estimation procedure. This approach pursues a dual goal. First, to obtain a more robust estimation procedure by adding large amounts of data. Secondly, to incorporate certain 'session' variability characteristics not present in the limited available target domain data but that could appear in the target domain. The three techniques explored in [Gonzalez-Dominguez *et al.*, 2010a] for combining information from a data-rich domain and limited target domain data are presented in the remainder of this section.

7.2.1. Joining Matrices

A simple way to combine different session variability subspaces is to join session variability subspaces estimated on different datasets. This process is carried out by simply stacking the session variability directions estimated in each one of them in a bigger subspace. This approach has the major advantage that subspaces can be treated and trained independently. From a practical point of view, this property is highly desirable because it allows us to keep a well-trained reference subspace trained on accumulated data that can be refined by simply appending new session variability information from new domains. On the other hand, it has several shortcomings. Firstly, it is necessary to restrict the size of each contributing subspace, loosing potentially useful directions of variability, in order to keep the overall size of the joined subspace relatively small as stipulated by the principles of FA. Second no particular emphasis is placed on the target domain data because all the directions play an equal role in the new subspace. Finally, even the main directions of session variability will tend to be poorly estimated for the target domain if there is severely limited data as the subspaces are estimated independently.

7.2.2. Pooled Sufficient Statistics

As an alternative to stacking two independently trained subspaces, the subspace estimation can also be supplemented with the data-rich telephone set simply by estimating a completely new session subspace. This time, estimation is performed by pooling all data. An obvious advantage of this method is that the estimation is performed using a substantial amount of data, making it potentially more robust. Unfortunately, there is no means of preventing the supplementary set dominating the estimation and having the biggest effect on the directions of variability.

7.2.3. Scaling Statistics

Based on the fact that we are usually most interested in the session variability present in a specific domain (the closest to the target domain conditions), it is reasonable to think that somehow these data should become more important in the subspace estimation procedure. Moreover, we should be able to get some advantage by using all the data available together rather than separately. The approach presented here is based on giving a specific weight to each dataset in the training session variability subspace with a dual purpose. First, allow the estimation procedure to learn from a broader set of data leading us to more robust subspace estimation, and second to highlight the type of data which is considered most important. This second point is especially necessary when not enough data of this type is available and the variability presented could be overshadowed by the other types. Specifically, first order statistics supervector extracted from each utterance is scaled by a previous fixed weight depending on the dataset to which it belongs. Thus, the matrix of first order statistics for training utterances $\boldsymbol{F}$, input in the EM procedure for training the variability subspace take the following form:

$$\boldsymbol{F} = [\alpha \boldsymbol{F}_{tgt};\ (1-\alpha)\boldsymbol{F}_{bckg}] \tag{7.1}$$

where $\boldsymbol{F}_{bkg}$ and $\boldsymbol{F}_{tgt}$ are the matrices whose columns are the first order statistics of utterances belonging background data similar to target data and other background data available respectively. More generally, this could be extend to:

$$\boldsymbol{F} = [\alpha_1 \boldsymbol{F}_1;\ \alpha_2 \boldsymbol{F}_2;\ ...;\ \alpha_N \boldsymbol{F}_N] \tag{7.2}$$

with $\sum_i^N \alpha_i = 1$ and N different background sets.

	Equal Error Rate (EER in %)			
	1conv4w/	1conv4w/	1mic/	1mic/
U **Training**	1conv4w	1mic	1conv4w	1mic
$U = 0$	5.97	8.20	7.81	11.03
$dTel$	3.49	4.31	3.95	6.79
$dMic$	5.80	5.19	5.30	6.64
$dMic_{10}$	5.99	5.69	5.50	7.51
$dMic_5$	5.93	6.06	5.72	8.07
$dMic_3$	5.99	6.13	5.72	8.33

Table 7.1: *Performance under restricted MIC data conditions in U training.*

	Equal Error Rate (EER in %)			
	1conv4w/	1conv4w/	1mic/	1mic/
U **Training**	1conv4w	1mic	1conv4w	1mic
$dMic \mid dTel$	3.41	3.63	3.12	5.14
$dMic_{10} \mid dTel$	3.55	3.72	3.32	5.43
$dMic_5 \mid dTel$	3.55	4.15	3.63	5.74
$dMic_3 \mid dTel$	3.55	4.31	3.54	6.03

Table 7.2: *Performance using the joint matrices subspaces estimation approach.*

7.2.4. Results

As a starting point of this study, the effect of using restricted datasets in order to estimate a session variability subspace was analysed. For this purpose, a baseline JFA without eigenvoices as that presented in 6.2.2 was evaluated using the differing restricted microphone datasets described in Section 5.4.1.1 as training data for the low-rank session matrix U. The results in Table 7.1 summarise the performance statistics of these restricted subspace training data experiments. Studying these results, it can be seen that when microphone data scarcity is simulated in the development stage (i.e. the amount of training data for $\boldsymbol{U}$ is reduced), system performance is degraded significantly. It is clear from these results alone that data availability for training the channel subspace has a large impact on overall performance.

For comparison purposes, results for a baseline system that does not include session compensation ($\boldsymbol{U} = 0$) were also included in Table 7.1. It is obvious from the results that incorporating session compensation leads to significant improvements in performance across all train/test conditions. Interestingly, even when the data used to estimate the session subspace is mismatched to the conditions (channel type) of the evaluation trials, the inclusion of session compensation always results in an improvement. A session matrix estimated using purely telephone data reduces the error rates in the *1mic-1mic* condition. Similarly, a session matrix estimated using

	Equal Error Rate (EER in %)			
	1conv4w/	1conv4w/	1mic/	1mic/
U **Training**	1conv4w	1mic	1conv4w	1mic
$dMic + dTel$	3.73	3.54	3.43	4.97
$dMic_{10} + dTel$	3.61	3.72	3.43	5.47
$dMic_5 + dTel$	3.42	3.88	3.66	5.78
$dMic_3 + dTel$	3.49	4.12	3.76	6.19

Table 7.3: *Performance under restricted microphone data conditions when statistics are pooled with devTel.*

	Equal Error Rate (EER in %)			
	1conv4w/	1conv4w/	1mic/	1mic/
Scaling (α)	1conv4w	1mic	1conv4w	1mic
–	3.49	4.12	3.76	6.19
0.6	3.46	4.15	3.74	5.95
0.7	3.55	4.15	3.67	**5.82**
0.8	3.80	4.49	3.32	**5.82**
0.9	4.30	4.64	3.78	6.50

Table 7.4: *Performance using scaled statistics during ML estimation. Results using 3 mic utterances per speaker.*

microphone data for telephone based trials provides some benefits over no session compensation at all. Expectedly, the best performance is achieved when the session subspace is trained using appropriate data (eg. $dMic$ used for *1mic-1mic*).

Experiments were then performed to examine whether the data rich sources - in this case the telephone data - could be used alongside the restricted data in the estimation of the session variability subspace $\boldsymbol{U}$, in order to improve the estimation and in turn, the overall performance. The first approach considered for this task was the joint subspace approach as outlined in Section 7.2.1. A new session variability subspace was generated simply by stacking two independently trained session subspaces, one estimated using the $dTel$ set and the other using the target domain data $dMic$. For this combination strategy, the top 50 and 20 eigenchannels from $\boldsymbol{U}_{dTel}$ and $\boldsymbol{U}_{dMic}$, respectively, were used to create a 70 eigenchannel joint subspace [1]. The performance using both the full and restricted datasets are presented in Table 7.2.

Comparing the results in Table 7.2 with those in Table 7.1, it can be seen that supplementing the subspace training data with telephone data has a positive effect across nearly all evaluated tasks. While this effect seems obvious in those conditions where telephone data is involved, it

[1] An analysis of the eigenvalues for the microphone data showed a very rapid decline in values in comparison to the telephone data. For this reason, a reduced number (20) of dimensions were retained.

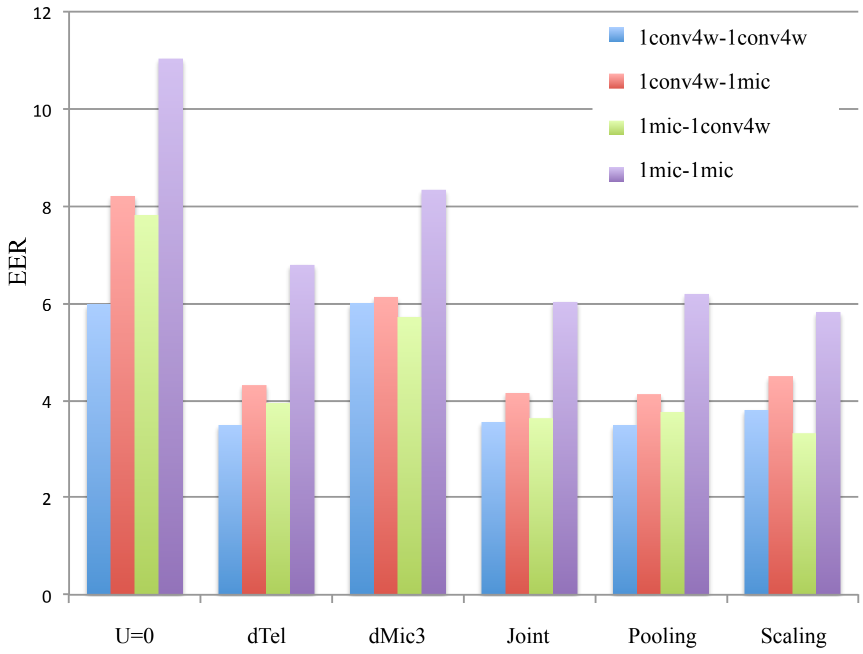

Figure 7.1: *Performance of the several proposed methods for the diverse conditions considered, 1conv4w-1conv4w, 1conv4w-1mic, 1mic-1conv4w and 1mic-1mic.*

is worth noting that even in the case condition *1mic-1mic*, including telephone data alongside the available microphone data in the subspace development stage is clearly beneficial. This suggests that it is possible to account for some session variability even in very apparently different acoustic subspaces. The biggest gains from supplementing the target domain microphone data with telephone data were observed when the target domain (microphone) data was restricted. For the most restricted training scenario $dMic_3$, a relative improvement of 28% resulted for the *1mic-1mic* condition when $dMic_3$ was supplemented using $dTel$.

As outlined in Section 7.2.2, a new subspace can also be estimated by pooling the statistics from both the data-rich set and target domain set. Results using this pooling method are presented in Table 7.3. An interesting point to highlight here is the case where the full microphone dataset $dMic$ is available for subspace estimation. In this case, an improvement in performance over the joint matrix technique is observed for the case condition *1mic-1mic*. When less target domain (microphone) data is available for the subspace estimation, we see that the effectiveness of the session compensation is reduced when pooled statistics rather than stacked matrices are used. This suggests that for the pooled approach, the subspace estimation is being overwhelmed by the larger quantity of telephone data, and is not able to best utilise the available (but restricted) target domain data.

Finally, the method proposed in Section 7.2.3, where more emphasis is placed on data from the target domain by performing a scaling of the statistics during subspace estimation, was evaluated. Results in Table 7.3 show the performance using various scaling weights, α. For

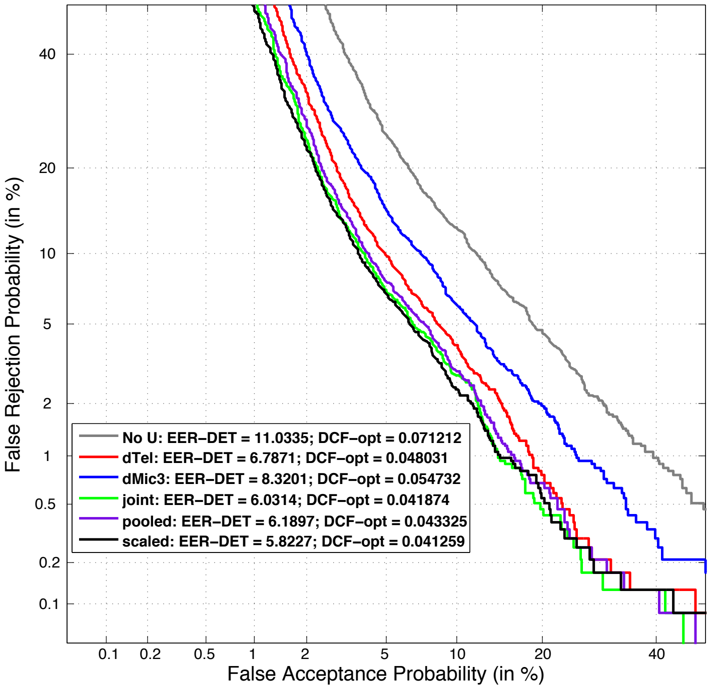

Figure 7.2: *DET curves of the several proposed methods for the diverse conditions considered, 1conv4w-1conv4w, 1conv4w-1mic, 1mic-1conv4w and 1mic-1mic.*

these experiments, the closest simulation of real forensic applications, where only 3 utterances per speaker in $dMic$ was made available for subspace estimation was studied ($dMic_3$). It can be seen from these results, that in general, placing a larger weighting on the $dMic_3$ statistics results in an improvement in performance over straight pooling (unweighted). For the case condition *1mic-1mic*, a scaled statistics estimation results in a 6% relative improvement in EER over the straight pooling.

Figure 7.1 shows a final comparison of the considered estimation strategies for the session subspace, evaluated on the 1mic-1mic condition with only a limited amount of target domain data available ($dMic_3$). This chart clearly demonstrates the benefit of session compensation, but also the problems associated with a direct estimation of the subspace on a small dataset. Better results are achieved when subspace estimation is performed using the data-rich $dTel$ rather than $dMic_3$ alone. Importantly though, benefits result from supplementing the $dMic_3$ with other data. Each of the strategies for combining the two sets in estimation give improvements over either alone. The joint estimation approach using stacked subspaces achieves a better result than a straight pooling of the data, however, this trend can be reversed by introducing a simple scaling of the statistics during estimation. By weighting the target domain data more heavily during estimation, the best performance out of the considered approaches is achieved.

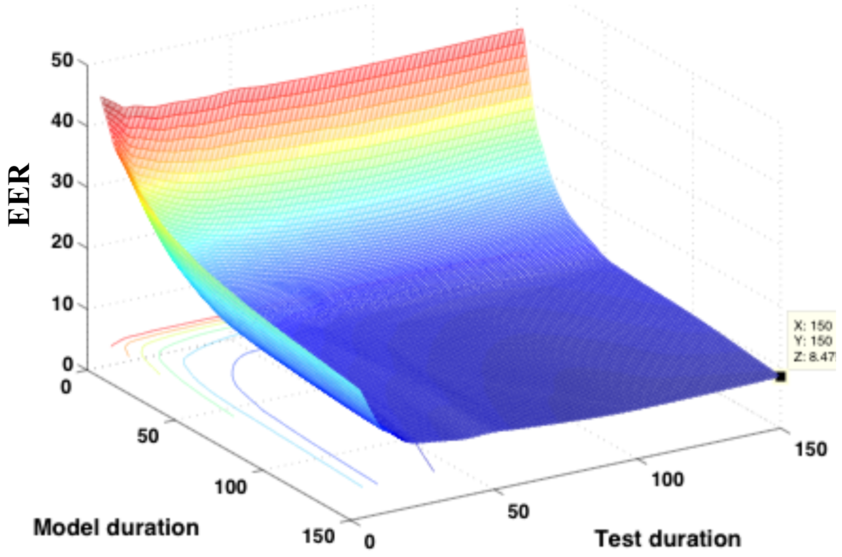

Figure 7.3: *Representation of the Equal Error Rate in the telephone part of SRE08 in function of the training and test recordings duration.*

7.3. Facing the short duration problem via FA

As expected, one of the major degrading variability factors concerning the SV or SLR systems is the length of the speech utterances involved in enrolment and testing processes [Pelecanos *et al.*, 2004]. However, although performance with extremely short utterances are of interest for the scientific community [Perez-Gomez *et al.*, 2010; Vogt *et al.*, 2008b], nowadays a scant amount of research has been conducted for compensating the effects of speech duration variability. This is mainly due to the configuration of tasks in NIST SRE, where the length of the enrolment and testing utterances present small variation in a single condition. Nevertheless, there is a wide range of scenarios where the length of the utterance involved in the recognition process may vary, e.g. forensic applications. Figure 7.3 depicted this effect in the telephone part of SRE'08, where the EER is presented in function of the length of both training and test recordings, which were artificially reduced from 150s to 10s.

A detailed study of the behaviour of FA when dealing with short durations was conducted in Vogt *et al.* [2008b]. Expectedly, experiments demonstrated that, as utterance lengths for both training and test utterances was reduced, the effectiveness of JFA was also diminished; but more interesting, it was observed that the inclusion of the session variability compensation term $\boldsymbol{Ux}$, when dealing with very short utterances ($\leqslant$ 20s) led to a significant degradation performance. Further experiments conducted in [Vogt *et al.*, 2008a] demonstrated that a match duration between the testing recordings and the development recordings, used to estimate $\boldsymbol{U}$, partially fixed this gap of performance (even in the case of development recordings were reduced to match testing recordings). Further experiments in the area identified as one of the major factors of those

Phonetic Class	**Phone**
Vowel	A: E e: i i: O o o: u u: y y: :2 _2
Occlusive	B b: d d_ d_: g k k: p t t: t1 t1:
Fricative	f h h1 S S: s s: v x Z z z:
Affricate	dz tS tS_ ts ts_
Nasal	F J J: m m: N n n:
Aproximant	j j:
Lateral	l l:

Table 7.5: *Broad phonetic decomposition.*

effects the phonetic content within a recording. In typical 150s NIST conversational recordings a reasonable coverage of the phonetic variability is found, and discarded when estimating the session variability subspace. However this is not the case when dealing with short utterances, and the phonetic content could largely vary among different recordings. To counteract this effect in [Scheffer *et al.*, 2009] the session variability was proposed to be disentangled into inter-session variability and intra-session variability, by estimating two different session variability subspaces considering the variations between sessions and those produced within same sessions respectively. However, although results slightly improved the match recording estimation, this strategy requires and additional computational cost, as well as the need of further development material.

We extend here those results conducted in [Scheffer *et al.*, 2009] by carefully analysing the impact of the phonetic-class composition within the training material used for estimating the speaker variability subspace $\boldsymbol{V}$ rather than the session one. To this aim, first a set of broad phonetic classes will be defined to second be included/excluded into the speaker variability subspace training material.

7.3.1. Broad phonetic classes defined

As the base phone recognizer for phone conditioning the Hungarian phone recognizer made available by Brno University of Technology (BUT) was used. One of the reasons for choosing this particular language among those available for this recognizer is that the phone set is very large, including 56 different phones. This makes it easier to make phone conditioning language-independent as the phone set covers most, if not all, possible broad phonetic classes.

Table 7.5 presents these 56 phones[1] classified into 7 broad phonetic classes defined according to the manner of articulation. Also, we evaluate some joining classes as presented in Table 7.6 were considered.

[1] Phones are represented in SAMPA (Speech Assessment Methods Phonetic Alphabet) notation.

Phonetic Class	**Phone**
AproxLateral	j j: l l:
Voiced	A: E e: i i: O o o: u u: y y: :2 _2 F J J: m m: N n n: j j: l l: B b: d d_ d_: g
Unvoiced	k k: p t t: t1 t1: f h h1 S S: s s: v x Z z z:

Table 7.6: *Broad phonetic classes originated by joining other phonetic classes.*

7.3.2. Effect of phonetic composition in the speaker variability subspace

This section presents a detailed study based on the use of the different phonetic classes described above in order to train the speaker variability subspace.

The study is focused on training a robust eigenvoice subspace ($\boldsymbol{V}$) given excerpts belonging to a single phonetic class or a combination of them, while the session variability matrix $\boldsymbol{U}$ is kept constant. In that sense a ”Pure-eigenvoice” model is used rather than a ”Classical MAP + eigenvoices”

Unlike [Scheffer *et al.*, 2009] where a number of four phonetic classes were considered, a total number of eight phonetic classes have been analyzed and different combination via the concatenation method described in [Scheffer *et al.*, 2009] has been analysed. This ”concatenation” method outperformed other proposed method in previous studies, including [Scheffer *et al.*, 2009].

Tables 7.7 and 7.10 present the results obtained using each single phonetic classes and different combinations of them when dealing with the 10s-10s task defined in Section 5.4.1.1. As it can be observed, best performance is reached by using all the speech or just using vowels. On the other hand, affricate phones show to be achieve the worst performance.

Those results motivated different compositions of the speaker variability subspace from different number of eigenvoices trained with different excepts belonging to the proposed phonetic classes. Specifically, five speaker variability subspaces as showed in Table 7.9. Results from those matrices are collected in Table 7.10.

The following conclusions can be extracted from the above results:

- Vowel is the most discriminant single phonetic class in order to train the eigenvoice subspace while affricate is the least discriminant.

- Similar results can be obtained using just the vowels phonetic class instead all the speech (see Table 7.7 vs Table 7.10).

- Using 200 eigenvectors outperforms in general the use of a smaller number of eigenvoices (100, 150)

- Removing the affricate phonetic class when composing the eigenvoice subspace shows a slight improvement with respect to include it.

	Equal Error Rate (EER in %)				
#Eigenvoices	*Vowel*	*Occlusive*	*Fricative*	*Affricate*	*Nasal*
100	23.77/0.088	27.70/0.093	26.60/0.093	33.03/0.098	26.46/0.092
150	23.03/0.087	26.88/0.093	26.74/0.092	32.92/0.097	26.91/0.092
200	**22.75/0.086**	26.74/0.092	26.58/0.092	32.62/0.098	27.42/0.091

Table 7.7: *Results on 10s-10s SRE'08 conditions by estimating the speaker variability subspace constrained to different phonetic classes.*

	Equal Error Rate (EER in %)			
#Eigenvoices	*All*	*Aproximant/Lateral*	*Voiced*	*Unvoiced*
100	22.89/0.085	26.46/0.092	23.43/0.086	25.78/0.092
150	22.65/0.085	26.02/0.092	22.92/0.086	25.49/0.091
200	**22.24/0.084**	25.49/0.091	22.75/0.085	25.69/0.091

Table 7.8: *Results on 10s-10s SRE'08 condition by estimating the speaker variability subspace constrained to different phonetic classes.*

- None of the results achieved either using phone classes alone or in combination improve the results attained in the baseline system (using all speech). However the use of vowels reach very close results without the need of considering all the speech.

7.4. Summary

This chapter has addressed main issues associated to the deployment of SV and SLR systems in real-world applications as forensic speaker recognition, namely the *database mismatch* and the *short durations* problems; their negative impact in terms of system performance and how can be FA strategies modified in order to mitigate that degradation performance.

The successful application of FA techniques is highly dependent on the proper estimation of session variability as represented by the variability subspaces. The problem of applying FA in situations where a scant amount of data similar to the expected operating conditions is available, has been largely analysed.

A range of experiments using the microphone condition of the well-known NIST SRE 2006 database and protocol were initially conducted exploring the effect of reducing the quantity of available development data. These experiments clearly demonstrated the importance of a well-estimated session variability subspace as using poorly matched telephone data or heavily restricting the available microphone development data resulted in significantly increased error rates. In these situations, current estimation procedures lead to poorly estimated subspaces and consequently far from optimal FA performance.

Phonetic Class	$\boldsymbol{V}_1$	$\boldsymbol{V}_2$	$\boldsymbol{V}_3$	$\boldsymbol{V}_4$	$\boldsymbol{V}_5$
Affricate	35	0	0	0	0
Aprox/Lateral	35	40	35	35	0
Fricative	35	40	50	35	0
Nasal	35	40	35	35	0
Occlusive	35	40	50	35	0
Unvoiced	0	0	0	0	100
Voiced	0	0	0	0	100
Vowel	35	40	50	70	0
Total	210	200	200	210	200

Table 7.9: *Eigenvoices used to compose five different speaker variability subspaces.*

System	**Equal Error Rate (EER in %)**
$\boldsymbol{V}_1$	23.16/0.085
$\boldsymbol{V}_2$	23.44/0.085
$\boldsymbol{V}_3$	23.03/0.086
$\boldsymbol{V}_4$	22.61/0.084
$\boldsymbol{V}_5$	22.62/0.084

Table 7.10: *Results on 10s-10s SRE'08 condition for five different phonetic-class composition of the speaker variability subspace.*

To deal with this problem, several methods were explored to combine different variability information obtained from different sources of data, including joining subspace matrices, and pooling estimation statistics. These techniques are based on the idea that variability present in different databases can be exploited in order to provide more robust subspace estimates. Experiments with these methods show that a suitable method of combining information from both the target domain and a data-rich development domain can be very useful in the restricted data scenarios, particularly if emphasis can be placed on the limited available target domain data.

On the other hand, a wide analysis of the phonetic class composition of the training material used for estimating the speaker variability subspace has been conducted. The phonetic content variability of short-durations has been identified as one of the main hurdles in the development of adequate FA systems. In that sense, it has been demonstrated that similar results in short durations can been obtained by taken into account only vowels that those obtained with all the phonetic content.

Chapter 8

Conclusions and Future Work

This chapter has addressed the problem of session variability in automatic speaker and language recognition, their negative impact in systems performance, and how this can be mitigated via new methods based on Factor Analysis. After a detailed vision of the state-of-the-art composition in both speaker and language recognition fields, a study of Factor Analysis modelling in the context of Latent Variables Models has been conducted, deep analysing the principles and mathematical grounds which sustain it. Diverse forms to build and incorporating Factor Analysis in state-of-the-art acoustic systems has been then explored and detailed with the main goal of yielding robust but also efficient speaker and language recognition systems. A wide set of experiments in both speaker and language challenging tasks has been conducted to give empirical support to the use of Factor Analysis dealing with the session variability problem. Besides, two primes challenges in the deployment of "real-world" speaker and language recognition systems, namely the *database mismatch* and the *short durations* problems, has been analysed, deep exploring possible counteracts based on Factor Analysis. Inherent in the different chapters, contributions of this thesis has been detailed and properly evaluated.

8.1. Conclusions

Chapter 1 introduced the basics of automatic speaker and language recognition systems framed into the biometric systems family, identified the problem of session variability as one the main cause of system performance degradation and exposed then the motivation of this Dissertation. The research contributions originated from this Thesis were also enunciated in this first chapter.

The most relevant works which conformed the state-of-the-art in speaker and language recognition field, previously to the incorporation of Factor Analysis methods to palliate the session variability problem, is summarized in Chapter 2. A review of the different modules which compose a speaker or language recognition system, from the speech signal to the final taken decisions about identity, as well as the most successful approaches in the literature associated to each of those modules were described. Also most successful techniques to counteract the session variability problem before the appearance of FA were described in this chapter.

Chapter 3 deeply analysed the mathematical grounds of the Factor Analysis model in the context of Latent Variable Models, as well as its extension from a single Gaussian, as use to be referred in the literature, to a mixture of Gaussians as it is incorporated in speaker and language acoustic recognition systems. In this chapter also a chronological review of the use of subspaces in order to represent variability in related fields, such as face or speech recognition, to its use in speaker and language recognition was conducted; analysing in this manner, common links among different techniques which arose in related fields as well as identifying the specificities of the speaker and language tasks.

Chapter 4 detailed how can be Factor Analysis integrated into the the well-known GMM and SVM acoustic systems, detailing different strategies to incorporate Factor Analysis at three different levels in the architecture of those kind of systems, namely the feature, the model and the statistics domain. A special effort was focused on yielding robust but also efficient systems, rescuing for the literature efficient ways and possible simplifications incorporated to the original Factor Analysis model that which an acceptable loss of performance, achieve very efficient recognition systems. In that sense, this chapter ends detailing efficient recipes to build speaker and language recognition systems based on Factor Analysis in both speaker and language recognition task. Those algorithms are supported by several and novel contributions conducted during the research which has originated this Dissertation.

The databases and experimental protocols used later on in Chapters 6 and 7 are described in Chapter 5. The protocols adopted in this Dissertation are those established by NIST in the speaker and language recognition evaluation series. This fact ensures that any of the experiments presented through this Thesis can be either fairly compared with other proposed techniques or replicated by other researches to a major benefit of the area. Specifically, the tel-tel and 10s-10s condition extracted from SRE'08 were used to evaluate the Factor Analysis and proposed methods in speaker verification, and the challenging LRE'09 was utilised to assess the performance of language recognition systems presented. Also, a simulated adverse speaker recognition scenario was simulated from data belonging to SRE'05 and SRE'06.

The experimental part of this Dissertation started in Chapter 6, where a wide set of experiments were conducted to evidence Factor Analysis as an effective and efficient tool to deal with the session variability problem. Experiments conducted on the telephone part of the challenging NIST SRE'08 evaluation largely proved that the Factor Analysis application to explicitly modelling both speaker and session variability lead to a major benefit of systems perform. Specifically, an outstanding global improvement of 40% and 48% for female and male conditions was achieved respect a non-compensated classical GMM-UBM system. Those great results were then confirmed in the context of language recognition, in the LRE'09 evaluation, where FA was proved to be critical in the development of accurate acoustic language recognition systems. In that sense, improvements up to a 82% were achieved over a baseline GMM system without session variability compensation. Also the global and complete ATVS Biometric Recognition group system presented to the LRE'09 evaluation and which obtained an excellent 2^{nd} rank in the core 30s open-set condition, was detailed; evaluating in this manner most of the strategies presented in Chapters 2 and 4, and showing how all the identities levels within the speech signal can be jointly exploited to reach high performance recognition results.

Chapter 7 addressed main issues associated to the deployment of SV and SLR systems in ”real-world” applications as forensic speaker recognition, namely the *database mismatch* and the *short durations* problems; their negative impact in terms of system performance and how can be FA strategies modified in order to mitigate that degradation performance. Several novel contributions to deal with the database mismatch were evaluated and a deep study of the phonetic content of recording used to estimate the speaker variability subspace in the context of the short durations problems was conducted.

In summary, the main conclusions that can be extracted and have been highlighted through in this Thesis are:

- The session variability problem is one of the main causes of system performance degradation in both automatic speaker and language recognition systems.

- The session variability should be treated as continuous rather than in a discrete way, since it is the result of the conjunction of a numberless of sources which cannot be properly quantified.

- Most of the session and speaker/language variability associated to a given recording can be explained by a reduced number of variability directions and corresponding weights. Those variability subspaces can be previously estimated from large amount of data and be used as strong priors in the modelling of speaker/language or session variability. This process fits with the theory of Latent Variable models, specifically with Factor Analysis modelling.

- The use of a complex mathematical framework as Factor Analysis is not incompatible with the development of efficient systems. FA can be incorporated in an properly manner in speaker and language recognition, leading to robust and very efficient systems.

- Factor Analysis should not be used as either a closed formula or as a *black box* to deal with session variability. A deep understanding of this modelling strategy as well as the target data (data in operational conditions) nature is needed in order to achieve significant results. A non-adequate use of FA could lead the global system to fail.

- The database mismatch and the short durations problem still being a challenge for speaker and language recognition systems, and although FA can be useful to deal with them, further research is needed to adequate its use in those scenarios.

Main contributions and results are:

- The compilation of the mathematical grounds of Factor Analysis, from its original formulation to its use in speaker and language recognition systems.

- The efforts made in achieving robust but also efficient Factor Analysis based acoustic systems for both speaker and language recognition

- The novel methods explored and proposed to incorporate Factor Analysis into speaker and language recognition systems.

- The study of the main problems in the deployment of speaker and language recognition systems in ”real-world” scenarios and the novel methods proposed to mitigate their negative impact in performance by using Factor Analysis.

8.2. Future Work

A number of research lines arise from the work conducted in this Thesis. Among then, following ones are highlighted:

- Exploring new forms of Factor Analysis applied to palliate the session variability problem and modelling speaker variability. Although JFA has demonstrated to be very effective the, new improved versions could achieve better performance results. Recent strategies as Total Variability [Dehak *et al.*, 2011] o Probabilistic Linear Discriminant Analysis [Kenny, 2010] are an example of those evolved FA methods.

- Including a full-Bayesian treatment on Factor Analysis methods. Although, it has been noted that speaker and channel factors (latent variables) has been well modelled rather than make use of a point estimate, other model parameters such as the variability subspaces are estimated via a Maximum Likelihood procedure. This fact could lead to problems as the over-fitting and may be solved via a full Bayesian treatment of all the model parameters involved in FA [Bishop, 2007]. Recent work in that sense is accomplished in the field of speaker recognition in [Villalba and Brummer, 2011].

- Exploring new forms to palliate the *short durations* problem via Factor Analysis. Although the problem of short durations is still being a challenge in the field, a scant amount of research has been conducted in the area [Perez-Gomez *et al.*, 2010], specially when durations of the recordings vary from one trial to the next, as usual occurs in tasks as forensic speaker recognition. Recent studies as this conducted in [Mandasari *et al.*, 2011] endorse this research line.

- Exploring new forms to palliate the *database mismatch* problem via Factor Analysis. Identified as one of the main challenges when dealing with ”real-world” systems [Ramos *et al.*, 2008], the database mismatch problem is an open research line where Factor Analysis has not been completely exploited. Initial works have already been conducted in [Gonzalez-Dominguez *et al.*, 2010a; Senoussaoui *et al.*, 2010].

- Studying the application of those session variability compensation schemes to other biometric recognition traits such as fingerprint or signature verification. The session variability is not a specific problem of speaker and language recognition but in general affects to any biometric trait, as the results of different aspects of their acquisition processes. For instance, the use of different sensors in the capture of fingerprints or signatures includes different session variations that could be faced via Factor Analysis.

- Considering the application of Factor Analysis in automatic speech recognition systems. In the line of the above point, the speech recognizers are strongly affected by a number of variability sources (note that in this case even the inter-speaker variability is considered as a nuisance source). Pioneer experiments in the area have been already conducted in [Burget *et al.*, 2010; Povey *et al.*, 2010].

- Exploring discriminative approaches based on Factor Analysis. Discriminative approaches as SVM has been proved to be very effective in both speaker and language recognition. The idea of derive a discriminative FA model rather than the generative presented in this Thesis is an open line in the development of Factor Analysis based systems. Initial research in this line endorses this future line [Glembek *et al.*, 2011].

- Combining traditional and automatic speaker/language recognition approaches. It is widely agreed upon the scientific community [Gonzalez-Rodriguez *et al.*, 2007b] that combining automatic and classical speaker/language recognition approaches [Kunzel, 1994; Rose, 2006] should lead to a major benefit of the recognition systems. Pioneer studies in this field has showed excellent results [de Castro *et al.*, 2009; Gonzalez-Rodriguez, 2011].

Appendix A

Factor Analysis

Note for the following proofs, that by definition of the FA model

$$\mathbb{E}[\boldsymbol{z}] = \boldsymbol{0} \tag{A.1}$$

$$Cov(\boldsymbol{z}) = \mathbb{E}[\boldsymbol{z}\boldsymbol{z}^T] = \boldsymbol{I} \tag{A.2}$$

$$\mathbb{E}[\boldsymbol{\epsilon}] = \boldsymbol{0} \tag{A.3}$$

$$Cov(\boldsymbol{\epsilon}) = \mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T] = \boldsymbol{\Psi} \tag{A.4}$$

and also that as $\boldsymbol{z}$ and $\boldsymbol{\epsilon}$ are considered independent

$$Cov(\boldsymbol{\epsilon}, \boldsymbol{z}) = \mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{z}^T] = \boldsymbol{0} \tag{A.5}$$

Proof of $p(x \mid \boldsymbol{z}) \sim N(\mu + \boldsymbol{L}\boldsymbol{z}, \boldsymbol{\Psi})$ (equation 3.8)

$$\begin{aligned}
\mathbb{E}[\boldsymbol{x} \mid \boldsymbol{z}] &= \mathbb{E}[\boldsymbol{\mu} + \boldsymbol{L}\boldsymbol{z} + \epsilon \mid \boldsymbol{z}] \\
&= \mathbb{E}[\boldsymbol{\mu} \mid \boldsymbol{z}] + \mathbb{E}[\boldsymbol{L}\boldsymbol{z} \mid \boldsymbol{z}] + \mathbb{E}[\boldsymbol{\epsilon} \mid \boldsymbol{z}] \\
&= \mathbb{E}[\boldsymbol{\mu}] + \boldsymbol{L}\mathbb{E}[\boldsymbol{z} \mid \boldsymbol{z}] + \mathbb{E}[\boldsymbol{\epsilon}] \\
&= \boldsymbol{\mu} + \boldsymbol{L}\boldsymbol{z} + \boldsymbol{0} = \boldsymbol{\mu} + \boldsymbol{L}\boldsymbol{z} \qquad (A.6)
\end{aligned}$$

so that:

$$\begin{aligned} Cov(\boldsymbol{x}) &= \mathbb{E}[(\boldsymbol{x} - \boldsymbol{L}\boldsymbol{z} - \boldsymbol{\mu})(\boldsymbol{x} - \boldsymbol{L}\boldsymbol{z}\boldsymbol{\mu})^T] \\ &= \mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon} \mid \boldsymbol{z}] \\ &= \boldsymbol{\Psi} \end{aligned} \tag{A.7}$$

□

Proof of $A = \boldsymbol{L}^T(\boldsymbol{L}\boldsymbol{L}^T + \boldsymbol{\Psi})^{-1} = (I + \boldsymbol{L}^T\boldsymbol{\Psi}^{-1}\boldsymbol{L})^{-1}\boldsymbol{L}^T\boldsymbol{\Psi}^{-1}$ (equation 3.10)

$$\begin{aligned} A &= \boldsymbol{L}^T(\boldsymbol{L}\boldsymbol{L}^T + \boldsymbol{\Psi})^{-1} \\ &= \boldsymbol{L}^T[\boldsymbol{\Psi}^{-1} - \boldsymbol{\Psi}^{-1}\boldsymbol{L}(\boldsymbol{I} + \boldsymbol{L}^T\boldsymbol{\Psi}^{-1}\boldsymbol{L})^{-1}\boldsymbol{L}^T\boldsymbol{\Psi}^{-1}] \\ &= \boldsymbol{L}^T\boldsymbol{\Psi}^{-1} - \boldsymbol{L}^T\boldsymbol{\Psi}^{-1}\boldsymbol{L}(\boldsymbol{I} + \boldsymbol{L}^T\boldsymbol{\Psi}^{-1}\boldsymbol{L})^{-1}\boldsymbol{L}^T\boldsymbol{\Psi}^{-1} \\ &= [\boldsymbol{I} - \boldsymbol{L}^T\boldsymbol{\Psi}^{-1}\boldsymbol{L}(\boldsymbol{I} + \boldsymbol{L}^T\boldsymbol{\Psi}^{-1}\boldsymbol{L})^{-1}]\boldsymbol{L}^T\boldsymbol{\Psi}^{-1} \\ &= [\boldsymbol{I} + (\boldsymbol{I} + \boldsymbol{L}^T\boldsymbol{\Psi}^{-1}\boldsymbol{L})^{-1} - (\boldsymbol{I} + \boldsymbol{L}^T\boldsymbol{\Psi}^{-1}\boldsymbol{L})(\boldsymbol{I} + \boldsymbol{L}^T\boldsymbol{\Psi}^{-1}\boldsymbol{L})^{-1}]\boldsymbol{L}^T\boldsymbol{\Psi}^{-1} \\ &= [\boldsymbol{I} + (\boldsymbol{I} + \boldsymbol{L}^T\boldsymbol{\Psi}^{-1}\boldsymbol{L})^{-1} - \boldsymbol{I}]\boldsymbol{L}^T\boldsymbol{\Psi}^{-1} \\ &= (\boldsymbol{I} + \boldsymbol{L}^T\boldsymbol{\Psi}^{-1}\boldsymbol{L})^{-1}\boldsymbol{L}^T\boldsymbol{\Psi}^{-1} \end{aligned}$$

where we have used the Binomial Inverse Matrix Theorem [Strang, 2003]

$$(\boldsymbol{A} + \boldsymbol{U}\boldsymbol{B}\boldsymbol{V})^{-1} = \boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}\boldsymbol{U}\boldsymbol{B}(\boldsymbol{B} + \boldsymbol{B}\boldsymbol{V}\boldsymbol{A}^{-1}\boldsymbol{U}\boldsymbol{B})^{-1}\boldsymbol{B}\boldsymbol{V}\boldsymbol{A}^{-1} \tag{A.8}$$

□

Proof of $p(\boldsymbol{x}) \sim N(\boldsymbol{\mu}, \boldsymbol{L}\boldsymbol{L}^T + \boldsymbol{\Psi})$ (equation 3.11)

$$\begin{aligned} \mathbb{E}[\boldsymbol{x}] &= \mathbb{E}[\boldsymbol{\mu} + \boldsymbol{L}\boldsymbol{z} + \boldsymbol{\epsilon}] \\ &= \boldsymbol{\mu} + \boldsymbol{L}\mathbb{E}[\boldsymbol{z}] + \mathbb{E}[\boldsymbol{\epsilon}] \\ &= \boldsymbol{\mu} \end{aligned}$$

$$\begin{aligned} Cov(\boldsymbol{x}) &= \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^T] = \mathbb{E}[(\boldsymbol{\mu} + \boldsymbol{L}\boldsymbol{z} + \boldsymbol{\epsilon})(\boldsymbol{\mu} + \boldsymbol{L}\boldsymbol{z} + \boldsymbol{\epsilon})^T] = \mathbb{E}[\boldsymbol{\mu}\boldsymbol{\mu}^T] + \boldsymbol{\mu}\mathbb{E}[\boldsymbol{z}^{\boldsymbol{T}}]\boldsymbol{L}^T + \boldsymbol{\mu}\mathbb{E}[\boldsymbol{\epsilon}^T] \\ &+ \boldsymbol{L}\mathbb{E}[\boldsymbol{z}]\boldsymbol{\mu}^T + \boldsymbol{L}\mathbb{E}[\boldsymbol{z}\boldsymbol{z}^T]\boldsymbol{L}^T + \boldsymbol{L}\mathbb{E}[\boldsymbol{z}]\boldsymbol{\epsilon}^{\boldsymbol{T}} + \mathbb{E}[\boldsymbol{\epsilon}]\boldsymbol{\mu}^{\boldsymbol{T}} + \mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{z}^T]\boldsymbol{L}^T + \mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T] \\ &= \boldsymbol{L}\mathbb{E}[\boldsymbol{z}\boldsymbol{z}^T]\boldsymbol{L} + \mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T] \\ &= \boldsymbol{L}\boldsymbol{L}^T + \boldsymbol{\Psi} \end{aligned}$$

□

Proof of $p(\boldsymbol{z} \mid \boldsymbol{x}) \sim N(A(\boldsymbol{x} - \boldsymbol{\mu}), (I + \boldsymbol{L}^T\boldsymbol{\Psi}^{-1}\boldsymbol{L})^{-1})$ (equation 3.9).

$$\begin{aligned} p(\boldsymbol{z} \mid \boldsymbol{x}) &= \frac{p(\boldsymbol{x},\boldsymbol{z})}{p(\boldsymbol{x})} \\ &= \frac{(2\pi)^{-(d+q)/2} \mid \boldsymbol{\Lambda} \mid^{-1/2} exp(-1/2\boldsymbol{y}^T\boldsymbol{\Lambda}^{-1}\boldsymbol{y})}{(2\pi)^{-(d+q)/2} \mid \ \boldsymbol{C} \mid^{-1/2} exp(-1/2\boldsymbol{y}^T\boldsymbol{C}^{-1}\boldsymbol{y})} \\ &\propto exp(-\frac{1}{2}(\boldsymbol{y}^T\boldsymbol{\Lambda}^{-1}\boldsymbol{y} - \boldsymbol{x}^T\boldsymbol{C}^{-1}\boldsymbol{x})) \end{aligned} \tag{A.9}$$

being $y = \begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{z} \end{bmatrix}$, $\boldsymbol{C} = Cov(\boldsymbol{x}) = \boldsymbol{L}\boldsymbol{L}^T + \boldsymbol{\Psi}$ and $\boldsymbol{\Lambda} = Cov(\boldsymbol{x},\boldsymbol{z})$ derived as

$$\begin{aligned} Cov(\boldsymbol{x},\boldsymbol{z}) &= Cov(\boldsymbol{y}) = \mathbb{E}[\begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{z} \end{bmatrix} [\boldsymbol{x}^T\boldsymbol{z}^T]] \\ &= \mathbb{E}\begin{bmatrix} \boldsymbol{x}\boldsymbol{x}^T & \boldsymbol{x}\boldsymbol{z}^T \\ \boldsymbol{z}\boldsymbol{x}^T & \boldsymbol{z}\boldsymbol{z}^T \end{bmatrix} \\ &= \begin{bmatrix} \boldsymbol{L}\boldsymbol{L}^T + \boldsymbol{\Psi} & \boldsymbol{L} \\ \boldsymbol{L}^T & \boldsymbol{I} \end{bmatrix} = \Lambda \end{aligned} \tag{A.10}$$

given that

$$\begin{aligned} \boldsymbol{\Lambda}^{-1} &= \begin{bmatrix} \boldsymbol{\Lambda}^{-1,11} & \boldsymbol{\Lambda}^{-1,12} \\ \boldsymbol{\Lambda}^{-1,21} & \boldsymbol{\Lambda}^{-1,22} \end{bmatrix} \\ &= \begin{bmatrix} (\boldsymbol{\Lambda}_{11} - \boldsymbol{\Lambda}_{12}\boldsymbol{\Lambda}_{22}^{-1}\boldsymbol{\Lambda}_{21})^{-1} & \boldsymbol{\Lambda}_{11}^{-1}\boldsymbol{\Lambda}_{12}(\boldsymbol{\Lambda}_{21}\boldsymbol{\Lambda}_{11}^{-1}\boldsymbol{\Lambda}_{12} - \boldsymbol{\Lambda}_{22})^{-1} \\ (\boldsymbol{\Lambda}_{21}\boldsymbol{\Lambda}_{11}^{-1}\boldsymbol{\Lambda}_{12} - \boldsymbol{\Lambda}_{22})^{-1}\boldsymbol{\Lambda}_{21}\Lambda_{11}^{-1} & (\boldsymbol{\Lambda}_{22} - \boldsymbol{\Lambda}_{21}\boldsymbol{\Lambda}_{11}^{-1}\boldsymbol{\Lambda}_{12})^{-1} \end{bmatrix} \end{aligned}$$

Consider now the term inside the exponent in equation A.9

$$\begin{aligned} \boldsymbol{y}^T\boldsymbol{\Lambda}^{-1}\boldsymbol{y} - \boldsymbol{x}^T\boldsymbol{C}^{-1}\boldsymbol{x} &= \begin{bmatrix} \boldsymbol{x}^T & \boldsymbol{z}^T \end{bmatrix}\Lambda^{-1}\begin{bmatrix} \boldsymbol{x}^T \\ \boldsymbol{z} \end{bmatrix} - \boldsymbol{x}^T\boldsymbol{C}^{-1}\boldsymbol{x} \\ &= \boldsymbol{x}^T\boldsymbol{\Lambda}^{-1,11}\boldsymbol{x} + \boldsymbol{x}^T\boldsymbol{\Lambda}^{-1,12} + \boldsymbol{z}^T\boldsymbol{\Lambda}^{-1,21}\boldsymbol{x} + \boldsymbol{z}^T\boldsymbol{\Lambda}^{-1,22}\boldsymbol{z} - \boldsymbol{x}^T\boldsymbol{C}^{-1}\boldsymbol{x} \\ &= \boldsymbol{x}^T(\boldsymbol{\Lambda}^{-1,11} - \boldsymbol{C}^{-1})\boldsymbol{x} + 2\boldsymbol{x}\boldsymbol{\Lambda}^{-1,12}z + \boldsymbol{z}^T\boldsymbol{\Lambda}^{-1,22}\boldsymbol{z} \end{aligned} \tag{A.11}$$

Analysing the term

$$\boldsymbol{\Lambda}^{-1,11} - \boldsymbol{C}^{-1} = (\boldsymbol{\Lambda}_{11} - \boldsymbol{\Lambda}_{12}\boldsymbol{\Lambda}_{22}^{-1}\boldsymbol{\Lambda}_{21})^{-1} \tag{A.12}$$

$$\begin{aligned} &= \boldsymbol{\Lambda}_{11}^{-1} + \boldsymbol{\Lambda}_{11}^{-1}\boldsymbol{\Lambda}_{12}(\boldsymbol{\Lambda}_{22} - \boldsymbol{\Lambda}_{21}\boldsymbol{\Lambda}_{11}^{-1}\boldsymbol{\Lambda}_{12})\boldsymbol{\Lambda}_{21}\boldsymbol{\Lambda}_{11}^{-1} - \boldsymbol{\Lambda}_{11}^{-1} \\ &= \boldsymbol{\Lambda}_{11}^{-1}\boldsymbol{\Lambda}_{12}(\boldsymbol{\Lambda}_{22} - \boldsymbol{\Lambda}_{21}\boldsymbol{\Lambda}_{11}^{-1}\boldsymbol{\Lambda}_{12})\boldsymbol{\Lambda}_{21}\boldsymbol{\Lambda}_{11}^{-1} \\ &= \boldsymbol{\beta}\boldsymbol{\Lambda}^{-1,22}\boldsymbol{\beta} \end{aligned} \tag{A.13}$$

being $\boldsymbol{\beta}$

$$\boldsymbol{\beta} = \boldsymbol{\Lambda}_{21}\boldsymbol{\Lambda}_{11}^{-1} = \boldsymbol{L}^T(\boldsymbol{L}\boldsymbol{L}^T + \boldsymbol{\Psi})^{-1} \tag{A.14}$$

Substituting equation A.13 in A.11 then

$$\begin{aligned}
\boldsymbol{y}^T\boldsymbol{\Lambda}^{-1}\boldsymbol{y} - \boldsymbol{x}^T\boldsymbol{C}^{-1}\boldsymbol{x} &= \boldsymbol{x}^T\boldsymbol{\beta}^T\boldsymbol{\Lambda}^{-1,22}\boldsymbol{\beta}\boldsymbol{x} + 2\boldsymbol{x}^T\boldsymbol{\Lambda}^{\boldsymbol{-1,12}}\boldsymbol{z} + \boldsymbol{z}^T\boldsymbol{\Lambda}^{\boldsymbol{-1,22}}\boldsymbol{z} \\
&= (\boldsymbol{z} - \boldsymbol{\beta}^T\boldsymbol{x})^T\boldsymbol{\Lambda}^{\boldsymbol{-1,22}}(\boldsymbol{z} - \boldsymbol{\beta}^T\boldsymbol{x}) + 2\boldsymbol{x}^T\boldsymbol{\beta}^T\boldsymbol{\Lambda}^{\boldsymbol{-1,22}}\boldsymbol{z} + 2\boldsymbol{x}^T\boldsymbol{\Lambda}^{\boldsymbol{-1,12}}\boldsymbol{z} \\
&= (\boldsymbol{z} - \boldsymbol{\beta}\boldsymbol{x})^T\boldsymbol{\Lambda}^{\boldsymbol{-1,22}}(\boldsymbol{z} - \boldsymbol{\beta}\boldsymbol{x}) + 2\boldsymbol{x}^T(\boldsymbol{\beta}^T\boldsymbol{\Lambda}^{\boldsymbol{-1,22}} + \boldsymbol{\Lambda}^{\boldsymbol{-1,12}})\boldsymbol{z} \qquad \text{(A.15)}
\end{aligned}$$

Noting that

$$\begin{aligned}
\boldsymbol{\beta}^T\boldsymbol{\Lambda}^{\boldsymbol{-1,22}} + \boldsymbol{\Lambda}^{\boldsymbol{-1,12}} &= \boldsymbol{\Lambda}_{11}^{-1}\boldsymbol{\Lambda}_{12}(\boldsymbol{\Lambda}_{22} - \boldsymbol{\Lambda}_{21}\boldsymbol{\Lambda}_{11}^{-1}\boldsymbol{\Lambda}_{12})^{-1} \\
&+ \boldsymbol{\Lambda}_{11}^{-1}\boldsymbol{\Lambda}_{12}(\boldsymbol{\Lambda}_{21} - \boldsymbol{\Lambda}_{11}^{-1}\boldsymbol{\Lambda}_{12} - \boldsymbol{\Lambda}_{22})^{-1} \\
&= 0 \qquad \text{(A.16)}
\end{aligned}$$

Hence

$$\boldsymbol{y}^T\boldsymbol{\Lambda}^{-1}\boldsymbol{y} - \boldsymbol{x}^T\boldsymbol{C}^{-1}\boldsymbol{x} = (\boldsymbol{z} - \boldsymbol{\beta}\boldsymbol{x})^T\boldsymbol{\Lambda}^{\boldsymbol{-1,22}}(\boldsymbol{z} - \boldsymbol{\beta}\boldsymbol{x}) \tag{A.17}$$

then

$$p(\boldsymbol{z} \mid \boldsymbol{x}) \propto exp(-\frac{1}{2}(\boldsymbol{z} - \boldsymbol{\beta}\boldsymbol{x})^T\boldsymbol{\Lambda}^{\boldsymbol{-1,22}}(\boldsymbol{z} - \boldsymbol{\beta}\boldsymbol{x})) \tag{A.18}$$

and finally

$$\mathbb{E}[\boldsymbol{z} \mid \boldsymbol{x}] = \boldsymbol{\beta}\boldsymbol{x} = (\boldsymbol{I} + \boldsymbol{L}^T\boldsymbol{\Psi}^{-1}\boldsymbol{L})^{-1}\boldsymbol{L}^T\boldsymbol{\Psi}^{-1}\boldsymbol{x} = \boldsymbol{A}\boldsymbol{x} \tag{A.19}$$

(note that $\boldsymbol{\mu}$ has been considered 0 without loss of generality).
Regarding the covariance term

$$\begin{aligned}
Cov(\boldsymbol{z} \mid \boldsymbol{x}) &= \boldsymbol{(\Lambda^{-1,22})^{-1}} \\
&= (\boldsymbol{\Lambda}_{22} - \boldsymbol{\Lambda}_{21}\boldsymbol{\Lambda}_{11}^{-1}\boldsymbol{\Lambda}_{12})^{-1} \\
&= \boldsymbol{\Lambda}_{22}^{-1} - \boldsymbol{\Lambda}_{22}^{-1}\boldsymbol{\Lambda}_{21}(-\boldsymbol{\Lambda}_{11} + \boldsymbol{\Lambda}_{12}\boldsymbol{\Lambda}_{22}^{-1}\boldsymbol{\Lambda}_{21})^{-1}\boldsymbol{\Lambda}_{12}\boldsymbol{\Lambda}_{22}^{-1} \\
&= \boldsymbol{I} - \boldsymbol{L}^T(-\boldsymbol{\Psi} - \boldsymbol{L}\boldsymbol{L}^T + \boldsymbol{L}\boldsymbol{L}^T)^{-1}\boldsymbol{L} \\
&= \boldsymbol{I} + \boldsymbol{L}^T\boldsymbol{\Psi}^{-1}\boldsymbol{L} \qquad \text{(A.20)}
\end{aligned}$$

and therefore

$$Cov(\boldsymbol{z} \mid \boldsymbol{x}) = (\boldsymbol{I} + \boldsymbol{L}^T\boldsymbol{\Psi}^{-1}\boldsymbol{L})^{-1} \tag{A.21}$$

□

Derivation of complete-data log-likelihood form, equation 3.14

$$\begin{aligned}
\mathcal{L} &= \sum_{i=1}^{N} log \frac{1}{(2\pi)^{d/2} \mid \boldsymbol{\Psi} \mid^{1/2}} exp\{-\frac{1}{2}(\boldsymbol{x}_i - \boldsymbol{L}\boldsymbol{z}_i)^T \boldsymbol{\Psi}^{-1}(\boldsymbol{x}_i - \boldsymbol{L}\boldsymbol{z}_i)\} \\
&= -\frac{Nd}{2} log(2\pi) - \frac{N}{2} log \mid \boldsymbol{\Psi} \mid - \frac{1}{2}\sum_{i}^{n}(\boldsymbol{x}_i^T \boldsymbol{\Psi}^{-1}\boldsymbol{x}_i - 2\boldsymbol{x}_i^T \boldsymbol{\Psi}^{-1}\boldsymbol{L}\boldsymbol{z}_i + \boldsymbol{z}_i \boldsymbol{L}^T \boldsymbol{\Psi}^{-1}\boldsymbol{L}\boldsymbol{z}_i) \quad \text{(A.22)} \\
&= -\frac{Nd}{2} log(2\pi) - \frac{N}{2} log \mid \boldsymbol{\Psi} \mid - \frac{1}{2}\sum_{i}^{n}(\boldsymbol{x}_i^T \boldsymbol{\Psi}^{-1}\boldsymbol{x}_i - 2\boldsymbol{x}_i^T \boldsymbol{\Psi}^{-1}\boldsymbol{L}\boldsymbol{z}_i + tr[\boldsymbol{L}^T \boldsymbol{\Psi}^{-1}\boldsymbol{L}\boldsymbol{z}_i\boldsymbol{z}_i])
\end{aligned}$$

where the equality $\boldsymbol{z}^T\boldsymbol{L}\boldsymbol{z} = tr[\boldsymbol{L}\boldsymbol{z}\boldsymbol{z}^T]$ has been used in the last step.

□

Derivation of M-step equation 3.35

$$\begin{aligned}
\frac{\partial \mathbb{E}_{p(\boldsymbol{z}|\boldsymbol{x})}[\mathcal{L}]}{\partial \boldsymbol{L}} &= \frac{\partial \left[C - \frac{N}{2} ln \mid \boldsymbol{\Psi} \mid - \frac{1}{2}\sum_{i=1}^{N}\{\boldsymbol{x}_i^T \boldsymbol{\Psi}^{-1}\boldsymbol{x}_i - 2\boldsymbol{x}_i^T \boldsymbol{\Psi}^{-1}\boldsymbol{L}\mathbb{E}[\boldsymbol{z}_i \mid \boldsymbol{x}_i] + tr[\boldsymbol{L}^T \boldsymbol{\Psi}^{-1}\boldsymbol{L}\mathbb{E}[\boldsymbol{z}_i\boldsymbol{z}_i^T \mid \boldsymbol{x}_i]]\}\right]}{\partial \boldsymbol{L}} \\
&= -\frac{1}{2}\sum_{i=1}^{N}\{-2\boldsymbol{\Psi}^{-1}\boldsymbol{x}_i\mathbb{E}[\boldsymbol{z}_i \mid \boldsymbol{x}_i] + 2\boldsymbol{\Psi}^{-1}\boldsymbol{L}\mathbb{E}[\boldsymbol{z}_i\boldsymbol{z}_i^T \mid \boldsymbol{x}_i]]\} \quad \text{(A.23)}
\end{aligned}$$

where relations $\frac{\partial A^T XB}{\partial \boldsymbol{X}} = A^TB$ and $\frac{\partial tr[X^TAXB]}{\partial \boldsymbol{X}} = AXB + A^TXB^T$ have been used.

Hence, setting equation A.23 to zero

$$\frac{\partial \mathbb{E}_{p(\boldsymbol{z}|\boldsymbol{x})}[\mathcal{L}]}{\partial \boldsymbol{L}} = 0 \Rightarrow \boldsymbol{L}^* = \left(\sum_{i}^{N}\boldsymbol{x}_i\mathbb{E}[\boldsymbol{z}_i \mid \boldsymbol{x}_i]^T\right)\left(\sum_{i}^{N}\mathbb{E}[\boldsymbol{z}_i\boldsymbol{z}_i^T \mid \boldsymbol{x}_i]\right)^{-1} \quad \text{(A.24)}$$

□

Derivation of M-step equation 3.36

$$\begin{aligned}
\frac{\partial \mathbb{E}_{p(\boldsymbol{z}|\boldsymbol{x})}[\mathcal{L}]}{\partial \boldsymbol{\Psi}} &= \frac{\partial \left[C - \frac{N}{2} ln \mid \boldsymbol{\Psi} \mid - \frac{1}{2}\sum_{i=1}^{N}\{\boldsymbol{x}_i^T \boldsymbol{\Psi}^{-1}\boldsymbol{x}_i - 2\boldsymbol{x}_i^T\boldsymbol{\Psi}^{-1}\boldsymbol{L}\mathbb{E}[\boldsymbol{z}_i \mid \boldsymbol{x}_i] + tr[\boldsymbol{L}^T\boldsymbol{\Psi}^{-1}\boldsymbol{L}\mathbb{E}[\boldsymbol{z}_i\boldsymbol{z}_i^T \mid \boldsymbol{x}_i]]\}\right]}{\partial \boldsymbol{\Psi}} \\
&= \frac{N}{2}\boldsymbol{\Psi} - \frac{1}{2}\sum_{i=1}^{N}\{\boldsymbol{x}_i\boldsymbol{x}_i^T - 2\boldsymbol{x}_i\mathbb{E}[\boldsymbol{z}_i \mid \boldsymbol{x}_i]\boldsymbol{L}^T + \boldsymbol{L}\mathbb{E}[\boldsymbol{z}_i\boldsymbol{z}_i^T \mid \boldsymbol{x}_i]\boldsymbol{L}^T\} \\
&= \frac{N}{2}\boldsymbol{\Psi} - \frac{1}{2}\sum_{i=1}^{N}\boldsymbol{x}_i\boldsymbol{x}_i^T + \left(\sum_{i=1}^{N}\boldsymbol{x}_i\mathbb{E}[\boldsymbol{z}_i \mid \boldsymbol{x}_i]^T\right)\boldsymbol{L}^T - \frac{1}{2}\boldsymbol{L}\left(\sum_{i=1}^{N}\mathbb{E}[\boldsymbol{z}_i\boldsymbol{z}_i^T \mid \boldsymbol{x}_i]\right)\boldsymbol{L}^T \qquad \text{(A.25)}
\end{aligned}$$

where relations $\frac{\partial A^T X B}{\partial \boldsymbol{X}} = A^T B$ and $\frac{\partial log|X|}{\partial \boldsymbol{X}} = \left(X^{-1}\right)^T$ have been used.

Hence, setting equation A.25 to zero

$$\boldsymbol{\Psi} = \frac{1}{N}\left[\sum_{i=1}^{N}\boldsymbol{x}_i\boldsymbol{x}_i^T - 2\left(\sum_{i=1}^{N}\boldsymbol{x}_i\mathbb{E}[\boldsymbol{z}_i \mid \boldsymbol{x}_i]^T\right) + \boldsymbol{L}\left(\sum_{i=1}^{N}\mathbb{E}[\boldsymbol{z}_i\boldsymbol{z}_i^T \mid \boldsymbol{x}_i]\right)\boldsymbol{L}^T\right] \qquad \text{(A.26)}$$

and replacing $\boldsymbol{L}$ by its update equation given in A.24, we obtain

$$\boldsymbol{\Psi} = \frac{1}{N} diag\left[\sum_{i=1}^{N}\boldsymbol{x}_i\boldsymbol{x}_i^T - \left(\sum_{i=1}^{N}\boldsymbol{x}_i\mathbb{E}[\boldsymbol{z}_i \mid \boldsymbol{x}_i]^T\right)\boldsymbol{L}\right] \qquad \text{(A.27)}$$

□

Appendix B

Linear Scoring

The Taylor series of a real or complex function $f(x)$ that is infinitely differentiable in a neighborhood of a real or complex number a is defined as

$$\sum_{n=0}^{\infty} \frac{f^{(n)}(a)}{n!}(x-a)^n \tag{B.1}$$

where $f^{(n)}(a)$ denotes the nth derivative of $f(x)$ evaluated at the point a; and n defines the order of the Taylor series, or in other words, the sum terms used.

Let $\boldsymbol{O} = \boldsymbol{o}_1, ..., \boldsymbol{o}_t$ be a set of test observations and λ_s a GMM model for a given speaker s, with mean supervector $\boldsymbol{\mu}_s$; by Linear Scoring the likelihood $P(\boldsymbol{O} \mid \lambda_s)$ is approximated via a 1st Taylor series evaluated at the UBM model mean supervector point, $\boldsymbol{\mu}$, as

$$\begin{aligned} P(\boldsymbol{O} \mid \lambda_s) & \sim \frac{f^0(\lambda_{ubm})}{0!}(\boldsymbol{\mu}_s - \boldsymbol{\mu})^0 + \frac{f^1(\lambda_{ubm})}{1!}(\boldsymbol{\mu}_s - \boldsymbol{\mu})^1 \\ & = \frac{f^0(\lambda_{ubm})}{\cancelto{1}{0!}}\cancelto{1}{(\boldsymbol{\mu}_s - \boldsymbol{\mu})^0} + \frac{f^1(\lambda_{ubm})}{1!}(\boldsymbol{\mu} - \boldsymbol{\mu})^1 \\ & = f^0(\lambda_{ubm}) + f^1(\lambda_{ubm})(\boldsymbol{\mu}_s - \boldsymbol{\mu}) \\ & = P(\boldsymbol{O} \mid \lambda_{ubm}) + \bigtriangledown P(\boldsymbol{O} \mid \lambda_s)[\boldsymbol{\mu}](\boldsymbol{\mu}_s - \boldsymbol{\mu}) \end{aligned} \tag{B.2}$$

where the second term, the gradient of the likelihood versus the target model, λ_s, evaluated at the UBM mean supervector, $\boldsymbol{\mu}_s$, can be developed as

$$\begin{aligned}
\triangledown\, log\,(P(\boldsymbol{O} \mid \lambda_s))[\boldsymbol{\mu}] &= \triangledown \sum_{t=1}^{T} log(P(\boldsymbol{o}_t \mid \lambda_s))[\boldsymbol{\mu}] \\
&= \triangledown \sum_{t=1}^{T} log\left(\sum_{k=1}^{K} w_k p_k(\boldsymbol{o}_t)\right) \\
&= \sum_{t=1}^{T} \frac{1}{\sum_{k=1}^{K} w_k p_k(\boldsymbol{o}_t)} \triangledown \sum_{k=1}^{K} w_k p_k(\boldsymbol{o}_t) \\
&= \sum_{t=1}^{T} \frac{1}{\sum_{k=1}^{K} w_k p_k(\boldsymbol{o}_t)} \sum_{k=1}^{K} w_k \triangledown p_k(\boldsymbol{o}_t) \\
&= \sum_{t=1}^{T} \frac{1}{\sum_{k=1}^{K} w_k p_k(\boldsymbol{o}_t)} \sum_{k=1}^{K} w_k p_k(\boldsymbol{o}_t)(\boldsymbol{o}_t - \boldsymbol{\mu}_k)\boldsymbol{\Sigma}_k^{-1} \qquad \text{(B.3)}
\end{aligned}$$

taken into account that the Gaussian occupation probability is defined as

$$P_{kt} = \frac{w_k p_k(\boldsymbol{o}_t)}{\sum_{k=1}^{K} w_k p_k(\boldsymbol{o}_t)} \qquad \text{(B.4)}$$

it can be readily seen that B.3 reduces to the first order normalized statistics defined as

$$1st_{norm} \longrightarrow \boldsymbol{f}_k = \sum_{t} \boldsymbol{\Sigma}_k^{-1} P_{kt}(\boldsymbol{o}_t - \boldsymbol{\mu}_k) \qquad \text{(B.5)}$$

Under this analysis, classical scoring defined as the log-likelihood ratio can be computed as

$$\begin{aligned}
score_{\boldsymbol{O},\lambda_s} &= log(P(\boldsymbol{O} \mid \lambda_s)) - log(P(\boldsymbol{O} \mid \lambda_{ubm})) \\
&= P(\boldsymbol{O} \mid \lambda_{ubm}) + \triangledown P(\boldsymbol{O} \mid \lambda_s)[\boldsymbol{\mu}](\boldsymbol{\mu}_s - \boldsymbol{\mu}) - log(P(\boldsymbol{O} \mid \lambda_{ubm})) \\
&= \triangledown P(\boldsymbol{O} \mid \lambda_s)[\boldsymbol{\mu}](\boldsymbol{\mu}_s - \boldsymbol{\mu}) \\
&= \boldsymbol{f}(\boldsymbol{\mu}_s - \boldsymbol{\mu})
\end{aligned}$$

(B.6)

Appendix C

Extended Results

Extended results for both speaker and language recognition systems presented in Chapter 6 are included in this Appendix. Particularly, complete data-table results for different configurations of the JFA SV systems used besides a by-language decomposition of SLR results are presented.

((**a**)) $\boldsymbol{U}: PCA$

$\#Eigs$	Raw	Tnorm	Znorm	Ztnorm
0	13.09/0.064	12.86/0.056	12.65/0.056	12.17/0.053
50	8.93/0.044	8.30/0.041	8.29/0.043	7.59/0.041
100	8.85/0.045	8.26/0.043	8.45/0.044	7.87/0.042
150	9.01/0.045	8.38/0.043	8.46/0.044	7.91/0.042
200	8.89/0.044	8.34/0.042	8.48/0.044	7.95/0.042
250	9.01/0.045	8.46/0.043	8.62/0.044	7.99/0.043
300	8.97/0.044	8.35/0.042	8.52/0.044	8.04/0.043

((**b**)) $\boldsymbol{U}: PCA + 1EMiteration$

$\#Eigs$	Raw	Tnorm	Znorm	Ztnorm
0	13.09/0.064	12.86/0.056	12.65/0.056	12.17/0.053
50	8.89/0.044	8.14/0.040	8.23/0.043	7.40/0.040
100	8.51/0.044	7.91/0.042	8.21/0.043	7.56/0.041
150	8.62/0.044	8.07/0.042	8.20/0.043	7.67/0.041
200	8.66/0.044	8.00/0.042	8.22/0.043	7.56/0.042
250	8.62/0.044	8.11/0.042	8.30/0.043	7.59/0.042
300	8.52/0.044	7.99/0.042	8.26/0.043	7.53/0.042

((**c**)) $\boldsymbol{U}: PCA + 5EMiteration$

$\#Eigs$	Raw	Tnorm	Znorm	Ztnorm
0	13.09/0.064	12.86/0.056	12.65/0.056	12.17/0.053
50	9.09/0.044	8.19/0.041	8.30/0.044	7.44/0.041
100	8.51/0.045	8.10/0.042	8.11/0.044	7.56/0.041
150	8.62/0.044	8.11/0.042	8.11/0.043	7.59/0.041
200	8.66/0.044	8.06/0.042	8.16/0.043	7.51/0.041
250	8.43/0.044	8.07/0.042	8.24/0.043	7.48/0.042
300	8.34/0.043	7.99/0.041	8.25/0.043	7.48/0.042

((**d**)) $\boldsymbol{U}: PCA + 10EMiteration$

$\#Eigs$	Raw	Tnorm	Znorm	Ztnorm
0	13.09/0.064	12.86/0.056	12.65/0.056	12.17/0.053
50	9.13/0.044	8.22/0.041	8.22/0.044	7.44/0.041
100	8.46/0.045	8.05/0.042	8.08/0.044	7.63/0.042
150	8.70/0.044	8.11/0.042	8.12/0.043	7.52/0.041
200	8.60/0.044	8.10/0.042	8.16/0.043	7.53/0.042
250	8.36/0.044	8.11/0.042	8.26/0.043	7.55/0.042
300	8.30/0.043	7.99/0.041	8.20/0.042	7.49/0.042

Table C.1: *System: MAP-SVC. Results on SRE'08 female tel-tel condition by using different ML iterations on training the session variability subspace U as well as different number of eigenchannels.*

((a)) $\boldsymbol{V}: PCA$

$\#Eigs$	Raw	Tnorm	Znorm	Ztnorm
0	8.93/0.044	8.30/0.041	8.29/0.043	7.59/0.041
50	14.27/0.069	12.15/0.059	12.35/0.060	9.77/0.046
100	11.84/0.060	10.35/0.050	10.37/0.050	8.86/0.041
150	11.05/0.056	9.88/0.046	9.81/0.047	8.50/0.040
200	10.82/0.054	9.52/0.045	9.36/0.047	8.22/0.040
250	10.66/0.053	9.29/0.045	9.25/0.046	7.94/0.040
300	10.41/0.052	9.24/0.044	9.21/0.045	8.07/0.039

((b)) $\boldsymbol{V}: PCA + 1EMiteration$

$\#Eigs$	Raw	Tnorm	Znorm	Ztnorm
0	8.93/0.044	8.30/0.041	8.29/0.043	7.59/0.041
50	15.10/0.070	13.05/0.061	12.34/0.060	10.19/0.047
100	12.35/0.060	10.96/0.052	10.35/0.049	9.01/0.042
150	11.40/0.056	10.17/0.049	9.80/0.047	8.62/0.041
200	11.09/0.054	9.95/0.047	9.64/0.046	8.51/0.041
250	10.91/0.053	9.83/0.046	9.55/0.046	8.54/0.040
300	10.86/0.053	9.72/0.046	9.46/0.046	8.46/0.040

((c)) $\boldsymbol{V}: PCA + 5EMiteration$

$\#Eigs$	Raw	Tnorm	Znorm	Ztnorm
0	8.93/0.044	8.30/0.041	8.29/0.043	7.59/0.041
50	15.28/0.070	13.18/0.062	12.39/0.060	10.27/0.047
100	12.45/0.060	11.14/0.051	10.31/0.050	9.09/0.042
150	11.42/0.056	10.11/0.049	9.72/0.048	8.81/0.041
200	10.97/0.054	9.95/0.047	9.64/0.046	8.62/0.040
250	10.82/0.054	9.83/0.047	9.48/0.045	8.54/0.041
300	10.78/0.052	9.72/0.045	9.40/0.045	8.40/0.041

((d)) $\boldsymbol{V}: PCA + 10EMiteration$

$\#Eigs$	Raw	Tnorm	Znorm	Ztnorm
0	8.93/0.044	8.30/0.041	8.29/0.043	7.59/0.041
50	15.18/0.070	13.37/0.062	12.51/0.060	10.26/0.048
100	12.35/0.060	11.01/0.052	10.46/0.050	9.01/0.042
150	11.33/0.056	10.20/0.049	9.73/0.048	8.70/0.041
200	10.89/0.055	9.98/0.047	9.64/0.046	8.54/0.040
250	10.78/0.054	9.84/0.046	9.52/0.046	8.46/0.040
300	10.70/0.052	9.68/0.045	9.38/0.045	8.42/0.040

Table C.2: *System: EV-SVC. Results on SRE'08 female tel-tel condition by using different ML iterations on training the speaker variability subspace V as well as different number of eigenvoices. U is fixed over all the experiments and was trained via PCA keeping 50 eigenchannels.*

((a)) $\boldsymbol{V} : PCA$

$\#Eigs$	Raw	Tnorm	Znorm	Ztnorm
0	8.93/0.044	8.30/0.041	8.29/0.043	7.59/0.041
50	9.40/0.045	8.43/0.041	8.37/0.041	7.40/0.040
100	9.06/0.044	8.32/0.040	8.22/0.041	7.42/0.039
150	9.04/0.044	8.26/0.040	8.18/0.041	7.40/0.039
200	9.01/0.044	8.19/0.040	8.22/0.041	7.32/0.039
250	9.05/0.045	8.22/0.040	8.30/0.041	7.29/0.039
300	9.03/0.044	8.22/0.040	8.18/0.041	7.29/0.039

((b)) $\boldsymbol{V} : PCA + 1EMiteration$

$\#Eigs$	Raw	Tnorm	Znorm	Ztnorm
0	8.93/0.044	8.30/0.041	8.29/0.043	7.59/0.041
50	9.74/0.046	8.79/0.042	8.58/0.041	7.56/0.040
100	9.32/0.045	8.50/0.041	8.25/0.042	7.47/0.039
150	9.29/0.044	8.50/0.041	8.32/0.042	7.52/0.039
200	9.23/0.044	8.51/0.041	8.41/0.042	7.65/0.039
250	9.24/0.045	8.42/0.041	8.38/0.042	7.67/0.039
300	9.24/0.045	8.42/0.041	8.36/0.042	7.52/0.039

((c)) $\boldsymbol{V} : PCA + 5EMiteration$

$\#Eigs$	Raw	Tnorm	Znorm	Ztnorm
0	8.93/0.044	8.30/0.041	8.29/0.043	7.59/0.041
50	9.83/0.046	8.84/0.041	8.59/0.042	7.61/0.040
100	9.36/0.045	8.50/0.041	8.30/0.042	7.48/0.039
150	9.40/0.045	8.50/0.041	8.42/0.042	7.59/0.039
200	9.25/0.045	8.62/0.041	8.46/0.042	7.72/0.039
250	9.29/0.045	8.49/0.041	8.41/0.042	7.74/0.039
300	9.25/0.045	8.53/0.041	8.48/0.042	7.70/0.039

((d)) $\boldsymbol{V} : PCA + 10EMiteration$

$\#Eigs$	Raw	Tnorm	Znorm	Ztnorm
0	8.93/0.044	8.30/0.041	8.29/0.043	7.59/0.041
50	9.81/0.046	8.86/0.042	8.66/0.042	7.64/0.039
100	9.40/0.045	8.46/0.041	8.34/0.042	7.47/0.039
150	9.32/0.045	8.50/0.041	8.42/0.042	7.67/0.039
200	9.23/0.045	8.59/0.041	8.54/0.042	7.75/0.039
250	9.24/0.045	8.58/0.041	8.46/0.042	7.67/0.039
300	9.25/0.045	8.62/0.041	8.53/0.042	7.68/0.039

Table C.3: *System: JFA. Results on SRE'08 female tel-tel condition by using different ML iterations on training the speaker variability subspace V as well as different number of eigenvoices. U is fixed over all the experiments and was trained via PCA keeping 50 eigenchannels.*

((**a**)) $\boldsymbol{V} : PCA$

$\#Eigs$	Raw	Tnorm	Znorm	Ztnorm
0	8.93/0.044	8.30/0.041	8.29/0.043	7.59/0.041
50	9.36/0.045	8.42/0.040	8.40/0.040	7.35/0.039
100	9.12/0.044	8.30/0.040	8.22/0.040	7.27/0.038
150	9.21/0.044	8.22/0.040	8.25/0.041	7.24/0.038
200	9.17/0.044	8.19/0.040	8.21/0.041	7.26/0.038
250	9.17/0.044	8.22/0.040	8.22/0.040	7.26/0.038
300	9.29/0.044	8.19/0.040	8.22/0.041	7.37/0.038

((**b**)) $\boldsymbol{V} : PCA + 1EMiteration$

$\#Eigs$	Raw	Tnorm	Znorm	Ztnorm
0	8.93/0.044	8.30/0.041	8.29/0.043	7.59/0.041
50	9.69/0.046	8.73/0.041	8.38/0.041	7.46/0.039
100	9.45/0.045	8.42/0.041	8.26/0.041	7.28/0.039
150	9.36/0.045	8.48/0.041	8.24/0.041	7.32/0.039
200	9.36/0.045	8.55/0.041	8.29/0.041	7.35/0.038
250	9.39/0.045	8.53/0.041	8.38/0.041	7.43/0.039
300	9.36/0.045	8.50/0.041	8.28/0.041	7.59/0.039

((**c**)) $\boldsymbol{V} : PCA + 5EMiteration$

$\#Eigs$	Raw	Tnorm	Znorm	Ztnorm
0	8.93/0.044	8.30/0.041	8.29/0.043	7.59/0.041
50	9.72/0.046	8.76/0.041	8.70/0.041	7.76/0.039
100	9.52/0.045	8.45/0.041	8.38/0.041	7.57/0.039
150	9.32/0.045	8.54/0.041	8.38/0.041	7.67/0.038
200	9.29/0.045	8.57/0.041	8.38/0.041	7.52/0.038
250	9.29/0.045	8.59/0.041	8.45/0.042	7.59/0.038
300	9.32/0.045	8.66/0.041	8.40/0.042	7.59/0.039

((**d**)) $\boldsymbol{V} : PCA + 10EMiteration$

$\#Eigs$	Raw	Tnorm	Znorm	Ztnorm
0	8.93/0.044	8.30/0.041	8.29/0.043	7.59/0.041
50	9.75/0.046	8.73/0.041	8.73/0.041	7.67/0.040
100	9.51/0.045	8.53/0.041	8.38/0.041	7.49/0.039
150	9.29/0.045	8.58/0.041	8.43/0.041	7.60/0.039
200	9.25/0.045	8.58/0.041	8.46/0.041	7.59/0.038
250	9.17/0.045	8.62/0.041	8.39/0.041	7.66/0.038
300	9.21/0.045	8.26/0.040	8.24/0.040	7.54/0.036

Table C.4: *System: JFA D trained on data. Results on SRE'08 female tel-tel condition by using different ML iterations on training the speaker variability subspace V as well as different number of eigenvoices. U is fixed over all the experiments and was trained via PCA keeping 50 eigenchannels.*

((**a**)) $\boldsymbol{U} : PCA$

$\#Eigs$	Raw	Tnorm	Znorm	Ztnorm
0	11.07/0.050	10.68/0.042	9.80/0.044	9.82/0.039
50	6.79/0.038	6.01/0.030	5.99/0.031	5.33/0.027
100	6.32/0.035	5.89/0.030	5.93/0.031	5.46/0.029
150	6.25/0.034	5.93/0.030	5.85/0.031	5.53/0.029
200	6.50/0.034	6.08/0.030	6.17/0.032	5.63/0.029
250	6.32/0.034	6.12/0.030	6.17/0.032	5.62/0.030
300	6.36/0.035	6.28/0.030	6.08/0.032	5.70/0.029

((**b**)) $\boldsymbol{U} : PCA + 1EMiteration$

$\#Eigs$	Raw	Tnorm	Znorm	Ztnorm
0	11.07/0.050	10.68/0.042	9.80/0.044	9.82/0.039
50	6.71/0.038	5.93/0.030	6.00/0.032	5.23/0.027
100	6.25/0.035	5.93/0.029	6.01/0.031	5.27/0.029
150	6.13/0.034	5.77/0.030	5.93/0.031	5.21/0.029
200	6.40/0.034	5.75/0.030	6.17/0.031	5.44/0.029
250	6.46/0.034	5.85/0.030	6.26/0.031	5.62/0.029
300	6.17/0.034	5.90/0.029	6.20/0.031	5.62/0.029

((**c**)) $\boldsymbol{U} : PCA + 5EMiteration$

$\#Eigs$	Raw	Tnorm	Znorm	Ztnorm
0	11.07/0.050	10.68/0.042	9.80/0.044	9.82/0.039
50	6.92/0.039	6.01/0.030	5.94/0.032	5.31/0.027
100	6.36/0.035	5.90/0.030	5.93/0.032	5.31/0.028
150	6.33/0.034	5.70/0.030	6.03/0.031	5.31/0.029
200	6.32/0.033	5.77/0.030	6.17/0.031	5.47/0.028
250	6.35/0.033	6.00/0.030	6.24/0.031	5.69/0.029
300	6.25/0.033	5.85/0.029	6.21/0.031	5.62/0.028

((**d**)) $\boldsymbol{U} : PCA + 10EMiteration$

$\#Eigs$	Raw	Tnorm	Znorm	Ztnorm
0	11.07/0.050	10.68/0.042	9.80/0.044	9.82/0.039
50	6.92/0.039	6.01/0.030	5.94/0.032	5.31/0.027
100	6.36/0.035	5.90/0.030	5.93/0.032	5.31/0.028
150	6.30/0.034	5.70/0.030	6.02/0.031	5.39/0.029
200	6.34/0.033	5.88/0.030	6.21/0.031	5.54/0.029
250	6.47/0.033	6.01/0.030	6.08/0.031	5.70/0.029
300	6.13/0.033	5.84/0.029	6.18/0.031	5.70/0.028

Table C.5: *System: MAP-SVC. Results on SRE'08 male tel-tel condition by using different ML iterations on training the session variability subspace U as well as different number of eigenchannels.*

((a)) $\boldsymbol{V} : PCA$

$\#Eigs$	Raw	Tnorm	Znorm	Ztnorm
0	6.79/0.038	6.01/0.030	5.99/0.031	5.33/0.027
50	13.56/0.071	10.76/0.052	11.07/0.053	8.58/0.040
100	10.76/0.060	8.89/0.043	8.87/0.045	7.10/0.032
150	9.66/0.056	8.11/0.038	8.19/0.041	6.32/0.030
200	9.35/0.053	7.67/0.036	7.88/0.038	6.17/0.030
250	8.96/0.051	7.48/0.036	7.57/0.038	6.01/0.029
300	8.81/0.049	7.33/0.034	7.48/0.037	6.01/0.028

((b)) $\boldsymbol{V} : PCA + 1EMiteration$

$\#Eigs$	Raw	Tnorm	Znorm	Ztnorm
0	6.79/0.038	6.01/0.030	5.99/0.031	5.33/0.027
50	14.45/0.074	12.08/0.055	11.15/0.055	9.05/0.041
100	11.26/0.062	10.04/0.044	9.13/0.046	7.33/0.034
150	10.53/0.058	9.28/0.041	8.11/0.042	7.01/0.032
200	9.98/0.055	8.50/0.039	7.88/0.039	6.79/0.031
250	9.81/0.052	8.50/0.037	7.73/0.038	6.48/0.030
300	9.51/0.051	8.27/0.036	7.48/0.037	6.56/0.030

((c)) $\boldsymbol{V} : PCA + 5EMiteration$

$\#Eigs$	Raw	Tnorm	Znorm	Ztnorm
0	6.79/0.038	6.01/0.030	5.99/0.031	5.33/0.027
50	14.55/0.073	12.25/0.056	11.22/0.056	9.13/0.041
100	11.38/0.062	10.29/0.045	9.28/0.046	7.42/0.035
150	10.45/0.058	9.44/0.041	8.27/0.042	7.13/0.032
200	10.13/0.054	8.84/0.039	8.05/0.039	6.94/0.031
250	9.84/0.052	8.58/0.037	7.80/0.038	6.63/0.030
300	9.59/0.050	8.42/0.036	7.57/0.038	6.48/0.030

((d)) $\boldsymbol{V} : PCA + 10EMiteration$

$\#Eigs$	Raw	Tnorm	Znorm	Ztnorm
0	6.79/0.038	6.01/0.030	5.99/0.031	5.33/0.027
50	14.50/0.073	12.16/0.057	11.27/0.056	9.13/0.041
100	11.45/0.063	10.30/0.045	9.20/0.046	7.31/0.035
150	10.41/0.057	9.36/0.041	8.42/0.042	7.16/0.031
200	10.13/0.054	8.81/0.038	7.88/0.039	7.04/0.031
250	9.77/0.052	8.44/0.037	7.80/0.038	6.56/0.029
300	9.44/0.050	8.42/0.035	7.41/0.038	6.63/0.030

Table C.6: *System: EV-SVC. Results on SRE'08 male tel-tel condition by using different ML iterations on training the speaker variability subspace V as well as different number of eigenvoices. U is fixed over all the experiments and was trained via PCA keeping 50 eigenchannels.*

((a)) $\boldsymbol{V}: PCA$

$\#Eigs$	Raw	Tnorm	Znorm	Ztnorm
0	6.79/0.038	6.01/0.030	5.99/0.031	5.33/0.027
50	8.27/0.044	6.44/0.032	6.19/0.033	5.16/0.026
100	7.93/0.043	6.23/0.031	6.32/0.032	5.08/0.025
150	7.65/0.042	6.25/0.030	6.15/0.032	5.16/0.025
200	7.36/0.042	6.16/0.030	6.08/0.031	5.02/0.025
250	7.28/0.041	6.25/0.030	6.08/0.031	5.08/0.025
300	7.25/0.041	6.26/0.030	6.01/0.031	5.09/0.025

((b)) $\boldsymbol{V}: PCA + 1EMiteration$

$\#Eigs$	Raw	Tnorm	Znorm	Ztnorm
0	6.79/0.038	6.01/0.030	5.99/0.031	5.33/0.027
50	8.58/0.046	7.33/0.034	6.48/0.034	5.39/0.027
100	8.34/0.044	7.02/0.033	6.17/0.034	5.31/0.027
150	7.96/0.043	6.94/0.032	6.08/0.034	5.44/0.026
200	7.88/0.043	6.87/0.031	6.06/0.033	5.39/0.026
250	7.96/0.042	6.87/0.031	6.13/0.032	5.39/0.027
300	7.73/0.041	6.79/0.031	6.07/0.032	5.39/0.026

((c)) $\boldsymbol{V}: PCA + 5EMiteration$

$\#Eigs$	Raw	Tnorm	Znorm	Ztnorm
0	6.79/0.038	6.01/0.030	5.99/0.031	5.33/0.027
50	8.42/0.046	7.28/0.034	6.52/0.034	5.31/0.027
100	8.42/0.044	7.02/0.033	6.22/0.034	5.47/0.027
150	8.19/0.043	6.98/0.032	6.24/0.034	5.39/0.026
200	7.96/0.042	6.85/0.031	6.17/0.033	5.59/0.027
250	7.85/0.042	6.93/0.031	6.17/0.033	5.62/0.027
300	7.85/0.041	6.79/0.031	6.17/0.033	5.62/0.026

((d)) $\boldsymbol{V}: PCA + 10EMiteration$

$\#Eigs$	Raw	Tnorm	Znorm	Ztnorm
0	6.79/0.038	6.01/0.030	5.99/0.031	5.33/0.027
50	8.47/0.046	7.32/0.034	6.48/0.034	5.34/0.027
100	8.34/0.044	7.07/0.033	6.29/0.034	5.39/0.027
150	8.23/0.043	6.92/0.032	6.25/0.034	5.47/0.026
200	7.85/0.042	6.94/0.032	6.17/0.033	5.47/0.027
250	7.73/0.041	6.79/0.031	6.25/0.033	5.57/0.026
300	7.85/0.041	6.79/0.031	6.21/0.033	5.62/0.026

Table C.7: *System: JFA. Results on SRE'08 male tel-tel condition by using different ML iterations on training the speaker variability subspace V as well as different number of eigenvoices. U is fixed over all the experiments and was trained via PCA keeping 50 eigenchannels.*

((**a**)) $\boldsymbol{V} : PCA$

$\#Eigs$	Raw	Tnorm	Znorm	Ztnorm
0	6.79/0.038	6.01/0.030	5.99/0.031	5.33/0.027
50	7.57/0.043	6.25/0.029	6.01/0.033	5.23/0.026
100	7.61/0.043	6.32/0.030	6.08/0.033	5.20/0.026
150	7.60/0.042	6.40/0.030	6.14/0.032	5.31/0.026
200	7.48/0.042	6.48/0.030	6.01/0.032	5.31/0.026
250	7.33/0.042	6.36/0.030	5.96/0.031	5.23/0.025
300	7.18/0.041	6.08/0.029	5.86/0.032	5.47/0.026

((**b**)) $\boldsymbol{V} : PCA + 1EMiteration$

$\#Eigs$	Raw	Tnorm	Znorm	Ztnorm
0	6.79/0.038	6.01/0.030	5.99/0.031	5.33/0.027
50	14.45/0.074	12.08/0.055	11.15/0.055	9.05/0.041
100	11.26/0.062	10.04/0.044	9.13/0.046	7.33/0.034
150	10.53/0.058	9.28/0.041	8.11/0.042	7.01/0.032
200	9.98/0.055	8.50/0.039	7.88/0.039	6.79/0.031
250	9.81/0.052	8.50/0.037	7.73/0.038	6.48/0.030
300	9.51/0.051	8.27/0.036	7.48/0.037	6.56/0.030

((**c**)) $\boldsymbol{V} : PCA + 5EMiteration$

$\#Eigs$	Raw	Tnorm	Znorm	Ztnorm
0	6.79/0.038	6.01/0.030	5.99/0.031	5.33/0.027
50	8.63/0.048	7.73/0.034	6.49/0.036	5.31/0.027
100	8.36/0.046	7.35/0.032	6.29/0.035	5.47/0.027
150	8.19/0.044	7.02/0.032	6.36/0.034	5.62/0.027
200	8.03/0.044	6.96/0.032	6.25/0.033	5.51/0.027
250	7.93/0.043	6.93/0.031	6.25/0.033	5.69/0.027
300	7.82/0.042	6.94/0.031	6.25/0.033	5.54/0.027

((**d**)) $\boldsymbol{V} : PCA + 10EMiteration$

$\#Eigs$	Raw	Tnorm	Znorm	Ztnorm
0	6.79/0.038	6.01/0.030	5.99/0.031	5.33/0.027
50	8.54/0.048	7.41/0.033	6.56/0.035	5.39/0.026
100	8.34/0.045	7.25/0.032	6.20/0.035	5.47/0.026
150	8.04/0.044	7.02/0.031	6.28/0.035	5.47/0.026
200	7.79/0.044	6.78/0.031	6.01/0.033	5.47/0.025
250	7.80/0.043	6.81/0.031	6.17/0.033	5.54/0.026
300	7.33/0.041	6.25/0.030	5.93/0.031	5.31/0.025

Table C.8: *System: JFA D trained on data. Results on SRE'08 male tel-tel condition by using different ML iterations on training the speaker variability subspace V as well as different number of eigenvoices. U is fixed over all the experiments and was trained via PCA keeping 50 eigenchannels.*

	Equal Error Rate (EER in %)											
	ATVS4			ATVS3			ATVS2			ATVS1		
	03s	10s	30s	03s	10s	30s	03s	10s	30s	03s	10s	30s
amha	11.53	3.72	1.12	11.41	2.36	0.47	7.40	1.64	0.40	7.51	0.89	0.07
bosn	12.02	5.23	2.17	10.84	3.62	1.28	7.56	2.19	1.30	7.93	2.58	1.32
cant	17.00	8.72	2.84	17.62	10.08	2.09	15.14	6.21	1.20	14.80	5.68	1.14
creo	15.68	5.54	2.48	20.21	4.84	2.16	13.29	2.68	1.97	12.97	2.79	1.26
croa	17.50	8.61	5.83	12.80	4.99	1.21	11.27	5.46	1.60	11.82	4.24	0.63
dari	16.68	9.12	4.20	19.35	7.37	2.74	15.70	6.31	2.22	16.69	5.36	1.48
fars	23.69	10.59	3.47	25.63	12.79	2.17	21.15	8.32	1.67	21.35	6.31	0.71
fren	26.20	14.71	6.83	27.07	16.77	7.67	22.74	11.39	4.32	23.79	9.78	1.81
geor	16.77	7.38	3.14	20.44	5.08	1.18	13.14	2.89	0.45	12.05	1.00	0.05
haus	13.94	4.54	1.96	17.77	5.27	1.17	11.81	2.09	0.71	10.32	1.18	0.04
hind	30.58	19.11	8.31	30.96	19.66	8.59	29.48	15.64	5.42	29.13	15.63	5.75
inen	23.08	12.71	4.49	36.16	26.20	11.71	24.42	12.63	3.18	24.29	10.92	2.13
kore	20.31	12.01	4.33	17.30	12.81	3.79	16.21	9.83	1.95	16.04	9.29	2.10
mand	21.93	11.52	1.81	21.24	11.68	1.43	19.13	9.34	0.60	19.64	8.16	0.66
pash	18.30	8.24	3.83	21.76	8.52	2.83	15.92	5.97	1.64	15.08	4.45	0.85
port	11.45	4.11	0.88	14.41	4.24	1.16	10.05	2.73	0.25	11.92	2.39	0.30
ruse	22.23	10.72	3.28	24.93	14.92	3.53	20.07	10.49	2.35	21.59	9.84	2.10
span	19.64	10.40	1.62	25.46	15.74	2.10	17.62	9.60	0.68	17.74	9.53	0.83
turk	15.58	3.66	1.69	13.13	2.72	0.38	8.83	0.88	0.15	10.17	0.70	0.02
ukra	8.65	2.23	0.80	10.23	2.65	0.73	6.96	0.83	0.36	8.24	0.89	0.31
urdu	30.08	20.27	11.93	27.92	19.11	9.08	27.75	19.21	8.03	27.82	18.13	8.45
usen	21.21	11.65	4.53	20.55	11.85	3.57	19.58	10.20	3.18	19.59	10.35	1.57
viet	18.30	11.59	4.33	21.35	13.57	3.71	16.77	10.27	2.71	18.95	8.94	2.22

Table C.9: *LRE'09 ATVS submitted systems performance (meanCavg x 100) on development dataset.*

	Equal Error Rate (EER in %)											
	ATVS4			ATVS3			ATVS2			ATVS1		
	03s	10s	30s	03s	10s	30s	03s	10s	30s	03s	10s	30s
amha	19.53	7.83	3.07	17.06	5.57	1.43	14.24	4.38	1.02	14.63	3.40	0.75
bosn	29.64	18.53	13.07	24.56	13.31	6.56	23.28	13.04	6.82	21.36	12.31	6.03
cant	16.12	7.48	4.24	19.39	10.05	5.69	13.42	5.58	2.76	11.80	4.32	2.91
creo	18.97	8.52	4.54	21.49	7.83	2.86	16.88	6.49	2.62	16.86	5.66	1.48
croa	24.58	15.81	10.88	19.72	10.50	6.21	20.37	12.01	6.77	20.99	11.89	6.21
dari	22.15	10.22	6.32	24.05	11.41	6.45	19.65	9.41	6.58	21.54	11.43	6.18
fars	18.66	6.82	3.70	21.05	8.31	3.67	15.38	5.74	1.91	17.21	6.54	2.15
fren	20.29	10.31	5.27	22.09	9.47	3.22	19.02	8.48	3.07	20.37	7.31	1.82
geor	17.99	8.96	4.87	18.58	6.70	2.24	13.05	5.57	1.66	14.24	4.21	1.03
haus	18.76	8.57	5.69	17.74	5.83	2.34	15.17	5.48	1.94	15.58	3.61	0.89
hind	23.32	10.42	7.54	25.73	13.19	8.20	21.40	9.19	5.65	21.21	8.65	6.00
inen	24.43	9.92	6.15	33.78	23.00	15.30	23.89	8.81	5.40	20.99	7.35	3.61
kore	16.81	6.22	3.17	17.84	6.08	2.66	13.42	3.73	1.27	12.54	3.27	1.09
mand	16.40	5.92	3.00	17.14	6.01	2.89	12.49	3.59	1.67	12.41	3.52	1.39
pash	22.87	12.15	5.23	24.91	11.82	5.12	21.03	9.91	3.75	20.95	9.84	3.28
port	17.34	7.53	1.86	18.10	6.70	1.49	14.68	5.10	1.31	14.93	4.88	0.73
ruse	19.44	6.69	2.82	19.74	7.85	5.18	16.32	3.96	2.01	16.06	4.62	1.61
span	16.37	6.43	3.39	15.71	4.80	0.79	11.34	3.72	0.57	12.92	3.12	0.62
turk	22.95	10.77	4.44	21.72	8.57	2.02	17.62	5.62	1.02	16.52	4.45	0.45
ukra	22.73	13.43	8.89	24.25	13.86	5.81	19.58	10.00	4.90	20.76	11.64	4.64
urdu	26.19	14.93	8.51	25.39	12.20	7.25	22.52	11.43	6.23	24.48	10.49	6.09
usen	16.64	7.95	6.31	18.00	7.59	5.11	14.04	5.32	4.08	15.89	5.85	4.00
viet	13.21	5.63	3.47	20.06	9.29	4.52	12.48	4.00	2.23	11.14	4.09	1.33

Table C.10: *LRE'09 ATVS submitted systems performance (meanCavg x 100) on evaluation dataset.*

References

K. A. and M. S. Intelligent Speaker Verification Based Biometric System for Electronic Commerce Applications. 14, August 25-27 2006. 5

A. Adami, R. Mihaescu, D. Reynolds, and J. Godfrey. Modeling Prosodic Dynamics for Speaker Recognition. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 4, pages IV–788–91, 2003. 26

W. E. Arnoldi. The Principle of Minimized Iterations in the Solution of the Matrix Eigenvalue Problem. *Quarterly of Applied Mathematics*, 9:17–29, 1951. 53

B. Atal. Automatic Speaker Recognition Based on Pitch Contours. *The Journal of the Acoustical Society of America*, 52(6):1687–1697, 1972. 3

B. Atal. Automatic Recognition of Speakers from Their Voices. In *Proceedings of the IEEE*, volume 64, pages 460–475, 1976. 3, 71

K. Atkinson. Language Identification from Nonsegmental Cues. *The Journal of the Acoustical Society of America*, 44:378A, 1968. 5

R. Auckenthaler, M. Carey, and H. Lloyd-Thomas. Score Normalization for Text-Independent Speaker Verification Systems. *Digital Signal Processing*, 10(1/2/3):42–54, 2000. 18

D. Bartholomew. *Latent Variable Models and Factor Analysis.* Charles Griffin Co. Ltd, London., 1987. 2, 8, 41, 42, 43

D. Bartholomew, M. Knott, and I. Moustaki. *Latent Variable Models and Factor Analysis: A Unified Approach*, volume 2nd editio. John Wiley Sons, 2011. URL http://eu.wiley.com/WileyCDA/. 8, 41

F. Bimbot, J.-F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-Garcia, D. Petrovska-Delacretaz, and D. A. Reynolds. A Tutorial on Text-Independent Speaker Verification. *Journal on Applied Signal Processing*, 2004(4):430–451, 2004. 1, 4, 8, 27

C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics).* Springer, 1st ed. 2006. corr. 2nd printing edition, Oct. 2007. ISBN 0387310738. URL http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20&path=ASIN/0387310738. 8, 22, 42, 52, 116

C. M. Bishop, M. Svensén, and C. K. I. Williams. GTM: The Generative Topographic Mapping. Technical Report NCRG/96/015, Neural Computing Research Group. Dept of Computer Science & Applied Mathematics. Aston University, Birminghan B4 7ET. United Kingdon, Apr. 1997. 42

P. D. Bricker and S. Pruzansky. Effects of Stimulus Content and Duration on Talker Identification. *The Journal of the Acoustical Society of America*, 40(6):1441–1449, 1966. URL http://scitation.aip.org/getabs/servlet/GetabsServlet?prog=normal&id=JASMAN000040000006001441000001&idtype=cvips&gifs=yes. 3

J. S. Bridle and M. D. Brown. An Experimental Automatic Word Recognition System. Technical report, Joint Speech Research Unit, Ruislip, England, 1974. 16

N. Brümmer, L. Burget, J. Cernocky, O. Glembek, F. Grezl, M. Karafiat, D. A. van Leeuwen, P. Matejka, P. Scwartz, and A. Strasheim. Fusion of Heterogeneous Speaker Recognition Systems in the STBU Submission for the NIST Speaker Recognition Evaluation 2006. *IEEE Transactions on Audio, Speech and Signal Processing*, 15(7):2072–2084, 2007. 19

N. Brümmer and J. du Preez. Application-Independent Evaluation of Speaker Detection. *Computer Speech & Language*, 20(2-3):230–275, 2006. 19, 20

N. Brümmer, A. Strasheim, V. Hubeika, P. Matejka, L. Burget, and . Glembek. Discriminative Acoustic Language Recognition Via Channel-Compensated GMM Statistics. In *Interspeech*, 2009. 62, 63

C. J. C. Burges. A Tutorial on Support Vector Machines for Pattern Recognition. *Data Min. Knowl. Discov.*, 2: 121–167, June 1998. ISSN 1384-5810. URL `http://portal.acm.org/citation.cfm?id=593419.593463`. 23

L. Burget, P. Schwarz, M. Agarwal, P. Akyazi, K. Feng, A. Ghoshal, O. Glembek, N. K. Goel, M. Karafiát, D. Povey, A. Rastrow, R. C. Rose, and S. Thomas. Multilingual Acoustic Modeling for Speech Recognition based on Subspace Gaussian Mixture Models. In *ICASSP 10*, pages 4334–4337, 2010. 117

W. Campbell, J. Campbell, D. Reynolds, D. Jones, and T. Leek. High-Level Speaker Verification with Support Vector Machines. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 73–76, 2004a. 26

W. Campbell, J. Campbell, D. Reynolds, D. Jones, and T. Leek. Phonetic Speaker Recognition with Support Vector Machines. *Advances in Neural Information Processing Systems*, 16, 2004b. 23

W. Campbell, J. Campbell, D. Reynolds, E. Singer, and P. Torres-Carrasquillo. Support Vector Machines for Speaker and Language Recognition. *Computer Speech & Language*, 20(2-3):210–229, 2006a. 1, 18, 23

W. M. Campbell, D. Sturim, and D. Reynolds. Support Vector Machines Using a GMM Supervectors for Speaker Verification. *IEEE Signal Processing Letters*, 13(5):308–311, 2006b. 24, 63

W. M. Campbell, D. Sturim, D. Reynolds, and A. Solomonoff. SVM Based Speaker Verification Using a GMM Supervector Kernel and NAP Variability Compensation. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 1, pages I–97–I–100, 2006c. 2, 60

W. M. Campbell, D. Sturim, P. Torres-Carrasquillo, and D. Reynolds. A Comparision of Subspace Feature-Domain Methods for Language Recognition. In *Proceedings of Interspeech 2008*, September 2008. 62, 63

F. Castaldo, D. Colibro, E. Dalmasso, P. Laface, and C. Vair. Compensation of Nuisance Factors for Speaker and Language Recognition. *IEEE Transactions on Audio, Speech and Language Processing*, 15(7):1969–1978, 2007. 1, 63

F. Castaldo, S. Cumani, P. Laface, and D. Colibro. Language Recognition using Language Factors. In *INTERSPEECH*, pages 176–179, 2009. 67

C. Cieri, W. Andrews, J. P. Campbell, and G. Doddington. The Mixer and Transcript Reading Corpora: Resources for Multilingual, Crosschannel Speaker Recognition Research,. LREC 2006: Fifth International Conference on Language Resources and Evaluation, 2006. 72, 73

C. Cieri, L. Corson, D. Graff, and K. Walker. Resources for New Research Directions in Speaker Recognition: The Mixer 3, 4 and 5 Corpora. Proc. Interspeech, 2007. 72, 73

P. Comon. Independent Component Analysis, a New Concept? *Signal Process.*, 36:287–314, April 1994. ISSN 0165-1684. URL http://portal.acm.org/citation.cfm?id=195302.195312. 42

C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995. URL http://dx.doi.org/10.1007/BF00994018. 23

A. de Castro, D. Ramos, and J. Gonzalez-Rodriguez. Forensic Speaker Recognition Using Traditional Features Comparing Automatic and Human-in-the-Loop Formant Tracking. In *INTERSPEECH*, pages 2343–2346, 2009. 117

N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet. Front-End Factor Analysis for Speaker Verification. *Audio, Speech, and Language Processing, IEEE Transactions on*, 19(4):788 – 798, February 2011. 116

J. R. Deller, J. H. L. Hansen, and J. L. Proakis. *Discrete-Time Processing of Speech Signals, 2nd Ed.* John Wiley and Sons, 1999. 12, 16

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977. ISSN 00359246. URL http://dx.doi.org/10.2307/2984875. 21, 46

G. Doddington. Speaker Recognition Based on Idiolectal Differences between Speakers. In *Eurospeech*, volume 4, pages 2521–2524, 2001. 12

G. Doddington, M. Przybocki, A. Martin, and D. Reynolds. The NIST Speaker Recognition Evaluation — Overview, Methodology, Systems, Results, Perspective. *Speech Communication*, 31(2-3):225–254, 2000. 71

R. Duda, P. Hart, and D. Stork. *Pattern Classification.* John Wiley and Sons Inc, New York City, New York, USA, 2001. 8

K. Farrell, R. Mammone, and K. Assaleh. Speaker Recognition using Neural Networks and Conventional Classifiers. *IEEE Trans. Speech Audio Process*, 2:194–205, 1994. 18

K. Fukunaga. *Introduction to Statistical Pattern Recognition.* Academic Press, San Diego, California, USA, 1990. 8, 52

S. Furui. Cepstral Analysis Technique for Automatic Speaker Verification. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 29:254–272, 1981. 2, 16, 29

M. Gales and P. Woodland. Mean and Variance Adaptation within the MLLR Framework. *Computer Speech Language*, 10:249–264, 1996. 36

J. Gauvain and C. Lee. Bayesian Learning of Gaussian Mixture Densities for Hidden Markov Models. In *DARPA Speech and Natural Language Workshop*, pages 272–277, 1991. 36

J. Gauvain and C. Lee. Maximum a Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains. *IEEE Transactions on Speech and Audio Processing*, 2(2):291–298, 1994. 22, 49

O. Glembek, L. Burget, N. Brummer, O. Plchot, and P. Matejka. Discriminatively trained i-vector extractor for speaker verification. In *INTERSPEECH*, 2011. 117

O. Glembek, L. Burget, N. Dehak, N. Brummer, and P. Kenny. Comparison of Scoring Methods Used in Speaker Recognition with Joint Factor Analysis. In *ICASSP '09: Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4057–4060, Washington, DC, USA, 2009. IEEE Computer Society. ISBN 978-1-4244-2353-8. 63

J. Godfrey and E. Holliman. *Switchboard-1 Release 2*. Linguistic Data Consortium, Philadelphia, USA, 1993. 72

J. Godfrey, E. Holliman, and J. Mcdaniel. Switchboard: Telephone Speech Corpus for Research and Development. In *Proceedings of ICASSP*, volume 1, pages 517–520, 1992. URL http://dx.doi.org/10.1109/ICASSP.1992.225858. 72

J. Gonzalez-Dominguez, B. Baker, R. Vogt, J. Gonzalez-Rodriguez, and S. Sridharan. On the Use of Factor Analysis with Restricted Target Data in Speaker Verification. In *Odyssey: The Speaker and Language Recognition Workshop, Brno, Czech Republic,*, June 28 - July 1 2010a. 6, 9, 102, 116

J. Gonzalez-Dominguez, I. Lopez-Moreno, J. Franco-Pedroso, D. Ramos, D. T. Toledano, and J. Gonzalez-Rodriguez. ATVS-UAM NIST SRE 2010 System. In *Proceedings of FALA 2010*, November 2010b. 8, 9, 65, 83

J. Gonzalez-Dominguez, I. Lopez-Moreno, J. Franco-Pedroso, D. Ramos, D. T. Toledano, and J. Gonzalez-Rodriguez. Eficiencia Computacional Y Alto Rendimiento En Reconocimiento Automático De Locutor. In *Actas de las V Jornadas de Reconocimiento Biométrico de Personas*, September 2010c. 9

J. Gonzalez-Dominguez, I. Lopez-Moreno, J. Franco-Pedroso, D. Ramos, D. T. Toledano, and J. Gonzalez-Rodriguez. Multilevel and Session Variability Compensated Language Recognition. *IEEE Journal on Selected Topics in Signal Processing*, 4(6):1084–1094, December 2010d. 9, 63, 65, 83, 94

J. Gonzalez-Dominguez, I. Lopez-Moreno, J. Franco-Pedroso, D. Ramos, D. T.Toledano, and J. Gonzalez-Rodriguez. Multilevel and Channel-Compensated Language Recognition: ATVS System at NIST LRE 2009. In *I Iberian SLTech - I Joint SIG-IL/Microsoft Workshop on Speech and Language*, September 2009. 9, 65

J. Gonzalez-Rodriguez. Speaker Recognition Using Temporal Contours in Linguistic Units: The Case of Formant and Formant-Bandwidth Trajectories. In *INTERSPEECH*, 2011. 117

J. Gonzalez-Rodriguez, D. Ramos-Castro, D. Torre-Toledano, A. Montero-Asenjo, J. Gonzalez-Dominguez, I. Lopez-Moreno, J. Fierrez-Aguilar, D. Garcia-Romero, and J. Ortega-Garcia. On the Use of High-Level Information for Speaker Recognition: The ATVS-UAM System at NIST SRE 2005. *IEEE Aerospace and Electronic Systems Magazine*, 22(1):15 – 21, January 2007a. 8, 27

J. Gonzalez-Rodriguez, P. Rose, D. Ramos, D. T. Toledano, and J. Ortega-Garcia. Emulating DNA: Rigorous Quantification of Evidential Weight in Transparent and Testable Forensic Speaker Recognition. *IEEE Transactions on Audio, Speech and Language Processing*, 15(7):2072–2084, 2007b. 3, 5, 6, 19, 101, 117

J. Gonzalez-Rodriguez, D. T. Toledano, and J. Ortega-Garcia. *Voice Biometrics*. Springer, 2007c. To appear. 4

D. Graff. CallHome and CallFriend Corpora in Various Languages. Linguistic Data Consortium, http://www.ldc.upenn.edu/Catalog/, 1996. 73

D. Graff. Voice of america. Linguistic Data Consortium, http://www.ldc.upenn.edu/Catalog/, 2009. 74

D. Graff, A. Canavan, and G. Zipperlen. Switchboard-2 Phase I. Available from: http://www.ldc.upenn.edu., 1998. 72

D. Graff, D. Miller, and K. Walker. Switchboard-2 Phase III. Available from: http://www.ldc.upenn.edu., 2002. 72

D. Graff, K. Walker, and A. Canavan. Switchboard-2 Phase II. Available from: http://www.ldc.upenn.edu., 1999. 72

D. Graff, K. Walker, and D. Miller. Switchboard cellular corpora: Parts 1 and 2. Available from: http://www.ldc.upenn.edu., 2001. 72

H. Hermansky, B. A. Hanson, and H. Wakita. Perceptually based Linear Predictive Analysis of Speech. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 509–512, 1985. 16

H. Hermansky and N. Morgan. RASTA Processing of Speech. *IEEE Transactions on Speech and Audio Processing*, 2(4):578–589, 1994. 2, 30

H. Hotelling. Analysis of a Complex of Statistical Variables into Principal Components. *Journal of Educational Psychology*, 24, 1933. 42

X. Huang, A. Acero, and H.-W. Hon. *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*. Prentice Hall PTR, Upper Saddle River, NJ, 2001. 12, 15, 16

A. Hyvärinen and E. Oja. Independent Component Analysis: Algorithms and Applications. *Neural Netw.*, 13: 411–430, May 2000. ISSN 0893-6080. URL `http://portal.acm.org/citation.cfm?id=351654.351659`. 42

D. James, H. peter Hutter, and F. Bimbot. CAVE project - Speaker Verification in Banking and Telecommunications, 1997. 5

P. Kenny. Bayesian Speaker Verification with Heavy-Tailed Priors. In *Odyssey: The Speaker and Language Recognition Workshop, Brno, Czech Republic,*, June 28 - July 1 2010. 116

P. Kenny, G. Boulianne, and P. Dumouchel. Eigenvoice Modeling With Sparse Training Data. *IEEE Trans. on Speech and Audio Processing*, 13(3):345–354, 2005a. 37, 51

P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel. Factor Analysis Simplified. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 1, pages 637–640, 2005b. 2, 58

P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel. Joint factor analysis versus eigenchannels in speaker recognition. *IEEE Transactions on Audio, Speech & Language Processing*, 15(4):1435–1447, 2007. 58

P. Kenny, N. Dehak, R. Dehak, V. Gupta, and P. Dumouchel. The Role of Speaker Factors in the NIST Extended Data Task. In *Odyssey: The Speaker and Language Recognition Workshop*, 2008a. 59

P. Kenny and P. Dumouchel. Disentangling Speaker and Channel Effects in Speaker Verification. In *in Proc. ICASSP*, pages 37–40, 2004a. 56

P. Kenny and P. Dumouchel. Experiments in Speaker Verification Using Factor Analysis Likelihood Ratios. In *Odyssey: The Speaker and Language Recognition Workshop*, pages 219–226, 2004b. 2, 27, 38, 51, 56

P. Kenny, M. Mihoubi, and P. Dumouchel. New Map Estimators for Speaker Recognition. In *INTERSPEECH*, 2003. 37

P. Kenny, P. Oullet, V. Dehak, N. Gupta, and P. Dumouchel. A Study of Interspeaker Variability in Speaker Verification. *IEEE Trans. on Audio, Speech and Language Processing*, 16(5):980–988, 2008b. 2, 59, 60

T. Kinnunen and H. Li. An Overview of Text-Independent Speaker Recognition: From Features to Supervectors, 2009. xix, 1, 4, 8, 13, 27

M. Kirby and L. Sirovich. Application of the Karhunen-Loeve Procedure for the Characterization of Human Faces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12:103–108, 1990. ISSN 0162-8828. 35

R. Kuhn, J. C. Junqua, P. Nguyen, and N. Niedzielski. Rapid Speaker Adaptation in Eigenvoice Space. *IEEE Transactions on Speech and Audio Processing*, 8(6):695–707, 2000. URL http://dx.doi.org/10.1109/89.876308. 36

H. J. Kunzel. Current Approaches to Forensic Speaker Recognition. In *ESCA Workshop on Automatic Speaker Recognition, Identification and Verification*, pages 135–141, Martigny, 1994. 117

C. Lanczos. An Iteration Method for the Solution of the Eigenvalue Problem of Linear Differential and Integral Operators. *Journal of Research of the National Bureau of Standards*, 45:255–282, 1950. 53

D. C. Lay. *Linear Algebra and Its Applications*. Addison Wesley, second edition, 1997. 8

P. F. Lazarsfeld and N. W. Henry. *Latent Structure Analysis*. Houghton-Mifflin, Boston., 1968. 41

K. F. Lee, H. W. Hon, and R. Reddy. An overview of the SPHINX speech recognition system. *Acoustics, Speech, and Signal Processing [see also IEEE Transactions on Signal Processing], IEEE Transactions on*, 38(1):35–45, 1990. URL http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=45616. 96

D. A. Leeuwen and N. Brümmer. An introduction to application-independent evaluation of speaker recognition systems. In C. Müller, editor, *Speaker Classification I*, pages 330–353. Springer-Verlag, Berlin, Heidelberg, 2007. ISBN 978-3-540-74186-2. URL http://dx.doi.org/10.1007/978-3-540-74200-5_19. 3, 20

R. B. Lehoucq, D. C. Sorensen, and C. Yang. Arpack Users Guide: Solution of Large Scale Eigenvalue Problems by Implicitly Restarted Arnoldi Methods., 1997. 53

Y. Linde, A. Buzo, and R. Gray. An Algorithm for Vector Quantizer Design. *Communications, IEEE Transactions on*, 28(1):84–95, Jan. 2003. URL http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1094577. 22

J. C. Loehlin. *Latent Variable Models: An Introduction to Factor, Path, and Structural Equation Analysis (Latent Variable Models: An Introduction to)*. Lawrence Erlbaum Associates, Inc., Jan. 2004. ISBN 0805849106. URL http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20&path=ASIN/0805849106. 8

I. Lopez-Moreno, D. Ramos, J. Gonzalez-Rodriguez, and D. T. Toledano. Anchor-Model Fusion for Language Recognition. In *Proceedings of Interspeech 2008*, September 2008. 19

S. Lucey and T. Chen. Improved Speaker Verification through Probabilistic Subspace Adaptation. In *Eurospeech*, pages 2021–2024, 2003. 37

J. Makhoul, F. Kubala, T. Leek, D. Liu, L. Nguyen, R. Schwartz, and A. Srivastava. Speech and language technologies for audio indexing and retrieval. *Proceedings of the IEEE*, 88(8):1338–1353, 2000. URL http://dx.doi.org/10.1109/5.880087. 6

J. I. Makhoul and J. J. Wolf. Linear Prediction and the Spectral Analysis of Speech. *IEEE Transactions on Audio Electroacoustic*, 21:140–148, 1973. 14, 15

M. I. Mandasari, M. McLaren, and D. van Leeuwen. Evaluation of i-vector Speaker Recognition Systems for Forensic Application. In *INTERSPEECH*, 2011. 116

J. D. Markel and S. B. Davis. Text-Independent Speaker Recognition from a Large Linguistically Unconstrained Time-Spaced Data Base. *IEEE Trans. on ASSP*, 27(1):74–82, 1979. 71

D. Matrouf, N. Scheffer, B. Fauve, and J.-F. Bonastre. A Straightforward and Efficient Implementation of the Factor Analysis Model for Speaker Verification. In *Proc. Interspeech*, pages 1242–1245, 2007. 63

A. Montero-Asenjo, J. Gonzalez-Dominguez, D. Ramos-Castro, I. Lopez-Moreno, D. T. Toledano, and J. Gonzalez-Rodriguez. On the Use of High-Level Information in Speaker and Language Recognition. In *Proceedings of the Spanish Network of Speech Technologies Workshop (ISBN 84-96214-82-6)*, pages 15–18, November 2006. 8, 9, 27

A. Moreno, D. Poch, A. Bonafonte, E. Lleida, J. Llisterri, J. Mariño, and C. Nadeu. ALBAYZÍN Speech Database: Design of the Phonetic Corpus. In *European Conference on Speech Communication and Technology, Eurospeech*, volume 1, pages 175–178, 1993. 96

Y. Muthusamy, K. Berkling, T. Arai, R. Cole, and E. Barnard. A comparison of approaches to automatic language identification using telephone speech. In *Proc Eurospeech*, pages 1307–1310, 1993. 5

Y. K. Muthusamy, E. Barnard, and R. A. Cole. Reviewing Automatic Language Identification, 1994. URL http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.56.7425. 5

NIST. The 2006 NIST SV Evaluation Plan. http://www.nist.gov/speech/tests/spk/2006/, 2006. 79

NIST. The 2009 NIST SLR Evaluation Plan. www.itl.nist.gov/iad/mig/tests/lre/2009/LRE09_EvalPlan_v6.pdf, 2009. 5, 79

NIST. NIST Speech Group Website. http://www.nist.gov/speech, 2010. 59

J. Ortega-Garcia, J. Gonzalez-Rodriguez, and V. Marrero-Aguiar. Ahumada: A large speech corpus in spanish for speaker characterization and identification. *Speech Communication*, 31(2-3):255 – 264, 2000. ISSN 0167-6393. URL http://www.sciencedirect.com/science/article/pii/S0167639399000813. 72

B. Paltridge. Thesis and dissertation writing: an examination of published advice and actual pratice. *English for specific purposes*, 21:125–143, 2002. 7

K. Pearson. On Lines and Planes of Closest Fit to Systems of Points in Space. *Philosophical Magazine*, 2(6): 559–572, 1901. 35, 42

J. Pelecanos, U. Chaudhari, and G. Ramaswamy. Compensation of Utterance Length for Speaker Verification. In S. Proc. of Odyssey 2004, Toledo, editor, *A Speaker Odyssey, The Speaker Recognition Workshop*, pages 161–164, 2004. 108

J. Pelecanos and S. Sridharan. Feature Warping for Robust Speaker Verification. In *Proceeding Odyssey*, pages 213–218, 2001. 2, 30

O. Perez-cruz and O. Bousquet. Kernel methods and their potential use in signal processing. In *IEEE Signal Processing Magazine*, volume 21, pages 57–65, 2004. 23

S. Perez-Gomez, D. Ramos, J. Gonzalez-Dominguez, and J. Gonzalez-Rodriguez. Score-Level Compensation of Extreme Speech Duration Variability in Speaker Verification. In *Proceedings of Interspeech 2010*, September 2010. 9, 108, 116

D. Povey, L. Burget, M. Agarwal, P. Akyazi, K. Feng, A. Ghoshal, O. Glembek, N. K. Goel, M. Karafiát, A. Rastrow, R. C. Rose, P. Schwarz, and S. Thomas. Subspace Gaussian Mixture Models for Speech Recognition. In *ICASSP*, pages 4330–4333, 2010. 117

M. A. Przybocki, A. F. Martin, and A. N. Le. NIST Speaker Recognition Evaluations Utilizing the Mixer Corpora-2004, 2005, 2006. *IEEE Transactions on Audio, Speech and Language Processing*, 15(7):1951–1959, 2007. 5

L. Rabiner and H. Juang. An Introduction to Hidden Markov Models. *IEEE Acoustics, Speech and Signal Processing Magazine*, 3(1):4–16, 1986. 18

L. R. Rabiner and R. W. Schafer. *Digital Processing of Speech Signals*. Prentice Hall, 1978. 12

D. Ramos. *Forensic Evaluation of the Evidence Using Automatic Speaker Recognition Systems*. PhD thesis, Universidad Autonoma de Madrid, November 2007. 3, 6, 19, 101

D. Ramos, J. Gonzalez-Dominguez, E. Arevalo, and J. Gonzalez-Rodriguez. High-Performance Session Variability Compensation in Forensic Automatic Speaker Recognition. In *Proceedings of 2nd Pan American Meeting on Acoustics. Acoustical Society of America. ISSN 0001-4966.*, page 2378. Acoustical Society of America., November 2010. 6, 9

D. Ramos, J. Gonzalez-Dominguez, D. T. Toledano, and J. Gonzalez-Rodriguez. *Speaker Feature*. Springer, July 2009. 8, 12, 13

D. Ramos, J. Gonzalez-Rodriguez, and J. Gonzalez-Dominguez, J. Lucena. Adressing Database Mismatch in Forensic Speaker Recognition with Ahumada III: A Public Real-Casework Database in Spanish. In *Interspeech*, 2008. 7, 9, 73, 81, 102, 116

D. Reynolds. The Effects of Handset Variability on Speaker Recognition Performance: Experiments on the Switchboard Corpus. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 113–116, 1996. 1

D. Reynolds. An Overview of Automatic Speaker Recognition Technology. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 4, pages 4072–4075, 2002. 1, 8

D. Reynolds. Channel Robust Speaker Verification Via Feature Mapping. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 53–56, 2003. 2, 31, 61

D. Reynolds, W. Andrews, J. Campbell, J. Navratil, B. Peskin, A. Adami, Q. Jin, D. Klusacek, J. Abramson, R. Mihaescu, J. Godfrey, D. Jones, and B. Xiang. The SuperSID Project: Exploiting High-Level Information for High-Accuracy Speaker Recognition. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 4, pages 784–787, 2003. 12

D. Reynolds, T. Quatieri, and R. Dunn. Speaker Verification Using Adapted Gaussian Mixture Models. *Digital Signal Processing*, 10(1/2/3):19–41, 2000. 20, 22, 56

D. Reynolds and R. Rose. Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models. *IEEE Transactions on Speech and Audio Processing*, 3(1):72–83, 1995. 18

P. Rose. Technical Forensic Speaker Recognition: Evaluation, Types and Testing of Evidence. *Computer Speech & Language*, 20(2-3):159–191, 2006. 117

D. Rubin and D. Thayer. Em Algorithms for Ml Factor Analysis. *Psychometrika*, 47(1):69–76, March 1982. URL http://ideas.repec.org/a/spr/psycho/v47y1982i1p69-76.html. 8

H. Sakoe. Dynamic Programming Algorithm Optimization for Spoken Word Recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26:43–49, 1978. 17

N. Scheffer, R. Vogt, S. Kajarekar, and J. Pelecanos. Combination strategies for a factor analysis phone-conditioned speaker verification system. *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, 0:4053–4056, 2009. 109, 110

M. Senoussaoui, P. Kenny, N. Dehak, and P. Dumouchel. An I-Vector Extractor Suitable for Speaker Recognition with Both Microphone and Telephone Speech. In *Odyssey: The Speaker and Language Recognition Workshop, Brno, Czech Republic,*, June 28 - July 1 2010. 116

E. Shriberg, L. Ferrer, S. Kajarekar, A. Venkataraman, and A. Stolcke. Modeling Prosodic Feature Sequences for Speaker Recognition. *Speech Communication*, 46(3-4):455–472, 2005. 23

A. Solomonoff, W. Campbell, and I. Boardman. Advances in Channel Compensation for SVM Speaker Recognition. In *ICASSP*, volume I, pages 629–632, 2005. 63

A. Solomonoff, C. Quillen, and W. Campbell. Channel Compensation for SVM Speaker Recognition. In *Odyssey: The Speaker and Language Recognition Workshop*, pages 57–62, 2004. 60

F. Soong and A. Rosenberg. A vector quantization approach to speaker recognition. *AT T Technical Journal*, 66: 14–26, 1987. URL http://maxwell.me.gu.edu.au/spl/teach/adsp/Proj/Lit/2.pdf. 17

D. C. Sorensen and F. M. Gomes. Arpack++: A C++ Implementation of Arpack Eigenvalue Package. 1997. URL http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.53.8506. 53

C. Spearman. General Intelligence, Objectively Determined and Measured. *American Journal of Psychology*, 15 (2):201–293, 1904. URL http://www.jstor.org/stable/1412107. 41

G. Strang. *Introduction to Linear Algebra*. Wesley-Cambridge Press, 3rd edition, 2003. 8, 120

D. Sturim, W. Campbell, D.A.Reynolds, R.B.Dunn, and TF.Quatieri. Robust Speaker Recognition with Cross-channel Data: Mit-ll Results on the 2006 nist sre Auxiliary Microphone Task. In *ICASSP*, volume I, pages 629–632, 2007. 1

R. Teunen, B. Shahshahani, and L. Heck. A Model-Based Transformational Approach to Robust Speaker Recognition. In *International Conference on Spoken Language Processing*, volume 2, pages 495–498, 2000. 2, 31

S. Theodoridis and K. Koutroumbas. *Pattern Recognition*. Academic Press, 2003. 8

O. Thyes, R. Kuhn, P. Nguyen, and J. Junqua. Speaker Identification and Verification Using Eigenvoices. In *International Conference on Spoken Language Processing*, volume 2, pages 242–245, 2000. 37

D. T. Toledano, J. Gonzalez-Dominguez, A. Abejón-Gonzalez, D. Spada, I. Mateos-Garcia, and J. Gonzalez-Rodriguez. Improved Language Recognition Using Better Phonetic Decoders and Fusion with MFCC and SDC Features. In *INTERSPEECH*, pages 194–197, 2007. 9, 27

D. T. Toledano, D. Ramos, J. Gonzalez-Dominguez, and J. Gonzalez-Rodriguez. *Speech Analysis*. Springer, July 2009. 8

P. A. Torres-Carrasquillo, E. Singer, M. A. Kohler, and J. R. Deller. Approaches to Language Identification Using Gaussian Mixture Models and Shifted Delta Cepstral Features. In *ICSLP*, volume 1, pages 89–92, 2002. 1, 16

S. Tsekeridou and I. Pitas. Speaker dependent video indexing based on audio-visual interaction. In *ICIP (1)*, pages 358–362, 1998. 5

M. A. Turk and A. P. Pentland. Eigenfaces for Recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86, 1991. 34, 35

C. Vair, D. Colibro, F. Castaldo, E. Dalmasso, and P. Laface. Channel Factors Compensation in Model and Feature Domain for Speaker Recognition. In *Speaker and Language Recognition Workshop, 2006. IEEE Odyssey 2006*, pages 1 – 6, 2006. 2, 59, 61

J. Villalba and N. Brummer. Towards Fully Bayesian Speaker Recognition: Integrating out the between-Speaker Covariance. In *INTERSPEECH*, 2011. 116

M. Viswanathan, H. S. M. Beigi, S. Dharanipragada, F. Maali, and A. Tritschler. Multimedia Document Retrieval Using Speech and Speaker Recognition. *International Journal on Document Analysis and Recognition*, 2(4): 147–162, 2000-06-01. URL `http://dx.doi.org/10.1007/s100320050002`. 5

R. Vogt, B. Baker, and S. Sridharan. Factor analysis subspace estimation for speaker verification with short utterances. In *INTERSPEECH*, pages 853–856, 2008a. 108

R. Vogt, C. Lustri, and S. Sridharan. Factor Analysis Modelling for Speaker Verification with Short Utterances. In *Odyssey: The Speaker and Language Recognition Workshop*, 2008b. 51, 108

R. Vogt and S. Sridharan. Explicit Modeling of Session Variability for Speaker Verification. *Computer Speech & Language*, 22(1):17–38, 2008. ISSN 0885-2308. 2, 59

S. Watanabe. Karhunen-Loeve Expansion and Factor Analysis Theoretical Remarks and Applications. *Transactions of the Fourth Prague Conference*, (635–660), 1965. 35

S. J. Young, G. Evermann, M. J. F. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. C. Woodland. *The HTK Book, version 3.4*. Cambridge University Engineering Department, Cambridge, UK, 2006. 96

D. Zhang. *Biometrics Solutions for Authentication in an e-World*. Number ISBN 1-4020-7142-6. Kluwer Academic Publishers, Dordrecht, 2002. 5

M. Zissman. Comparison of Four Approaches to Automatic Language Identification of Telephone Speech. *IEEE Trans. Acoust., Speech, Signal Processing*, 4(1):31–44, 1996. 5, 27

M. Zissman and E. Singer. Automatic Language Identification of Telephone Speech Messages Using Phoneme Recognition and N-Gram Modelling. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 1, pages 305–308, 1994. 5

M. A. Zissman and K. Berkling. Automatic Language Identification. *Speech Communication*, 35(1-2):115–124, 2001. 5