ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

UNIVERSITAT POLITÈCNICA DE CATALUNYA

Faculté Sciences et Techniques de l'Ingénieur
Institut de Traitement des Signaux
Section de Génie Électrique et Électronique

Master Thesis

Joint Factor Analysis for Forensic Automatic Speaker Recognition

by

Víctor Alonso Moreno

Supervisor: Dr A. Drygajlo

Lausanne, EPFL
Juin, 2011

to my parents

Contents

List of Figures

List of Tables

Acknowledgements

I would like to thanks my thesis supervisor Dr. Andrzej Drygajlo who gave me the opportunity to work on the topic of forensic speaker recognition and on the speech processing state-of-the-art techniques.

My most special thanks to my parents, I thank them for believing in me and helping me believe in myself; for giving me the opportunity to study abroad and for always having supported me during my entire live.

I would also like to say a special thanks to all my friends and colleagues for their friendship and support that has seen me through the highs and lows of these last six month. In particular, I thank Bruno, Dídac, Johnny and Javi Martínez for coming to Lausanne to visit me. I also want to thank all my new friends I met here.

Abstract

Nowadays, under controlled recording conditions, the state-of-the-art automatic speaker recognition systems show very good performance in discriminating between voices of speakers. However, in investigative activities (e.g., anonymous calls and wire-tapping) the conditions in which recordings are made can not be controlled and pose a challenge to automatic speaker recognition. Some factors that introduce variability in the recordings can be the differences in the phone handset, in the transmission channel and in the recording devices.

The strength of evidence, estimated using statistical models of within-source variability and between-sources variability, is expressed as a likelihood ratio, i.e., the probability of observing the features of the questioned recording in the statistical model of the suspected speaker's voice given the two competing hypotheses: the suspected speaker is the source of the questioned recording and the speaker at the origin of the questioned recording is not the suspected speaker.

The main unresolved problem in forensic automatic speaker recognition today is that of handling mismatch in recording conditions. This mismatch has to be considered in the estimation of the likelihood ratio because it can introduce important errors.

In this work, we handle and analyze this state-of-the-art system. The forensic automatic speaker recognition system consists of many parts, such as feature extraction and modeling. We have focused on the modeling part, training models which can be decomposed in two spaces, the speaker and session subspace.

This technique, called Joint Factor Analysis, is the state-of-the-art in the speaker verification systems. Using the property of decomposition in two subspaces, we try to solve the problem of mismatched conditions adapting the session subspace of the train recordings to a new session subspace (which is under different conditions).

To estimate the speaker and session subspaces, we need some databases, e.g. one database containing the traces, and another containing recordings from the suspect. These databases must be recorded in several conditions to simulate a real forensic case where mismatched is present. Examples to such recording conditions are cellular phones or fixed telephone network.

Finally, an evaluation of the system is presented at the end of the work. Thanks to this evaluation, we see which recording conditions degrade more the results, what effect the mismatch have on the results and, how much the adaptation can fix these effects.

Version abrégée

A l'heure actuelle, quand les conditions sont contrôlés, les systèmes de reconnaissance automatique de locuteur possèdent d'excellentes performances lorsqu'il s'agit de discriminer entre des voix de locuteurs. Cependant, dans les activités d'investigation (par exemple, les appels anonymes et les écoutes téléphoniques) les conditions dans lesquelles les enregistrements sont effectués ne peuvent être contrôlés et posent un défi à la reconnaissance automatique de locuteurs. Certains facteurs qui introduisent une variabilité dans les enregistrements peuvent être les différences dans les combinés téléphoniques, dans le canal de transmission et dans l'appareil d'enregistrement.

La force de la preuve, estimée à l'aide de modèles statistiques des intra- et inter-variabilités de la source, est exprimée sous la forme d'un rapport de vraisemblance, i.e., la probabilité d'observer les caractéristiques de l'enregistrement en question dans le modèle statistique de la voix du suspect étant donné les deux hypothèses : le suspect est la source de l'enregistrement en question et le locuteur à l'origine de l'enregistrement en question n'est pas le suspect.

Le principal problème non résolu dans la reconnaissance automatique de locuteurs en sciences forensiques est la manière de traiter les conditions d'enregistrement différentes. Les conditions d'enregistrement différentes doivent être considérées dans l'estimation du rapport de vraisemblance, car elles peuvent introduire des erreurs importantes.

Dans ce travail, on traite et analyse ce système état d'oeuvre. Le système de reconnaissance automatique du locuteur pour des fins juridiques consiste de plusieurs parties, tel que l'extraction des caractéristiques et le modelage. Nous nous avons concentré sur la partie de modélisation, la formation des modèles qui peuvent être décomposés en deux espaces, le sous-espace de locuteur et de session.

Cette technique, appelée Analyse Factorielle Commune (JFA), est l'état d'ouvre dans les systèmes de vérification du locuteur. En utilisant la propriété de décomposition en deux sous-espace, nous essayons de résoudre le problème de conditions différentes en adaptant le sous-espace de session des enregistrements d'entraînement à un nouveau sous-espace de session (qui est sous des conditions différentes).

Pour estimer le sous-espace de locuteur et de session, nous avons besoin de certaines bases de données, par exemple une base de données contenant les traces, et un autre contenant des enregistrements du suspect. Ces bases de données doivent être enregistrées dans plusieurs conditions pour simuler un cas réel juridique où incompatibles est présent. Quelques exemples de telles conditions sont les téléphones portables et le reseau de téléphone fixe.

Dernièrement, une évaluation du système est présentée à la fin du travail. Grâce à cette évaluation, on peut voir quelles conditions d'enregistrement dégradent plus les résultats, quel effet a le désaccord sur les résultats, et combien l'adaptation peut fixer ces effets.

Chapter 1

Introduction

1.1 Forensic Automatic Speaker Recognition

Automatic speaker recognition systems, that have been shown to perform a high accuracy in controlled conditions, are an attractive option for forensic speaker recognition tasks because forensic cases often include large amounts of audio data which are difficult to evaluate within the time constraints of an investigation or analysis required by the courts. The traditional aural-perceptual and semi-automatic speaker recognition techniques used in forensic speaker recognition can be complemented by the automatic recognition systems. These traditional techniques require a high degree of mastery of a language and its nuances, and experience in extracting and comparing relevant characteristics. Modern criminal activity spans several countries, there may be cases in which there is a need to analyze speech in languages where sufficient expertise is unavailable.

The last years, the interest in the use of automatic speaker recognition techniques for forensic has increase and several research groups around the world have been working in this problem. One of the requirements of this systems is that the methods used and the results must be understandable an interpretable by the courts.

1.2 Mismatched recording conditions

In forensic speaker recognition caseworks, the recordings analyzed often differ due to telephone channel distortions, ambient noise in the recording environments, the recording devices, as well as their linguistic content and duration. These factors may influence aural, instrumental and automatic speaker recognition. In many cases, the recordings are provided by the police or the court and the forensic expert does not have a choice in defining the recording conditions for the suspect, and additional recordings cannot be made. If there is a mismatch in the technical (encoding and transmission) and acoustic conditions between the recordings of the databases used, erroneous or misleading results can appear when comparisons between them are made, and therefore, it is a prior necessity to reduce and quantify the effect of the mismatch.

In this project, we focus on forensic automatic speaker recognition and the effect of mismatched recording conditions of the databases used on the strength of evidence and how to solve the problem using the state-of-the-art techniques.

1.3 Joint Factor Analysis

In state-of-the-art methods of speaker verification, speaker variability is assumed to be of primary importance but it has long been recognized that session variability is a serious problem. If a systematic model of session variability is integrated with an effective model of speaker variability could prove to be useful in speaker verification. As a first attempt at this problem, a model of session variability was proposed in [11] which was referred as eigenchannel MAP. In [5] was showed how this model can be integrated with standard models of speaker variability, namely classical MAP [12] and eigenvoice MAP [10], to produce a model of speaker and session variability which was referred as joint factor analysis.

The purpose of this project is to find a method that allows us to obtain a speaker model in a way which is immune to the channel effects or at least one that allows us to adapt the mismatched conditions. We assume that each speaker- and channel-dependent supervector can be decomposed into a sum of two supervectors, one of which lies in the speaker space and the other in the channel space. Given an enrollment recording for a speaker we can disentangle the speaker and channel effects in the corresponding speaker- and channel-dependent supervector by calculating the joint posterior distribution of the speaker and channel factors. An estimate of the speaker supervector which is immune to the channel effects in the enrollment recording can be obtained (in theory at least) by suppressing the contribution of the channel factors, and this estimation can be adapted to new session conditions.

1.4 Objectives of the thesis

The main goal of this project is to measure and compensate the effects of mismatch that arise in forensic case conditions due to the technical (encoding and transmission) and acoustic conditions of the recordings of the databases used. For this purpose, the joint factor analysis technique was proposed to solve the problem.

1.5 Organization of the thesis

This thesis is organized as follows:

- Chapter 1: Introduction and presentation of the objectives and contributions of the thesis.

- Chapter 2: Discussion of the state-of-the-art technique for speaker recognition and verification systems, the joint factor analysis.

- Chapter 3: Description of the process followed to create and adapt a speaker model using the joint factor analysis.

- Chapter 4: Evaluation of the results obtained in all the cases, matched conditions, mismatched conditions and adapted conditions.

- Chapter 5: The summary and conclusion of the thesis, with a discussion of possible extensions of the present work.

- Appendix A: Description of forensic speaker recognition databases used for the validation of the methods.

Chapter 2

Joint Factor Analysis

In this chapter, the theory underlying the Joint Factor Analysis technique will be described. Furthermore, we present some algorithms and methods to compute and estimate the parameters of this kind of model (also called hyperparameters). The mathematic development and simplification of the algorithms can be found in [5] and [6].

2.1 The JFA model

The theory underlying classical MAP, eigenvoice MAP and eigenchannel MAP are combined to create the factor analysis model. We assume a fixed GMM structure containing a total of C mixture components and an acoustic feature vector of dimension F.

The decomposition of the speaker- and channel-dependent supervector into a sum of two supervectors, one of which depends on the speaker and the other on the channel, is the basic principle of this technique. The speaker and channel supervectors are statistically independent and normally distributed. The dimensions of the covariance matrices of these distributions are $(CF\,x\,CF)$.

Let $\mathbf{M}(s)$ be the speaker supervector for a speaker s and let $\mathbf{m}$ denote the speaker- and channel-independent supervector. The way to estimate $\mathbf{m}$ is to take the supervector from a Universal Background Model (UBM). In classical MAP it is assumed that, for a randomly chosen speaker s, $\mathbf{M}(s)$ is normally distributed with mean $\mathbf{m}$ and a diagonal covariance matrix $\mathbf{d}^2$. The next expression describe it in terms of hidden variables:

$$\mathbf{M}(s) = \mathbf{m} + \mathbf{dz}(s) \tag{2.1}$$

where $\mathbf{z}(s)$ is a hidden vector distributed according to the standard normal density, $N(\mathbf{z}|\mathbf{0},\ \boldsymbol{I})$. (The expectation of $\mathbf{M}(s)$ is $\mathbf{m}$ and its covariance is $\mathbf{d}^2$.)

Only the mixture components observed in the adaptation data can be updated using the MAP adaptation. Thus, if the number of mixture components C is large, classical MAP tends to saturate slowly in the sense that large amounts of enrollment data are needed to use it to full advantage.

The pressence of a rectangular matrix $\mathbf{v}$ of dimensions $CF\,x\,R$ where $R \ll CF$ is assumed in eigenvoice MAP. Thus, for a randomly chosen speaker s,

$$\mathbf{M}(s) = \mathbf{m} + \mathbf{v}\mathbf{y}(s) \tag{2.2}$$

where $\mathbf{y}(s)$ is a hidden $R\,x\,1$ vector having a standard normal distribution. Eigenvoice MAP tends to saturate much more quickly than classical MAP since the dimension of $\mathbf{y}(s)$ is smaller than that of $\mathbf{z}(s)$. This approach to speaker adaptation suffers from the drawback that, in estimating $\mathbf{v}$ from a given training corpus, it is necessary to assume that R is less than or equal to the number of training speakers [10]. Hence, to estimate $\mathbf{v}$ properly, a large number of training speakers are needed. The eigenvoice MAP estimate of the speaker's supervector is constrained to lie in the subspace spanned by the training speaker's supervectors even if the 'true' speaker supervector lies elsewhere.

Classical MAP and eigenvoice MAP complement each other due to the strengths and weaknesses of each other. (Eigenvoice MAP is preferable if small amounts of data are available for speaker adaptation and classical MAP if large amounts are available.) The next combination strategy assume a decomposition of the form

$$\mathbf{M}(s) = \mathbf{m} + \mathbf{v}\mathbf{y}(s) + \mathbf{d}\mathbf{z}(s) \tag{2.3}$$

where $\mathbf{y}(s)$ and $\mathbf{z}(s)$ are assumed to be independent and to have standard normal distributions. In other words, $\mathbf{M}(s)$ is assumed to be normally distributed with mean $\mathbf{m}$ and covariance matrix $\mathbf{v}\mathbf{v}^* + \mathbf{d}^2$. The components of $\mathbf{y}(s)$ are called *common speaker factors* and the components of $\mathbf{z}(s)$ are *special speaker factors*; $\mathbf{v}$ and $\mathbf{d}$ are *factor loading matrices.* The *speaker space* is the affine space defined by translating the range of $\mathbf{v}\mathbf{v}^*$ by $\mathbf{m}$. If $\mathbf{d}\mathbf{=0}$, then all speaker supervectors are contained in the speaker space; in the general case ($\mathbf{d} \neq \mathbf{0}$) the term $\mathbf{d}\mathbf{z}(s)$ serves as a residual which compensates for the fact that this type of subspace constraint may not be realistic.

In order to incorporate channel effects, we suppose a set of given recordings $h = 1, ..., H(s)$ of a speaker s. For each recording h, let $\mathbf{M}_{h(s)}$ denote the corresponding speaker- and channel-dependent supervector. As in [11], the difference between $\mathbf{M}_h(s)$ and $\mathbf{M}(s)$ can be accounted for by a vector of *common channel factors* $\mathbf{x}_h(s)$ having a standard normal distribution. That is, we assume that there is a rectangular matrix $\mathbf{u}$ of low rank (the *loading matrix* for the *channel factors*) such that

$$\begin{aligned} \mathbf{M}(s) &= \mathbf{m} + \mathbf{v}\mathbf{y}(s) + \mathbf{d}\mathbf{z}(s) \\ \mathbf{M}_h(s) &= \mathbf{M}(s) + \mathbf{u}\mathbf{x}_h(s) \end{aligned} \tag{2.4}$$

for each recording $h = 1, ..., H(s)$. An important detail is that the *speaker factors* are assumed to have the same values for all recordings of the speaker whereas the *channel factors* vary from one recording to another. We refer as the *channel space* the low-dimensional subspace of the supervector space, namely the range of $\mathbf{u}\mathbf{u}^*$.

Thus, in its current form, the joint factor analysis model is specified as follows. If R_C is the number of *channel factors* and R_S the number of *speaker factors*, the model is specified by a quintuple $\mathbf{\Lambda}$ of the form ($\mathbf{m}$, $\mathbf{u}$, $\mathbf{v}$, $\mathbf{d}$, $\mathbf{\Sigma}$) where $\mathbf{m}$ is $CF\,x\,1$, $\mathbf{u}$ is $CF\,x\,R_C$, $\mathbf{v}$ is $CF\,x\,R_S$, and $\mathbf{d}$ and $\mathbf{\Sigma}$ are $CF\,x\,CF$ diagonal matrices. To explain the role of $\mathbf{\Sigma}$, fix a mixture component c and let Σ_c be the corresponding block of $\mathbf{\Sigma}$. For each speaker s and recording h, let $M_{hc}(s)$ denote the subvector of $\mathbf{M}_h(s)$ corresponding to the given mixture component. We assume that, for all speakers s and recordings h, observations drawn from mixture component c are distributed with mean $M_{hc}(s)$ and covariance matrix Σ_c .

The factor analysis model can be reduced to the eigenvoice MAP in the case where $\mathbf{d{=}0}$ and $\mathbf{u{=}0}$. The classical MAP is obtained if $\mathbf{u{=}0}$ and $\mathbf{v{=}0}$. And finally, if we assume that $\mathbf{M}(s)$ has a point distribution instead of the Gaussian distribution specified by the equation 2.1 and that this point distribution is different for different speakers we obtain the eigenchannel MAP.

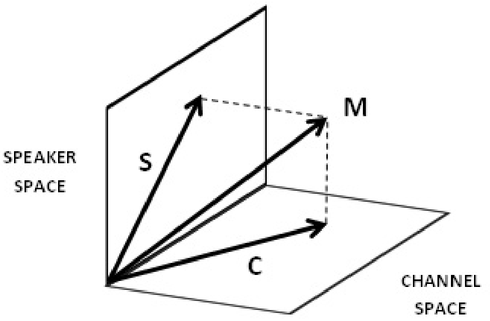

Figure 2.1: In the PCA case, a speaker- and channel-dependent supervector $\mathbf{M}$ can be written as a sum of two supervectors, one of which ($\mathbf{S}$) lies in the *speaker space* and the other ($\mathbf{C}$) lies in the *channel space* (in accordance with the parallelogram rule). In the general case, speaker supervectors are distributed in the neighborhood of the *speaker space.*

In order to ensure that the model inherits the asymptotic behavior of classical MAP the *special speaker factors* $\mathbf{z}(s)$ are included, but they are costly in terms of computational complexity. The reason for this is that, although the increase in the number of free parameters is relatively modest since (unlike $\mathbf{u}$ and $\mathbf{v}$) $\mathbf{d}$ is assumed to be diagonal, introducing $\mathbf{z}(s)$ greatly increases the number of hidden variables.

We will use the term Principal Components Analysis (PCA) to refer to the case where $\mathbf{d}$=$\mathbf{0}$. The model is quite simple in this case since the basic assumption is that each speaker- and channel-dependent supervector is a sum of two supervectors, one of which is contained in the *speaker space* and the other in the *channel space*. This decomposition is actually unique since the range of $\mathbf{uu}^*$ and the range of $\mathbf{vv}^*$, being low dimensional subspaces of a very high dimensional space, (typically) only intersect at the origin (see Fig. 2.1).

2.2 Speaker variability estimation

The supervector defined by a UBM can serve as an estimate of $\mathbf{m}$ and the UBM covariance matrices are good first approximations to the residual covariance matrices $\Sigma_c\,(c = 1, ..., C)$. The problem of estimating $\mathbf{v}$ in the case where $\mathbf{d}$=$\mathbf{0}$ was addressed in [10] and a very similar approach can be adopted for estimating $\mathbf{d}$ in the case where $\mathbf{v}$=$\mathbf{0}$. We first summarize the estimation procedures for these two special cases and then explain how they can be combined to tackle the general case, [8].

2.2.1 Baum-Welch statistics

Given a speaker s and acoustic feature vectors $Y_1, Y_2, ...$, for each mixture component c we define the Baum-Welch statistics in the usual way:

$$N_c(s) = \sum_t \gamma_t(c) \tag{2.5}$$

$$F_c(s) = \sum_t \gamma_t(c) Y_t \tag{2.6}$$

$$S_c(s) = diag(\sum_t \gamma_t(c) Y_t Y_t^*) \tag{2.7}$$

where, for each time t, $\gamma_t(c)$ is the posterior probability of the event that the feature vector Y_t is accounted for by the mixture component c. We calculate these posteriors using the UBM.

We denote the centralized first- and second order Baum-Welch statistics by $\tilde{F}_c(s)$ and $\tilde{S}_c(s)$:

$$\tilde{F}_c(s) = \sum_t \gamma_t(c)(Y_t - m_c) \tag{2.8}$$

$$\tilde{S}_c(s) = diag(\sum_t \gamma_t(c)(Y_t - m_c)(Y_t^* - m_c)) \tag{2.9}$$

where m_c is the subvector of $\mathbf{m}$ corresponding to the mixture component c. In other words,

$$\tilde{F}_c(s) = F_c(s) - N_c(s)m_c \tag{2.10}$$

$$\tilde{S}_c(s) = S_c - diag(F_c(s)m_c^* + m_c F_c(s)^* - N_c(s)m_c m_c^*) \tag{2.11}$$

Let $N(s)$ be the $CF\,x\,CF$ diagonal matrix whose diagonal blocks are $N_c(s)I$ $(c = 1, ..., C)$. Let $\tilde{F}(s)$ be the $CF\,x\,1$ supervector obtained by concatenating $\tilde{F}_c(s)$ $(c = 1, ..., C)$. Let $\tilde{S}(s)$ be the $CF\,x\,CF$ diagonal matrix whose diagonal blocks are $\tilde{S}_c(s)$ $(c = 1, ..., C)$.

2.2.2 Training an eigenvoice model

In this section we consider the problem of estimating $\mathbf{m}$, $\mathbf{v}$ and $\mathbf{\Sigma}$ under the assumption that $\mathbf{d{=}0}$. We assume that initial estimates of the hyperparameters are given. (Random initialization of $\mathbf{v}$ works fine in practice.)

2.2.2.1 The posterior distribution of the hidden variables

For each speaker s, set $l(s) = I + \mathbf{v}^*\mathbf{\Sigma}^{-1}N(s)\mathbf{v}$. Then the posterior distribution of $\mathbf{y}(s)$ conditioned on the acoustic observations of the speaker is Gaussian with mean $l^{-1}(s)\mathbf{v}^*\mathbf{\Sigma}^{-1}\tilde{F}(s)$ and covariance matrix $l^{-1}(s)$. (See [10], Proposition 1.)

We will use the notation $E[\cdot\,]$ to indicate posterior expectations; thus $E[\mathbf{y}(s)]$ denotes the posterior mean of $\mathbf{y}(s)$ and $E[\mathbf{y}(s)\mathbf{y}^*(s)]$ the posterior correlation matrix.

2.2.2.2 Maximum likelihood re-estimation

This entails accumulating the following statistics over the training set where the posterior expectations are calculated using initial estimates of $\mathbf{m}$, $\mathbf{d}$, $\mathbf{\Sigma}$ and s ranges over the training speakers:

$$N_c = \sum_S N_c \; (c = 1, ..., C) \tag{2.12}$$

$$\mathfrak{A}_c = \sum_S N_c(s)E[\mathbf{y}(s)\mathbf{y}^*(s)] \; (c = 1, ..., C) \tag{2.13}$$

$$\mathfrak{C} = \sum_S \tilde{F}(s)E[\mathbf{y}^*(s)] \tag{2.14}$$

$$N = \sum_s NS). \tag{2.15}$$

For each mixture component $c = 1, ..., C$ and for each $f = 1, ..., F$, set $i = (c-1)F + f$ let v_i denote the ith row of $\mathbf{v}$ and $\mathfrak{C}_i$ the ith row of $\mathfrak{C}$. Then $\mathbf{v}$ is updated by solving the equations

$$v_i \mathfrak{A}_c = \mathfrak{C}_i \ (i = 1, ..., CF) \tag{2.16}$$

The update formula for $\mathbf{\Sigma}$ is

$$\mathbf{\Sigma} = N^{-1}(\sum_S \tilde{S}(s) - diag(\mathfrak{C}\mathbf{v}^*)). \tag{2.17}$$

(See [10], Proposition 3.)

2.2.2.3 Minimum divergence re-estimation

Given initial estimates m_0 and v_0, the update formulas for $\mathbf{m}$ and $\mathbf{v}$ are

$$\mathbf{m} = m_0 + v_0 \mu_y \tag{2.18}$$

$$\mathbf{v} = v_0 T_{yy}^* \tag{2.19}$$

Here

$$\mu_y = \frac{1}{S} \sum_s E[\mathbf{y}(s)], \tag{2.20}$$

$$T_{yy}^* T_{yy} = \frac{1}{S} \sum_s E[\mathbf{y}(s)\mathbf{y}^*(s)] - \mu_y \mu_y^* \tag{2.21}$$

(i.e. Cholesky decomposition), S is the number of training speakers, and the sums extend over all speakers in the training set. (See [5], Theorem 7.) The role of this type of estimation is to get good estimates of the eigenvalues corresponding to the eigenvoices.

2.2.3 Training a diagonal model

An analogous development can be used to estimate $\mathbf{m}$, $\mathbf{d}$ and $\mathbf{\Sigma}$ if $\mathbf{v}$ is constrained to be $\mathbf{0}$.

2.2.3.1 The posterior distribution of the hidden variables

For each speaker s, set $l(s) = I + \mathbf{d}^2\mathbf{\Sigma}^{-1}N(s)$. Then the posterior distribution of $\mathbf{z}(s)$ conditioned on the acoustic observations of the speaker is Gaussian with mean $l^{-1}(s)\mathbf{d}\mathbf{\Sigma}^{-1}\tilde{F}(s)$ and covariance matrix $l^{-1}(s)$.

Again, we will use the notation $E[\cdot\,]$ to indicate posterior expectations; thus $E[\mathbf{z}(s)]$ denotes the posterior mean of $\mathbf{z}(s)$ and $E[\mathbf{z}(s)\mathbf{z}^*(s)]$ the posterior correlation matrix.

It is straightforward to verify that, in the special case where d is assumed to satisfy

$$\mathbf{d}^2 = \frac{1}{r}\mathbf{\Sigma}, \tag{2.22}$$

this posterior calculation leads to the standard relevance MAP estimation formulas for speaker supervectors (r is the relevance factor). The following two sections summarize data-driven procedures for estimating $\mathbf{m}$, $\mathbf{d}$ and $\mathbf{\Sigma}$ which do not depend on the relevance MAP assumption. It can be shown that when these update formulas are applied iteratively, the values of a likelihood function analogous to that given in Proposition 2 of [10] increase on successive iterations.

2.2.3.2 Maximum likelihood re-estimation

This entails accumulating the following statistics over the training set where the posterior expectations are calculated using initial estimates of $\mathbf{m}$, $\mathbf{d}$, $\mathbf{\Sigma}$ and s ranges over the training speakers:

$$N_c = \sum_S N_c \; (c = 1, ..., C) \tag{2.23}$$

$$\mathfrak{a} = \sum_S diag(N(s)E[\mathbf{z}(s)\mathbf{z}^*(s)]) \tag{2.24}$$

$$\mathfrak{b} = \sum_S diag(\tilde{F}(s)E[\mathbf{z}^*(s)]) \tag{2.25}$$

$$N = \sum_s N(S). \tag{2.26}$$

For $i = 1, ..., CF$ let and d_i the ith entry of $\mathbf{d}$ and similarly for $\mathfrak{a}_i$ and $\mathfrak{b}_i$ Then $\mathbf{d}$ is updated by solving the equation

$$d_i\mathfrak{a}_i = \mathfrak{b}_i \tag{2.27}$$

for each i. The update formula for $\mathbf{\Sigma}$ is

$$\mathbf{\Sigma} = N^{-1}(\sum_S \tilde{S}(s) - diag(\mathfrak{b}\mathbf{d})). \tag{2.28}$$

2.2.3.3 Minimum divergence re-estimation

Given initial estimates m_0 and d_0, the update formulas for $\mathbf{m}$ and $\mathbf{d}$ are

$$\mathbf{m} = m_0 + d_0 \mu_z \tag{2.29}$$

$$\mathbf{d} = d_0 T_{zz} \tag{2.30}$$

where

$$\mu_z = \frac{1}{S} \sum_s E[\mathbf{z}(s)], \tag{2.31}$$

T_{zz} is a diagonal matrix such that

$$T_{zz}^2 = diag(\frac{1}{S} \sum_s E[\mathbf{z}(s)\mathbf{z}^*(s)] - \mu_z \mu_z^*), \tag{2.32}$$

S is the number of training speakers, and the sums extend over all speakers in the training set.

We will need a variant of this update procedure which applies to the case where $\mathbf{m}$ is forced to be $\mathbf{0}$. In this case $\mathbf{d}$ is estimated from d_0 by taking T_{zz} to be such that

$$T_{zz}^2 = diag(\frac{1}{S} \sum_s E[\mathbf{z}(s)\mathbf{z}^*(s)]). \tag{2.33}$$

2.2.4 Joint estimation of v and d

There is no difficulty in principle in extending the maximum likelihood and minimum divergence training procedures to handle a general factor analysis model in which both $\mathbf{v}$ and $\mathbf{d}$ are non-zero (Theorems 4 and 7 in [5]).

In a general factor analysis model all of the hidden variables become correlated with each other in the posterior distributions, therefore joint estimation of $\mathbf{v}$ and $\mathbf{d}$ becomes computationally demanding. Given the Baum-Welch statistics, training a diagonal model runs very quickly and training a pure eigenvoice model can be made to run quickly (at the cost of some memory overhead) by suitably organizing the computation of the matrices $l(s)$ in Sec. 2.2.2.1. Unfortunately, in the general case, no such computational short cuts seem to be possible. Furthermore, many iterations of joint estimation are needed to estimate $\mathbf{d}$ properly even if the eigenvoice component $\mathbf{v}$ is carefully initialized, and, it is difficult to judge when the training algorithm has effectively converged because the contribution of $\mathbf{d}$ to the likelihood of the training data is minor compared with the contribution of $\mathbf{v}$.

2.2.5 Decoupled estimation of v and d

An alternative training regimen is presented to where the training speaker are divided into two disjoint sets. The larger set is used to estimate $\mathbf{m}$ and $\mathbf{v}$ and the smaller to estimate $\mathbf{d}$ and $\mathbf{\Sigma}$.

Specifically, a pure eigenvoice model to the larger training set is fit using the procedures described in Sec. 2.2.2.2 and 2.2.2.3. Then, for each speaker s in the residual training set, the MAP estimate of $\mathbf{y}(s)$ is calculated, namely $E[\mathbf{y}(s)]$, as in Sec. 2.2.2.1. This gives us a preliminary estimate of the speakers supervector s, namely

$$s = \mathbf{m} + \mathbf{v}E[\mathbf{y}(s)]. \tag{2.34}$$

The speakers Baum-Welch statistics is centralized by subtracting the speakers supervector (applying the formulas in Sec. 2.2.1 with $\mathbf{m}$ replaced by s). Finally, these centralized statistics are used together with the procedures described in Sec. 2.2.3.2 and 2.2.3.3 to estimate a pure diagonal model with $m = 0$. This gives us estimates of $\mathbf{d}$ and $\mathbf{\Sigma}$.

The training algorithm converges rapidly since it uses only the diagonal and eigenvoice estimation procedures.

2.3 Training the speaker and session variability subspaces

The speaker and session variability subspaces - described by the transformation matrices $\mathbf{v}$ and $\mathbf{u}$ - must be appropriately estimated in order to obtain an effective factor analysis model. These matrices should represent the types of inter- and intra-speaker variations expected within and between recording sessions. For this purpose, databases containing a large number of speaker each with several independently recorded sessions are needed to train the subspaces. This training database should include a variety of channels, handset types and environmental conditions that closely resembles the conditions on which the eventual system is to be used.

Estimates for the transformation matrices $\mathbf{v}$ and $\mathbf{u}$ can be obtained in different ways as it was explained in [14]. Four different options for how to obtain these subspace transformation matrices are examined in this section. The Alize's library, that implements an algorithm to train a JFA model, uses the disjoint estimation method.

2.3.1 Principal Component Analysis

Each utterance in the training dataset is first converted into a single observation by training a relevance MAP adapted GMM. From these observations, the within- and between-class scatter matrices are then calculated in order to capture the intra-speaker and inter-speaker variability, respectively. The principal components of these scatter matrices are determined through eigen decomposition, with the factors corresponding to the R_C and R_S largest eigenvalues retained and used to form the transform matrices $\mathbf{v}$ and $\mathbf{u}$ respectively.

Even if the PCA analysis is good starting point for further analysis, it has some shortcomings. Firstly, each utterance is reduced to a single point estimate through the relevance MAP adaptation process. This approach does not fully use all data available when calculating the transformation matrix. Secondly, the optimization criterion or training method used in speaker model training is not used by this approach and will therefore be suboptimal for this task.

2.3.2 Simultaneous estimation of v and u

The simultaneous approach use an EM algorithm with the *speaker* and *session factors* $\mathbf{y}(s)$ and $\mathbf{x}_h(s)$ as hidden variables to refine $\mathbf{u}$ and $\mathbf{v}$. A maximum likelihood criterion over the entire dataset is optimized with each speaker s optimized as per the speaker model training described above. This method is described in [9] with the transformation matrix optimization equations presented in [10].

This approach addresses the issues highlighted for the PCA approach, specifically using all data available in the matrix optimization as well as optimizing the same criterion as the speaker model enrollment. Compared to PCA, the simultaneous approach therefore provides a more refined and theoretically optimal solution for training both $\mathbf{u}$ and $\mathbf{v}$.

As the simultaneous method employs ML as the optimization criterion it will fit the training data as well as it can. Considering the purpose of having separate subspaces, this result may actually not be desirable. Specifically, these subspaces have been termed *speaker* and *channel subspaces* but there is no means within an ML framework to constrain the $\mathbf{u}$ to capture only session variability and not capture speaker variability. If, for instance, $\mathbf{v}$ is not of high enough rank, there will be significant speaker variability captured by $\mathbf{u}$.

As in speaker model training the value and information contained in $\mathbf{x}_h(s)$ is effectively discarded, any speaker information captured in $\mathbf{u}$ will also be discarded.

It should be noted that the simultaneous optimization is performed under the assumption that $\mathbf{d}$=$\mathbf{0}$, to ensure that as much of the observed variability is modeled by the low-rank speaker and session spaces.

2.3.3 Disjoint estimation of v and u

Matrices $\mathbf{u}$ and $\mathbf{v}$ can be optimized independently in an attempt to explicitly capture the variability they are intended to model.

The optimization equations presented in [10] are again used to train $\mathbf{u}$ but with $\mu(s)$ estimated by a very loosely constrained relevance MAP, that is, by setting it to be very small. The reason of using a small value is that the relevance MAP adaptation will be preferred to model any common speaker characteristics found across sessions for a given speaker s in the training dataset and that $\mathbf{u}$ will be preferred only to capture the differences between sessions of the same speaker, that is, the inter-session variability.

To train $\mathbf{v}$ we use the model without $\mathbf{u}$ and with no relevance MAP ($\mathbf{d}=\mathbf{0}$). This approach forces $\mathbf{v}$ to represent as much of the variability in the training dataset as possible.

As it is not directly optimizing the ML criterion, the disjoint optimization approach will generally produce a lower overall likelihood than the previous approach , however, $\mathbf{u}$ is more likely to fulfill its role in modeling only the session variability.

2.3.4 Coupled estimation of v and u

Similar to the disjoint approach, the coupled estimation has the exception that an attempt is made to explicitly remove session variation during the optimization of $\mathbf{v}$ by incorporating a pre-trained session variability component ($\mathbf{u}$). Under the coupled approach, variability likely to be caused by session conditions, as described by $\mathbf{u}$, is modeled explicitly.

The same procedure as in the disjoint estimation is used to train $\mathbf{u}$. Once optimization of $\mathbf{u}$ is complete, the procedure followed in the simultaneous approach is used to optimize $\mathbf{v}$ including $\mathbf{u}$ into the FA model for each speaker. However, to perform this optimization, $\mathbf{u}$ is held constant rather than re-estimated. The optimization of $\mathbf{v}$ is once again performed under the assumption of no relevance MAP component ($\mathbf{d}=\mathbf{0}$).

Chapter 3

Process Description

In this chapter, the whole process to adapt the speaker model is described from the feature extractor stage to the result stage. All the steps are explained and some alternatives are shown. In this project, we have focused in two systems to see the factor analysis performance, one is a speaker verification system and the other is a forensic automatic speaker recognition system.

3.1 Database

Data from the Polyphone IPSC-03 database was used to develop an experimental framework. This database was chosen for two main reasons. First, the datasets cover a wide range of acoustic (fix telephone, cell phone and microphone) allowing for vigorous testing under mismatched conditions. Secondly, this database contains three different kinds of recordings (reference database, controlled database and trace database). The database was recorded by Philipp Zimmerman, Damien Dessimoz and Filippo Botti from the Scientific Police Institute of Université de Lausanne (UNIL), and Anil Alexander and Andrzrej Drygajlo from the Signal Processing Intitute of École Polytechnique Fédérale de Lausanne (EPFL) [1].

This database has 62 useful speakers, all males. Each speaker has recordings in the three different conditions above. The recordings have three different modes; normal mode (reading a printed text), spontaneous mode (involving two simulated situations) and the dialog mode (reading a text in the tone of conversation).

To realize the experiments we have divided the database in three subsets of speakers. The first subset contains 35 speakers to train a UBM. The second is formed with 20 speakers to play the role of imposters or the population database. And the last one involves 7 speakers to play the role of suspects.

As said before, the database is divided in three parts (population, controlled and reference database). In the forensic framework, each subset has a special purpose which will be described below.

The population database (P) is used to model the variability of the speech of all the potentially relevant sources using the automatic speaker recognition method. The calculated between-sources variability pdf is then used to estimate the denominator of the likelihood ratio $p(E|H_1)$. Ideally, the technical characteristics of the recordings (e.g., signal acquisition and transmission) should be selected according to the characteristics analyzed in the trace.

The reference database (R) is recorded with the suspected speaker to model his/her speech with the automatic speaker recognition method. Speech utterances should be produced in the same way as those of the P database. The suspected speaker model obtained is then used to calculate the value of the evidence (E) by comparing the questioned recording with the model.

The controlled database (C) is recorded with the suspected speaker to evaluate her/his within-source variability when the utterances of this database are compared to the suspected speaker model. The calculated within-source variability pdf is then used to estimate the numerator of the likelihood ratio $p(E|H_0)$. The recording of the C database should be constituted of utterances as close as possible to the trace, according to the technical characteristics, quantity, and style of the speech.

The Polyphone IPSC-03 database is ideally organized to simulate a real forensic case. A complete description of the database can be found in appendix A.

3.2 Feature extraction

Mel-Frequency Cepstral Coefficients (MFCC) is one of the most popular feature extraction methods used in speaker verification and identification systems. It is based on the properties of human auditory system on the perception of frequencies. Thus, it analyses more in detail the lower frequencies and more roughly the highest frequencies. Here is the configuration parameters used in the feature extractor:

- Frame size of 25 ms.
- Frame shift of 10 ms.
- Hamming window.
- Number of filters 24.
- Number of cepstral coefficient 12
- Use of the log-energy.

- Use of the first and second derivative (delta and acceleration).
- Use of CMS (Cepstral Mean Subtraction).
- Use of energy normalization.
- The triangular filter bank has a low frequency of 100 Hz and a high frequency of 7,2 kHz for the recordings sampled at 16 kHz (GSM and PSTN) and 5 kHz for those sampled at 11,025 kHz (room microphones).

3.3 Universal Background Model

In order to estimate the speaker- and channel-independent supervector needed to create the joint factor analysis model, an Universal Background Model (UBM) was trained.

To train the UBM, all the speakers were used except those that will play the role of suspects and those that will play the role of imposters. Thus, a total of 35 speakers were used for the train. This UBM has a diagonal Gaussian distribution of 64 mixtures.

The most important part was to decide if train an UBM for each condition or train a global UBM regarding all possible conditions. Finally we decided to choose the second option. Using all the condition, the UBM will be channel-independent and will not produce a source of mismatch in the speaker model.

Forensic laboratories have databases each time larger and larger, and new channel conditions are now available. Thus, the idea of training a UBM, which is speaker-independent, can now have a sense of channel-independent.

We used the reference database of each speaker to create the UBM. So finally we have a UBM created with 35 speakers and 3 recordings per speaker, doing a total of 105 recordings.

3.4 Speaker and Session Models

To follow the methodology of a speaker verification system and a forensic automatic speaker recognition system, two subsets of speaker have been used.

The first subset was used to estimate the distribution curve of the H_0 hypothesis, where the suspected speaker is the source of the questioned recording. It contains a total of 7 speakers to play the role of suspects.

The second subset was used for the same purpose as the first one, but with this database we estimate the H_1 hypothesis, where the speaker at the origin of the questioned recording is not the suspected speaker. This subset has a total of 20 speakers to simulate the imposters or the population database.

For each speaker, we have train a set of 5 models for comparison between them. These models are known as Speaker model, True Speaker model, Session model, Substitution model and Joined model. The first three models will be described in this section and we will leave the Substitution and Joined model (Sec. 3.6) after describe the adaptation methods (Sec. 3.5).

To train the speaker model, we have used the reference database as in the case of the UBM. We have 3 recordings per speaker to create each model.

3.4.1 True Speaker Model

This model regards only the speaker subspace. It creates a GMM using the next expression:

$$\mathbf{M}(s) = \mathbf{m} + \mathbf{v}\mathbf{y}(s), \tag{3.1}$$

where $\mathbf{m}$ is the session-speaker dependent supervector mean of dimension CF, $\mathbf{v}$ is $CF\,x\,CF$ diagonal matrix and $\mathbf{y}(s)$ is the speaker vector ($\mathbf{y}(s)$ is normally distributed among $N(0|I)$). Matrix $\mathbf{v}$ satisfies the following equation:

$$I = \tau\mathbf{v}^{*}\mathbf{\Sigma}^{-1}\mathbf{v}, \tag{3.2}$$

where τ is the relevance factor required in the standard MAP adaptation.

3.4.2 Session Model

This is a model that takes only into account the channels subspace. The original Alize's source code creates a model using

$$\mathbf{M}_h(s) = \mathbf{m} + \mathbf{u}\mathbf{x}_h(s). \tag{3.3}$$

This source code was modified to obtain a model as

$$\mathbf{M}_h(s) = \mathbf{u}\mathbf{x}_h(s), \tag{3.4}$$

where $\mathbf{u}$ is the eigenchannel matrix of low rank R_c (a $CF\,x\,R_c$ matrix) and $\mathbf{x}_h(s)$ are the session factors ($\mathbf{x}_h(s)$ is normally distributed among $N(0|I)$ and theoretically does not dependent on s). The number of session factors was fixed to 40 (in [3] was shown that this rank obtain the best results).

Apart from the session model of each speaker, we have trained one global model per condition, using all the possible recordings, to create a more complete model of each condition.

3.4.3 Speaker Model

The speaker model is a complete model taking into account the speaker and channel subspaces. It creates a GMM as

$$\mathbf{M}_h(s) = \mathbf{m} + \mathbf{v}\mathbf{y}(s) + \mathbf{u}\mathbf{x}_h(s), \tag{3.5}$$

so finally we obtain the expression of the JFA model.

The parameters that compose this model were previously described in Sec. 3.4.1 and 3.4.2.

3.5 Adaptation

Due to the ability of joint factor analysis to decompose a speaker model into two well defined subspaces (speaker and channel subspace), we can adapt mismatched conditions in a simple way. There are different methods of adaptation that have been proven to have a good performance (as we can see in [3] and [4]). We will explain the most important strategies.

A training set in which there are multiple recordings of each speaker is needed in order to use the speaker-independent hyperparameter estimation algorithms, it seems very unlikely that speaker and session effects can ever be broken out using a training corpus in which there is just one recording for each speaker.

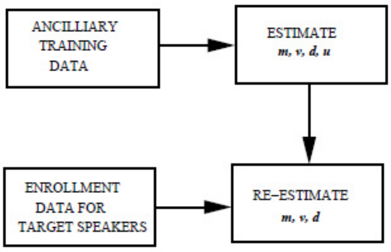

Figure 3.1: We estimate the speaker-independent hyperparameters on a larger ancillary training corpus that contains multiple recordings for each speaker. This is followed by adapting the hyperparameters that model inter-speaker variability (namely **m**, **v** and **d**) to the target speaker population; we assume that channel effects are invariant so we keep **u** fixed. [7]

To deal with this problem, first, we estimate a full set of hyperparameters **m**, **u**, **v**, **d** and $\mathbf{\Sigma}$ on the ancillary training corpus and then, holding **u** and $\mathbf{\Sigma}$ fixed, re-estimate **m**, **v** and **d** on the enrollment data (but not the test data) for the target speakers (Fig. 3.1). In other words, the hyperparameters associated with channel space are kept fixed and only the hyperparameters associated with the speaker space are re-estimated. It is necessary to keep the orientation of the speaker space fixed. This procedure is the base of the strategy called traditional approach.

There are also three techniques for combining information from a data-rich domain and limited target domain data [4]. The main idea is to deal with this limited data problem by exploiting data from a data-rich domain in the session subspace estimation procedure in order to achieve a dual goal. The first purpose is to obtain a more robust estimation procedure by adding large amounts of data. The second objective is to incorporate certain session variability characteristics not present in the limited available target domain but that could appear in the test domain. These techniques are called joining matrices, pooled statistics and scaling statistics.

Finally, from all the methods described below, we decided to implement the traditional approach technique and the joining matrices method.

3.5.1 Traditional Approach

In the traditional approach method [3], suppose we have a segment of speech Y to test and a targeted speaker s with a model learned from speech T. By using the session factor decomposition, we obtain

$$\mathbf{m}_{(h_Y,s_Y)} = \mathbf{m} + \mathbf{v}\mathbf{y}_{s_Y} + \mathbf{u}\mathbf{x}_{h_Y} \tag{3.6}$$

and

$$\mathbf{m}_{(h_T,s_T)} = \mathbf{m} + \mathbf{v}\mathbf{y}_{s_T} + \mathbf{u}\mathbf{x}_{h_T} \tag{3.7}$$

To compensate the session component in the score computation the next strategy is used. The test speaker and the target speaker are assumed to have the same identity. In this case, $\mathbf{y}_{s_Y}$ (speaker component in the test) is not estimated, but it is assumed to be equal to the speaker component in the target speaker $\mathbf{y}_{s_T}$. The channel component ($\mathbf{u}\mathbf{x}_{h_Y}$) is estimated in the test Y. To compensate the channel mismatch, the channel component in the target mean supervector ($\mathbf{u}\mathbf{x}_{h_T}$) is replaced by the one estimated in the test ($\mathbf{u}\mathbf{x}_{h_Y}$). The vector $\mathbf{m}$ from the UBM in the score equation remains unchanged. This strategy was adopted in [10] and [15]. In practice, a compensation is needed in the world model to avoid session mismatch in the same way as the target model. If world model compensation is not applied, a negative difference between the likelihoods of the test data can be observed: the likelihood estimated on the target model can be larger than the one estimated on the world model. For that reason, the UBM is trained with all the possible conditions to avoid the mismatch.

3.5.2 Alternative Approach

Taking into account the expressions in 3.6 and 3.7, another approach is proposed.

The alternative approach [3] is a strategy in which all the sessions are considered and treated separately. For each session, we estimate independently the session mismatch and the speaker of all other sessions. So, the channel mismatch can simply be eliminated from each session. However, the session mismatch is estimated in the model space, and the session compensation (for the test) must be performed in the feature space.

To do the latter, we adopt a strategy used in [13], namely

$$\hat{\mathbf{t}} = \mathbf{t} - \sum_{g=1}^{M} p(g|\mathbf{t})\mathbf{u}_g \mathbf{x}_{h_Y} \tag{3.8}$$

where $\mathbf{t}$ is a frame of size F, $p(g|\mathbf{t})$ is the Gaussian occupation probability of the component g, and $\mathbf{u}_g$ is a subset of $\mathbf{u}$ corresponding to g. Hence, two options are available in order to compensate the session mismatch.

1. *Feature Space Compensation.* All the compensations are performed in the feature space. This option is interesting because it operates in the feature space and is independent of the classifier.

2. *Symmetrical Compensation.* The target models are compensated by eliminating the session mismatch directly in the model and the compensation in the test is performed in the feature space. This new approach is called symmetrical factor analysis (SFA).

3.5.3 Joining Matrices

A simple way to combine different session variability subspaces is joining session variability subspaces that have been 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.

The major advantage of this approach is that subspaces can be treated and trained independently. From a practical point of view, this property is highly desirable because we can refine a well-trained reference subspace by simply appending new session variability information from other domains.

On the other hand, it has some deficiencies. The first one refers to the principle of keep the overall size of the joined subspace relatively small, thus, it is necessary to restrict the size of each contributing subspace, loosing potentially useful directions of variability. The second one concerns the importance of each subspace, no particular emphasis is placed on the target domain data because all the directions play an equal role in the new subspace.

3.5.4 Pooled Statistics

This time, all data is pooled to perform the estimation. An obvious advantage of this method is that the estimation is performed using a substantial amount of data, making it potentially more robust. Unfortunately, we can not prevent the supplementary set to dominate the estimation and to have the biggest effect on the variability directions.

3.5.5 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 idea of this approach 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 the most important data is highlighted. 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, a previously fixed weight depending on the dataset is used to scale the first order statistics supervector extracted from each utterance. Thus, the matrix of first order statistics $\mathbf{S}$, input in the EM procedure for training the variability subspace, takes the following form:

$$\mathbf{S} = \alpha \mathbf{S}_{tgt} + (1 - \alpha)\mathbf{S}_{bckg} \tag{3.9}$$

where $\mathbf{S}_{bkg}$ and $\mathbf{S}_{tgt}$ are the matrices whose columns are the first order statistics of utterances belonging to target data and other background data available respectively.

More generally, this could be extended to:

$$\mathbf{S} = \alpha_1 \mathbf{S}_1 + \alpha_2 \mathbf{S}_2 + ... + \alpha_n \mathbf{S}_n \tag{3.10}$$

with $\sum_{i=0}^{n} a_i$ and n different background sets.

In this way, the weight of each subset in the EM procedure can be balanced such that the available data can be combined in an optimal way. The problem of finding the optimal selection of weights is the main disadvantage, and is unique for each case. Although this can be solved empirically, choosing the weights in a proportional way to the quantity of data in target domain is a reasonable option, keeping at least a minimum weight for the rest of the sets.

3.6 Adapted Models

In this section, we will explain how to implement the theory seen previously (3.5.1 and 3.5.3). At the beginning, the alternative approach was supposed to be implemented also, but finally we discarded that option because it works in the feature space instead of the score domain. Another reason was the difficulty to implement it using the Alize's source code.

The idea is to move the model trained under certain conditions along the *channel subspace* to fit the conditions of the test recordings. In other words, we want to go from Fig.3.2 to Fig.3.3.

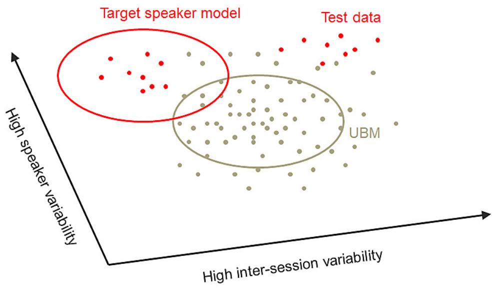

Figure 3.2: Initial trained model position

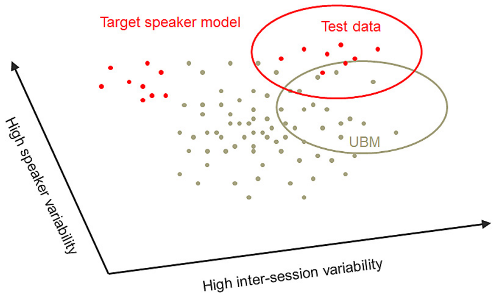

Figure 3.3: Final adapted model position

3.6.1 Substitution Model

This model tries to adapt the session conditions in a simple way. The principal idea is to obtain the channel conditions of the test recording (we will name it as C_2) and use this information to adapt the trained model (which is under conditions C_1).

The adaptation procedure was done as follows:

1. Obtainment of the true speaker model of a train recording under conditions C_1.

$$\mathbf{M}(s) = \mathbf{m} + \mathbf{vy}(s) \tag{3.11}$$

2. Obtainment of the session model of a test recording under conditions C_2.

$$\mathbf{M}'(s) = (\mathbf{ux}_h)'(s) \tag{3.12}$$

3. Addition of the session model to the true speaker model.

$$\mathbf{M}''(s) = \mathbf{m} + \mathbf{v}\mathbf{y}(s) + (\mathbf{u}\mathbf{x}_h)'(s) \tag{3.13}$$

4. Generation of the new GMM that we will call as Substitution model.

3.6.2 Joined Model

This approach is similar to the previous one, but instead than change the whole session model, we complement it with the new session conditions. We will refer, as in the previous strategy, the conditions in the training recording as C_1 and the conditions in the test recording as C_2.

The procedure is as follows:

1. Obtainment of the speaker model of a train recording in conditions C_1.

$$\mathbf{M}(s) = \mathbf{m} + \mathbf{v}\mathbf{y}(s) + \mathbf{u}\mathbf{x}_h(s) \tag{3.14}$$

2. Obtainment of the session model of a test recording in conditions C_2.

$$\mathbf{M}'(s) = (\mathbf{u}\mathbf{x}_h)'(s) \tag{3.15}$$

3. Creation of a new session model combining $\mathbf{u}\mathbf{x}_h(s)$ and $(\mathbf{u}\mathbf{x}_h)'(s)$. The result will be a new matrix with eigenchannels from C_1 and C_2, $(\mathbf{u}\mathbf{x}_h)''(s)$. The contribution of each condition in the new matrix can be chosen.

4. Removal of the session subspace from the original model $\mathbf{M}(s)$ and addition of this new session model to the first speaker model to obtain the Joined model.

$$\mathbf{M}''(s) = \mathbf{m} + \mathbf{v}\mathbf{y}(s) + (\mathbf{u}\mathbf{x}_h)''(s) \tag{3.16}$$

5. Generation of the new GMM that we will call as Joined model.

This adaptation does not allowed us to achieve good results in certain situations, so we decided to discard it when presenting the results and we will just focus in the substitution procedure.

3.7 Speaker Verification

In the speaker verification system, we compute the log-likelihood ratio between the speaker model and the UBM. For this purpose, we have created two datasets of test recordings, one to obtain scores for the H_0 hypothesis and the other for the H_1 hypothesis.

The H_0 database, contains recordings where the speaker is the same as the suspect. The trace database of the suspect was used as the H_0 database, it contains 11 recordings. Doing that, we assure that the suspected speaker is the source of the questioned recording.

For the H_1 database, we have used the trace database of each imposter. In that case we know that the speaker at the origin of the questioned recording is not the suspected speaker.

Once we have the 2 databases prepared, we compute the log-likelihood ratio between the suspect model and the UBM. With these scores, we can estimate the probability density function of each hypothesis.

We have used 0 as threshold so, for this system, the H_0 curve must be above 0 and the H_1 curve must be below 0. The performance of the system will be described in Sec. 4.1.

3.8 Forensic Automatic Speaker Recognition

The FASR system follows the methodology described in [2], where 3 databases are needed (trace, reference and control database). Fig. 3.4 shows all the steps and operations to obtain the log-likelihood ratio.

The first step is to obtain the evidence (E), for this purpose we obtain the log-likelihood between the trace and the suspect model. Then, to estimate the H_0 probability density function, we compute the scores between the control database and the suspect model. The last step is to estimate the H_1 probability density function, for that reason we compute the scores between the trace and the population models.

Once we have the H_0 and H_1 distribution and the evidence, we proceed to calculate the log-likelihood ratio. For this purpose we use the expression

$$LR = \frac{p(E|H_0)}{p(E|H_1)} \tag{3.17}$$

where LR is the strength of the evidence, and is the ratio that the expert will present to the court. The performance of the system can be found in Sec. 4.2.

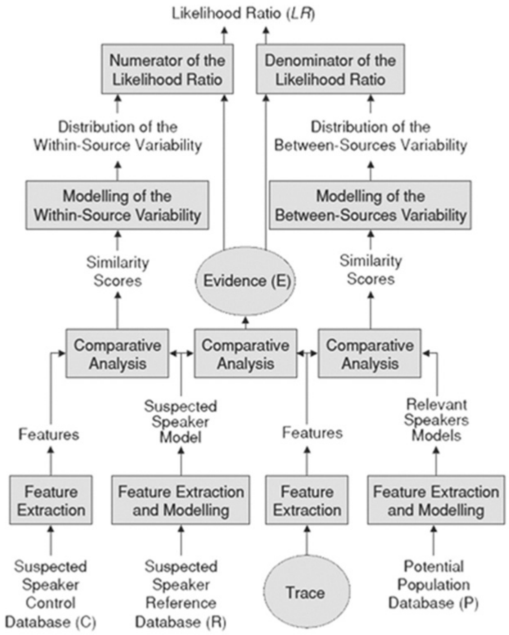
Likelihood Ratio (*LR*)
Numerator of the Likelihood Ratio
Denominator of the Likelihood Ratio
Distribution of the Within-Source Variability
Distribution of the Between-Sources Variability
Modelling of the Within-Source Variability
Modelling of the Between-Sources Variability
Similarity Scores
Evidence (E)
Similarity Scores
Comparative Analysis
Comparative Analysis
Comparative Analysis
Features
Suspected Speaker Model
Features
Relevant Speakers Models
Feature Extraction
Feature Extraction and Modelling
Feature Extraction
Feature Extraction and Modelling
Suspected Speaker Control Database (C)
Suspected Speaker Reference Database (R)
Trace
Potential Population Database (P)

Figure 3.4: Block diagram of the evidence processing and interpretation system. [2]

Chapter 4

Evaluation

We now present a set of results related to the different systems and situations developed for the JFA framework. First of all, we developed two well differentiated systems, the first one is a speaker verification system and the second one is a simulation of a real forensic casework. Each system was tested with three different situations; these situations are matched conditions, mismatched conditions with the non-adapted models and mismatched conditions using the adapted models. These conditions are GSM, PSTN and room microphones (Room acoustic).

To present the results, we used some plots to draw the scores. One of those plots are the Tippett plots that represents the proportion of the likelihood ratios greater than a given LR, i.e., $P(LR(H_i) > LR)$, for cases corresponding to the hypotheses H_0 (the suspected speaker is the source of the questioned recording) and H_1 (the speaker at the origin of the questioned recording is not the suspected speaker) true. The separation between the two curves in this representation is an indication of the performance of the system or method.

The likelihood ratio value of 1 is important for the forensic case, as it is the threshold between the support for the hypotheses H_0 and H_1. This is the reason why we look at the values of the Tippett curves at this point. The value of the curve at the point 0 (which is equal to $\log_{10}(1)$) give us the proportion of the distribution that is above 1. In principle, the results for an ideal system should be: the 100% of the H_0 distribution above 0 (this means that value in 0 is 1) and the 100% of the H_1 curve below 0 (the value of the point 0 is 0).

The first system developed was the speaker verification system; we used it to check is our adaptation method works correctly in a simple framework. If the results were not good we would not have proposed to use them in a forensic case.

The next step, and the original project idea, was to create a forensic casework to study the utility of the adaptation method in a strict and rigorous field. The process seems to improve the results but there is still work to do to obtain a perfect adaptation.

The performance of the system varies on many factors, as the number of speaker playing the role of imposter or population database, the number of recordings used to train the session model or the number of speakers to train the UBM. During this project, several configurations have been used in order to obtain the best performance of the system. Some of these values can be found at Sec. 3.

4.1 Speaker Verification

As described in Sec. 3.7, this system compares the log-likelihood ratio between the suspect and the UBM and makes a decision (the recording belongs to the suspect or not) comparing the score with the threshold. Thus, if the log-likelihood is above 0 we can say that the speaker in the recording is the same as the suspect and if it is below 0, they are two different speakers.

Translating the above to our case, that means that the H_0 distribution curve must be above 0 and the H_1 distribution curve below 0.

We will see in the next sections how the effect of mismatched conditions can influence in the results, and how the adapted model can solve the problem.

The model used to obtain the scores was the speaker model explained in Sec. 3.4.3, we used that model because it contains information related to the speaker and the channel subspace. In the case of adapted conditions we used the substitution model described in Sec. 3.6.1, that is the session adapted model.

We have made four different sets of experiments involving the three different conditions of the database. In each case, a plot with the three possible situations is shown (matched, mismatched and adapted conditions).

The situations of matched conditions are drawn in blue; this means that the speaker model was trained under the same condition as the test recordings. In color red we can find the mismatched situation, it means that the speaker model was trained under different condition of the test recordings. Finally, in black, we can see the results obtained by adapting the speaker model to the conditions of the test recordings.

Analyzing the results, we can conclude that the pair of conditions where the system works best is the one where we adapt from GSM to Room acoustic conditions, all the H_0 distribution is above 0 and all the H_1 distribution is below 0, so they are quite separated and there is no overlap between the two curves. Also, we can say that the worst system performance appears when we try to adapt from PSTN to GSM conditions, the H_0 mean is near 0 and the distribution has a high variance (a large proportion of the results are below 0).

The mentioned graphics are shown below.

4.1.1 Speaker in GSM and test in PSTN

In this case, we have trained two speaker models, one in PSTN conditions to perform a test in matched conditions and one in GSM conditions, the last one serves to make a test in mismatched conditions and in adapted conditions. Using the PSTN session model, we have adapted the speaker to the test conditions. The results are presented in the following plots.

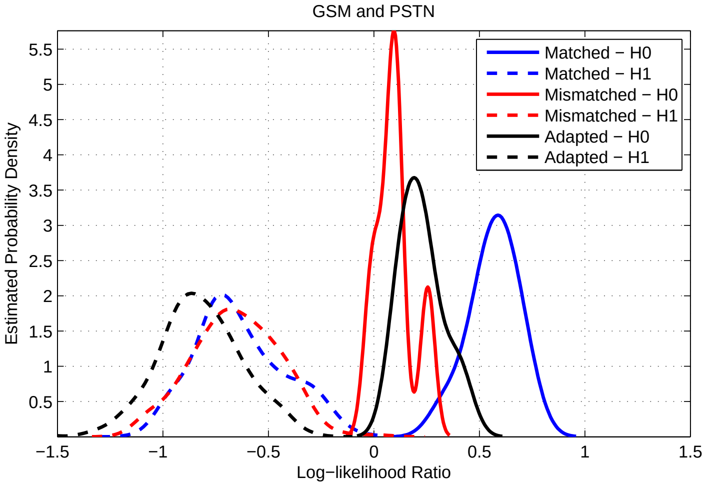

Figure 4.1: Matched, mismatched and adapted conditions (PDF Plot)

In this case, even the results in mismatched conditions are quite good. But our interest lies in the adaptation process, we can see that the H_0 distribution curve is almost all above 0 and the two distribution are quite far separated.

4.1.2 Speaker in PSTN and test in GSM

This situation is the opposite of the experiment presented in Sec. 4.1.1, the speaker model in GSM conditions serves us to perform the matched test and the speaker model in PSTN is used to the mismatched and adapted test. The following plots illustrate the performance of this system.

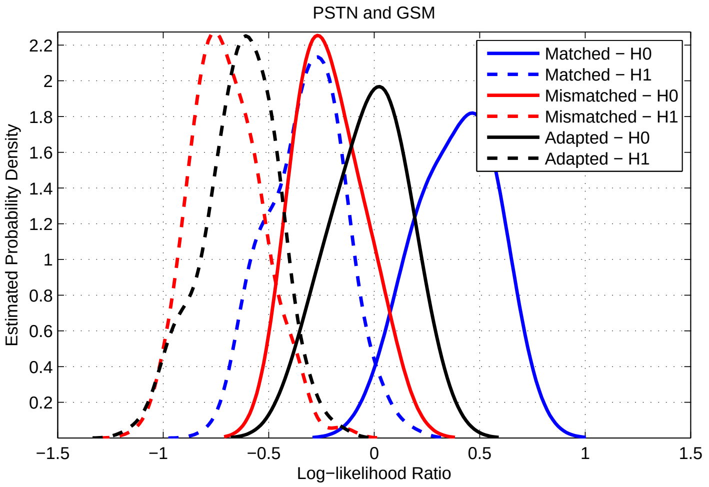

Figure 4.2: Matched, mismatched and adapted conditions (PDF Plot)

This is the worst case of all experiments done, the adapted H_0 is centered in 0 and the variance is quite high. Moreover, the two distributions are very close and the overlap is quite significant.

4.1.3 Speaker in GSM and test in Room acoustic

In this experiment, we trained a speaker model in room acoustic conditions and we used the speaker model in GSM conditions that we had trained for the previous tests. Once we had the two models, we compute the test using the speaker model in room acoustic conditions to evaluate the matched case and the speaker model in GSM conditions for the mismatched and adapted test.

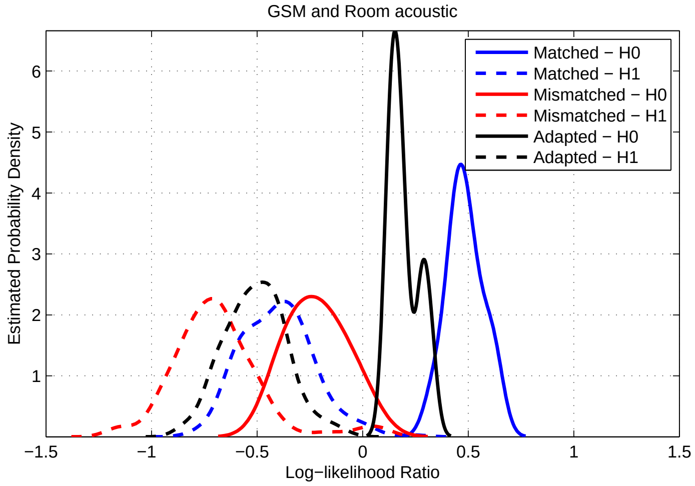

Figure 4.3: Matched, mismatched and adapted conditions (PDF Plot)

This experiment has achieved the best performance, in the case of mismatched conditions without adaptation, the two distributions are below 0, so the system will have a high error rate. Nevertheless, all the adapted H_0 distribution is above 0 and the overlap is negligible.

4.1.4 Speaker in Room acoustic and test in GSM

As in Sec. 4.1.2, this case is the opposite of the experiment in Sec. 4.1.3. The speaker models used are the same but the test conditions are GSM. The speaker model in GSM conditions is used to compute the matched test and the model in room acoustic conditions for the mismatched and adapted test.

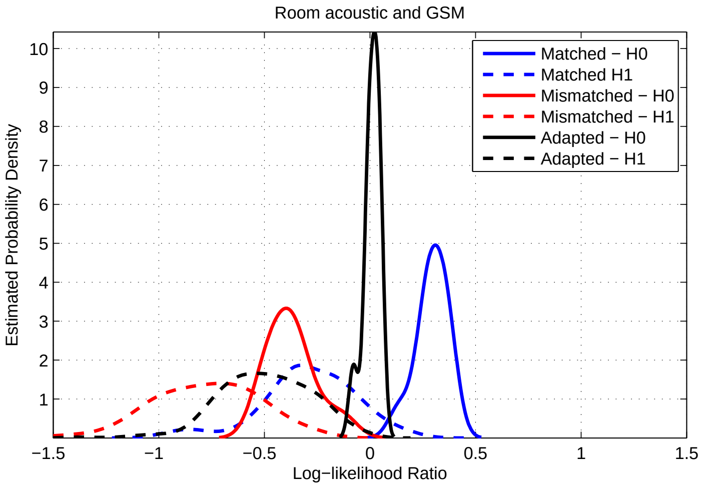

Figure 4.4: Matched, mismatched and adapted conditions (PDF Plot)

As in the case studied in Sec. 4.1.2, the adapted H_0 curve is centered in 0, but in this case, the overlap is not as high as in the referred case. Even so, the error rate can be high due to some scores are below 0, resulting in detection errors.

4.2 Forensic Automatic Speaker Recognition

The FASR used in a forensic framework was described in Sec. 3.8, the purpose of this system is to obtain the likelihood ratio that serves as the strength of the evidence. Three databases are needed for this methodology in order to obtain the evidence, the H_0 distribution and the H_1 distribution.

In a forensic casework, it is difficult to have a complete population database that matches with all the possible conditions or, at least, with the conditions under which the trace was recorded.

In this section, we simulate four forensic cases involving different conditions. Thus, as in the verification system (Sec. 4.1), we try to see the performance of adapting the joint factor analysis model under the case of mismatched conditions.

Each case was developed as follows: firstly, we have computed the results for the case of matched conditions, where the population database was recorded under the same conditions of the trace. Secondly, we have obtained the results of the mismatched conditions situation; there are a mismatch of conditions between the P database and the T database. And finally, we have adapted the P database conditions to match with the T database.

The following plots show the performance of the system developed. In blue we can see the H_0 distribution curve (it is the distribution of the scores between the suspect model and the control database). The H_1 distribution (that is the distribution of the scores between the population database and the trace) is represented in three colors: red for the matched conditions, magenta for the mismatched conditions and black for the adapted conditions case.

Theoretically, if the system works perfectly, the black curve should be the same, or at least very similar, to the red curve. We can see that our method can adapt quite well the mean but we have a problem with the variance.

As we done previously for the speaker verification system, we can conclude that the best adaptation is achieved when we adapt from GSM to PSTN conditions. In this case, the method can adapt quite good the H_1 mean but the shape of the distribution is not exactly the same. The worst situation is when we adapt from GSM to room acoustic conditions. We can see that the variance problem is also present but this time we must add the problem of mean mismatch.

The referred plots are presented below.

4.2.1 Population database in GSM and trace in PSTN

In this simulated case, we have trained two sets of population models, one in PSTN conditions to perform the test in matched conditions and the other in GSM conditions. As in the verification system, we have used each set of models to recreate a situation where there is a mismatch between the databases. The models under PSTN conditions serve as to simulate a case where the trace and the P database are in the same conditions and we used the other set of models to see the performance under mismatched and adapted conditions. The results are presented in the following plots.

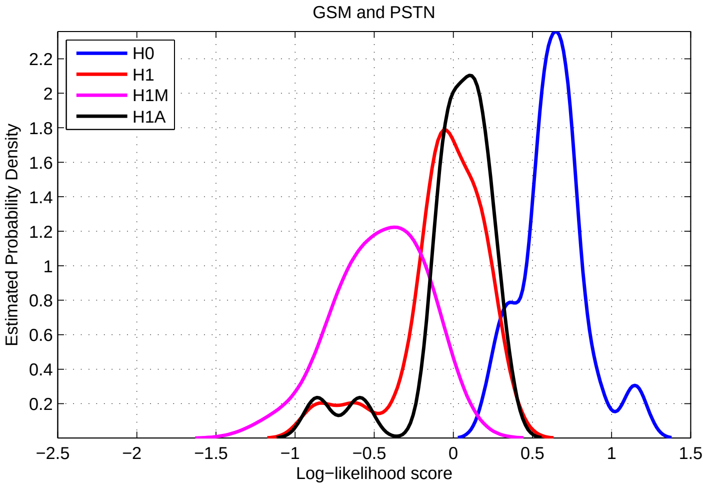

Figure 4.5: Matched, mismatched and adapted conditions (PDF Plot)

In this case, we can see how the mismatched H_1 distribution is shift to the left taking a different position than the matched case. With the adaptation technique, we fixed this displacement of the curve but we have problems with the distribution height. This adaptation is one of the most successful we have achieved, the adapted curve fits pretty well to the matched one. In Tab. 4.1 we can observe the percentage of the distribution which is above 0 for the H_0 hypothesis and the percentage below 0 for the H_1 hypothesis.

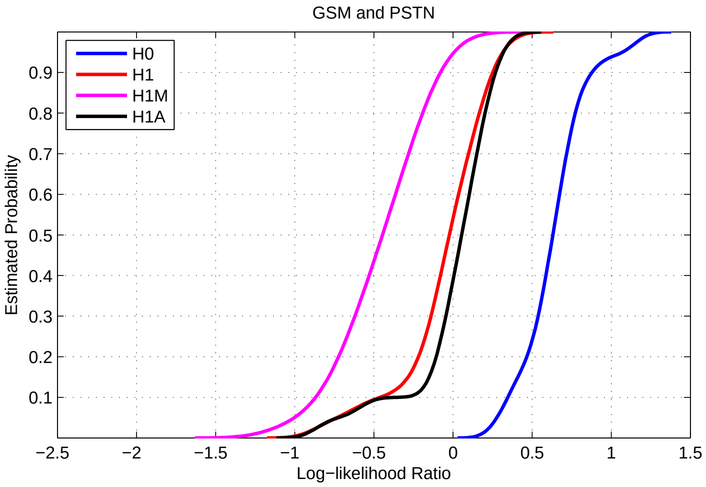

Figure 4.6: Matched, mismatched and adapted conditions (CDF Plot)

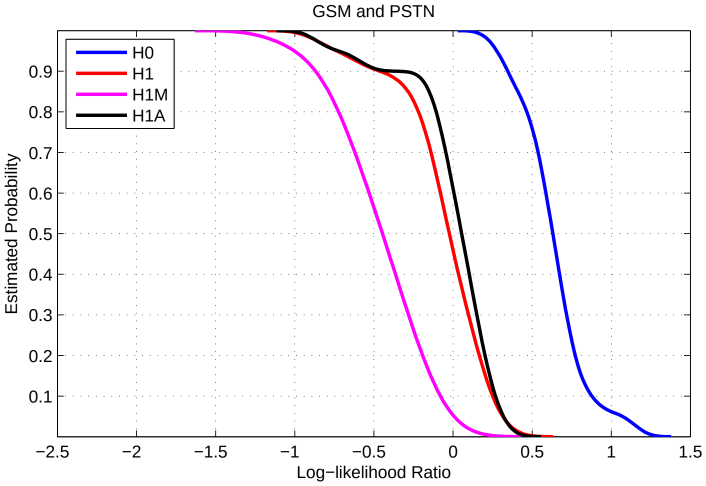

Figure 4.7: Matched, mismatched and adapted conditions (Tippett Plot)

H_0 (> 0)	H_1 Matched (< 0)	H_1 Mismatched (< 0)	H_1 Adapted (< 0)
100%	52.97%	94.82%	38.68%

Table 4.1: Population database in GSM and trace in PSTN (Results)

4.2.2 Population database in PSTN and trace in GSM

This situation is the opposite of the case presented in Sec. 4.2.1, the population models in GSM conditions serve us to perform the matched test and the models in PSTN are used to the mismatched and adapted tests. The following plots illustrate the performance of this system.

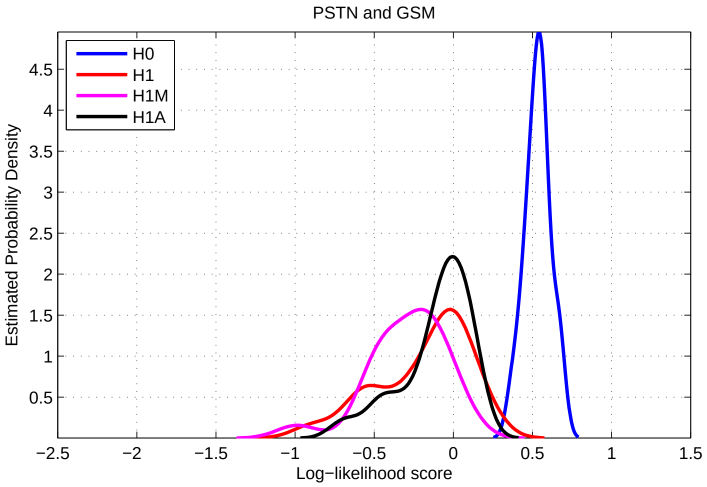

Figure 4.8: Matched, mismatched and adapted conditions (PDF Plot)

From this experiment we realize that the adapted distribution is shifted to the position of the matched conditions, however the height and the variance are different from the matched one. Nevertheless, the adapted curve fits pretty well to the matched one as in the previous case. The Tab. 4.2 presents the results as explained in the previous section.

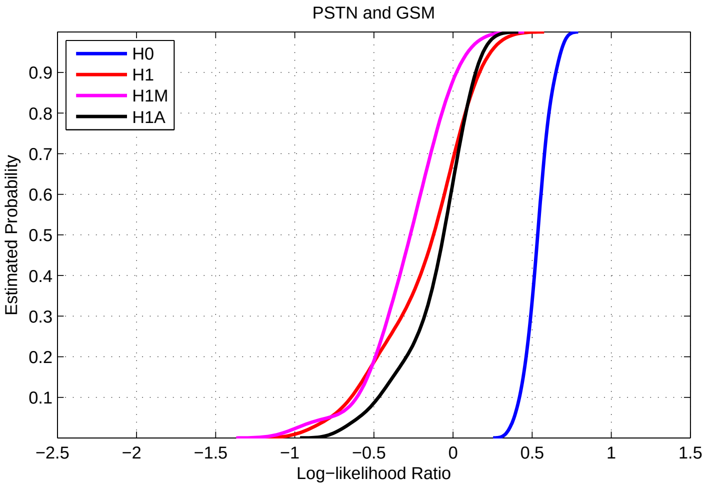

Figure 4.9: Matched, mismatched and adapted conditions (CDF Plot)

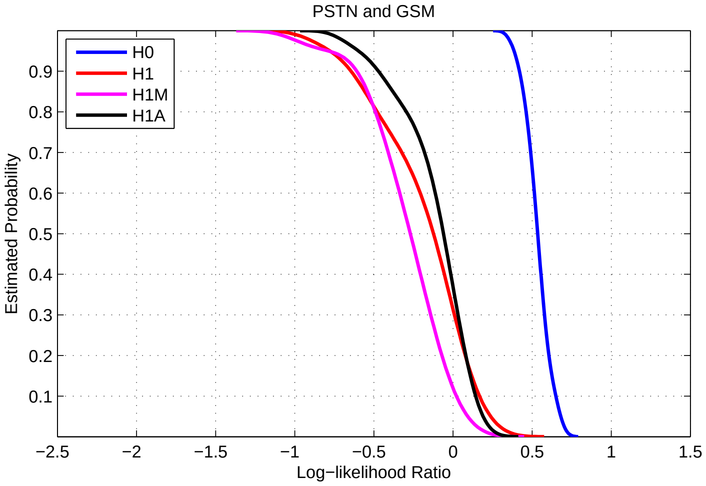

Figure 4.10: Matched, mismatched and adapted conditions (Tippett Plot)

H_0 (> 0)	H_1 Matched (< 0)	H_1 Mismatched (< 0)	H_1 Adapted (< 0)
100%	68.65%	88.83%	61.86%

Table 4.2: Population database in PSTN and trace in GSM (Results)

4.2.3 Population database in GSM and trace in Room acoustic

For this situation, we trained the population models in room acoustic conditions and we used the models in GSM conditions that we had trained for the previous tests. Once we had the two set of models, we compute the test using the models in room acoustic conditions to evaluate the matched case and the models in GSM conditions for the mismatched and adapted test.

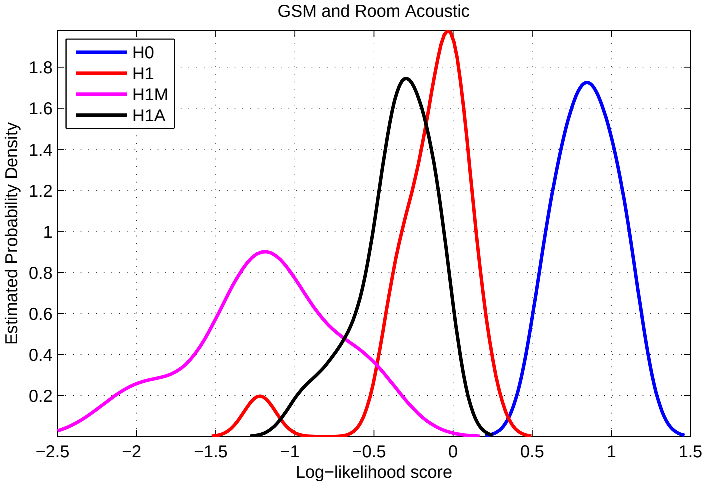

Figure 4.11: Matched, mismatched and adapted conditions (PDF Plot)

Surprisingly, this experiment has the worst performance of the four cases, even if the experiment developed in Sec. 4.1.3 has the best performance in the speaker verification system. The adapted curve does not fit, in terms of mean and variance, with the matched distribution and this produce important errors. The results can be found in Tab. 4.3.

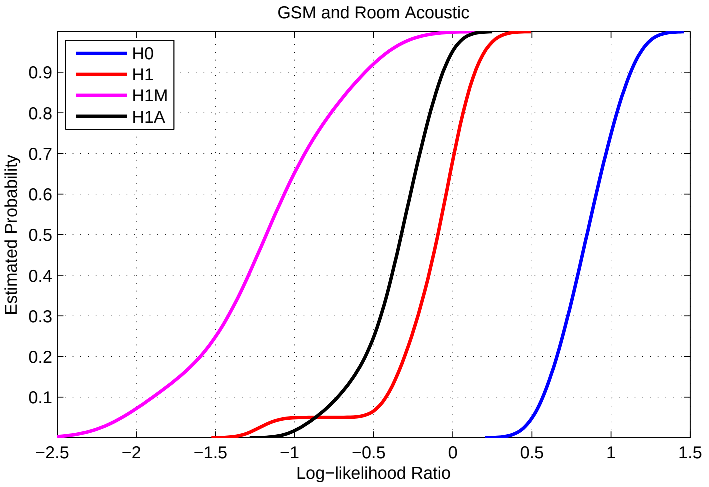

Figure 4.12: Matched, mismatched and adapted conditions (CDF Plot)

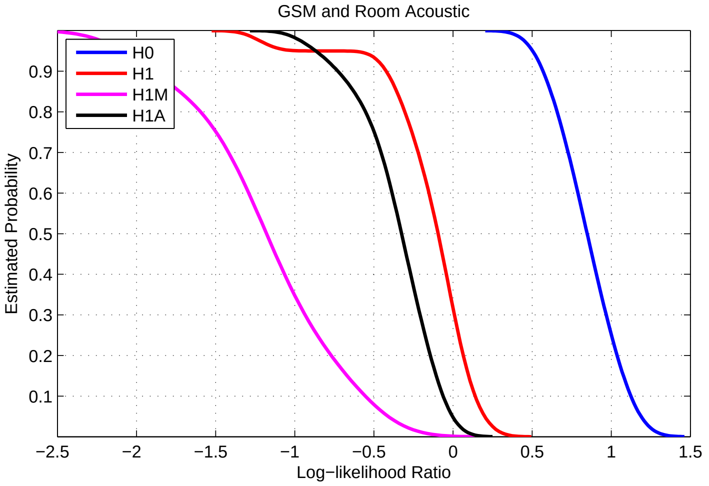

Figure 4.13: Matched, mismatched and adapted conditions (Tippett Plot)

H_0 (> 0)	H_1 Matched (< 0)	H_1 Mismatched (< 0)	H_1 Adapted (< 0)
100%	95.34%	99.84%	69.38%

Table 4.3: Population database in GSM and trace in Room acoustic (Results)

4.2.4 Population database in Room acoustic and trace in GSM

As in Sec. 4.2.2, this simulation is the opposite of the experiment in Sec. 4.2.3. The population models used are the same but the trace conditions are GSM. The population models under GSM conditions are used to compute the matched test and the models under room acoustic conditions serve for the mismatched and adapted test.

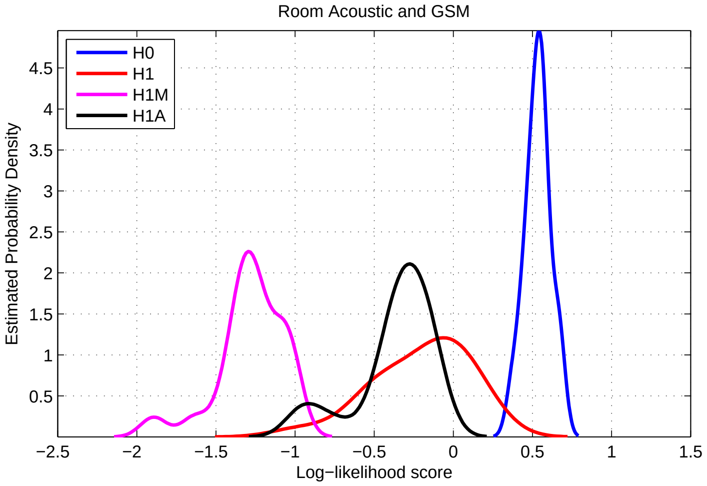

Figure 4.14: Matched, mismatched and adapted conditions (PDF Plot)

In this case, we have managed to adapt quite good the H_1 mean, compared to the mismatched conditions, but the problem of unequal variance and height has to be taken into account, the matched and adapted distributions have a totally different aspect. In Tab. 4.4 we show the performance of this case.

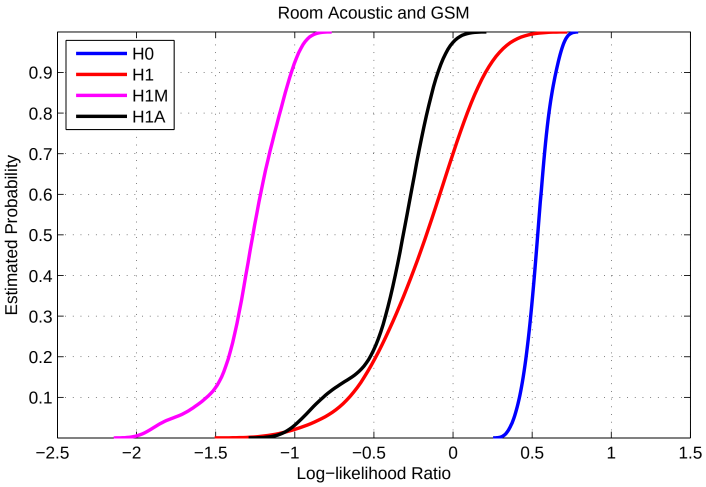

Figure 4.15: Matched, mismatched and adapted conditions (CDF Plot)

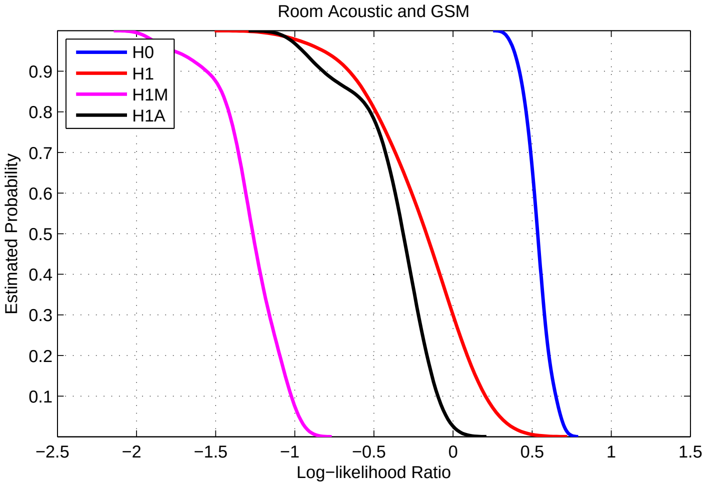

Figure 4.16: Matched, mismatched and adapted conditions (Tippett Plot)

H_0 (> 0)	H_1 Matched (< 0)	H_1 Mismatched (< 0)	H_1 Adapted (< 0)
100%	97.34%	100%	70.24%

Table 4.4: Population database in Room acoustic and trace in GSM (Results)

Chapter 5

Conclusions and future work

In this work, we handled the state-of-art method presented for modeling the speaker and session variability, the joint factor analysis. Ideally, in a forensic framework, all databases (reference, control and population databases) must be recorded in the same conditions for a better performance. As we can consider that practically every case is unique, it is almost impossible to have a database with recordings under the new conditions. This new approach is quite a convenient technique to solve the problem of mismatch in the databases. One important criterion that affects the performance of the system is the size of the database used to model the channel or session conditions, more than one recording per condition is needed to train a good session model. Furthermore, we need a database with recordings in several conditions in order to react to any possible situation.

In the evaluation chapter, we saw that the joint factor analysis models work well when the databases are recorded in the same conditions. However, in mismatched conditions, the performance degrades even reaching the point of working completely wrong. The performance of the system can be enhanced by applying an adaptation in the conditions. In this case, the system can work correctly but not how it would do ideally. We tested only the situation when only one database is in mismatch, an idea for the future could be evaluate the case of more than one database in mismatch.

As we said before, the performance of the technique depends not only on the number of recordings used to model the session but also to other factors. These factors can be, i.e., the number of speakers used to train the UBM, that is a hyperparameter of the joint factor analysis model. Another factor is the number of speakers used to estimate the H_1 distribution curve. The GMM parameters such the number of mixtures, iterations, the number of features, are important factors to improve the robustness of the speaker model. In this thesis we tried to obtain the best results optimizing some parameters. The future work can be focused on finding the optimal value of the parameters to improve the performance of the system.

To summarize, the joint factor analysis is a powerful technique to compensate the mismatch of conditions. In the speaker verification system we have seen that it can work perfectly, but there is still much work to do in the forensic case because this field requires a very high accuracy and rigor. The next step is obtain an adapted distribution that fits perfectly the curve under matched conditions in order to present it to the court and the experts.

Appendices

Appendix A

Description of the Polyphone IPSC-03 database

This database description has been taken from [1].

This database for forensic speaker recognition was recorded by the Institut de Police Scientifique (IPS), University of Lausanne, and the Signal Processing Institute, Swiss Federal Institute of Technology, Lausanne (EPFL). It contains speech from 73 male speakers, in three different recording conditions and several different controlled and uncontrolled speaking modes. This database was recorded between January and June 2005.

The recordings for the database were made in controlled conditions, in a quiet room, located in the IPS and École des Science Criminélles (ESC) building of UNIL.

The recording conditions of this database include transmission through a public switched telephone network (PSTN), a global system for mobile communications (GSM) network as well as calling-room acoustic conditions. The recording room contained a fixed line (PSTN) and mobile (GSM) telephone, and they both used the Swisscom ®telephone network provider.

The fixed line telephone instrument was a 'Meridian, Northern Telecom ®', and the mobile handset was a Nokia ®8310.

All the telephone calls were made from the recording room to an ISDN server located at the Signal Processing Institute, EPFL. The European ISDN (DSS1) transmission standard was used, and an answering machine application was used to record the calls. The transmitted speech was sampled at 16,000 Hz and recorded as 16-bit linear PCM Microsoft WAV files.

Along with these recordings, a third recording condition was simulated using a microphone and recorder placed directly in the recording room. The subjects spoke into a Sony ®electret condenser microphone (CARDIO ECM-23), placed at a distance of about 30cm from the mouth of the speaker and connected to a Sony ®portable digital recorder (ICD-MS1). This speech was recorded in MSV format, at a sampling rate of 11,025 Hz. The cues were presented to the subjects in the form of a printed Microsoft ®PowerPoint presentation (in order to avoid introducing the sound of a computer in the room), and care was taken to ensure that the recording room was free of any additional sound-adsorbing material.

The recorded speakers were male, aged between 18 and 50, with a majority being university-educated students, assistants (between 18 and 30 years of age) and faculty from within IPS and EPFL. All the utterances were in French.

The recordings of telephonic speech were made in two sessions with each of the subjects, the first using the PSTN (fixed) recording condition and the second using the GSM (mobile) recording condition. Additionally, two direct recordings, per speaker, were made on the microphone-recorder (digital) system described above. The length of each of these recordings was 10-15 minutes. Thus, four recordings were obtained, per speaker, in three different recording conditions (one in PSTN, one in GSM and two in room acoustic conditions). These conditions were called *Fixed*, *Cellular* and *Digital* respectively.

In addition to the actual text to be read out, the cue sheets contained detailed instructions for completing the task of speaking in three distinct styles. The first of these was the *normal* mode which involved simply reading printed text. The second *spontaneous* mode involved two simulated situations of a death threat call and a call to the police informing them of the presence of a bomb in a toilet. The third (*dialog*) mode involved reading a text in the tone of a conversation. The recordings (MSV format on the recorder and WAV on the answering machine) were edited with CoolEdit Pro 2 ®and grouped into various "subfiles" as described below.

The database contains 11 traces, 3 reference and 3 control recordings, grouped as the *T*, *R* and *C* sub-databases respectively.

- The *T* database consists of 9 files with read text and 2 *spontaneous* files. These 9 files are edited into three groups of 3 files, each having similar linguistic content. The *spontaneous* files are the simulations of calls as described earlier.

- The *R* database consists of recordings of read text only. Two of these recordings are identical one to the other. The content of the R database is similar to the IPSC-01 and IPSC-02 databases.

- The *C* database consists of recordings in the three different modes described above, viz., the *normal*, *spontaneous* and *dialog*.

A total of 73 speakers were recorded for this database, and it should be noted that the recordings for 63 of these are complete, with the four sets of recordings,

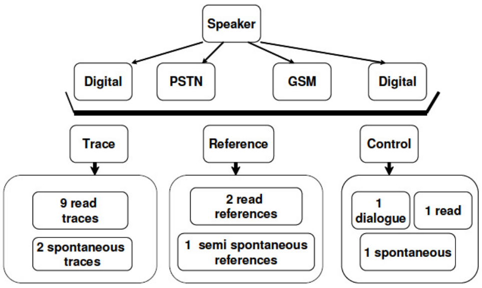

Figure A.1: Layout of the IPSC03 database.

i.e., PSTN, GSM, and two sets of acoustic room recordings. For the remaining 10 are only partially complete and for whom the fourth set of acoustic-room recordings, not available (for technical reasons).

The lengths of the recordings vary from a few seconds for the shortest (T40 and T50) to approximately two minutes for the longest (R01 and R02). This represents a total recording time of approximately 40 to 45 hours.

The nomenclature of the files can be described with an example:

Speaker No.1, in the *fixed* condition, for the file *Control*02, with speech in the *normal* mode, is called *M001FRFC02_NO.wav*.

The individual parts of the filename represent:

- M the sex of the speaker - Male
- 001 the chronological 'number' of the speaker; this number goes from 001 to 073.
- FR the language of speech - French
- C to denote it is a *control* recording. This is replaced by 'T' for the *trace* and by 'R' for the *reference* recordings.
- 02 the number of the subfile. This number can take the values 10, 11, 12, 20, 21, 22, 30, 31 and 32 for the *T* recordings; 00, 01 and 02 for the *R* recordings; and 01, 02 and 03 for the *C* recordings.
- NO the mode of the speech referring to the *normal* speaking mode. These letters are replaced by SP for the *spontaneous* and by DL for *dialog* mode.

The layout of this database is illustrated in Fig. A.1.

Bibliography

[1] A. Alexander. *Forensic automatic speaker recognition using bayesian interpretation and statistical compensation for mismatched conditions.* PhD thesis, EPFL, 2005. [cited at p. 17, 49]

[2] A. Drygajlo. Forensic automatic speaker recognition. *IEEE Signal Processing Magazine*, 24(2):132–135, 2007. [cited at p. 27, 28]

[3] B. G. B. Fauve and D. Matrouf. State-of-the-art performance in text-independent speaker verification through open-source software. *IEEE Trans. Speech Audio Process.*, 15(7):1960–1968, 2007. [cited at p. 20, 21, 22]

[4] J. Gonzalez and B. Baker. On the use of factor analysis with restricted target data in speaker verification. *The Speaker and Language Recognition Workshop*, pages 103–108, 2010. [cited at p. 21]

[5] P. Kenny. Joint factor analysis of speaker and session variability: Theory and algorithms. Tech. Report 06/08-13, CRIM, 2005. [cited at p. 2, 5, 10, 12]

[6] P. Kenny and G. Boulianne. Factor analysis simplified. *ICASSP2005*, 1:637–640, March 2005. [cited at p. 5]

[7] P. Kenny and G. Boulianne. Speaker and session variability in gmm-based speaker verification. *IEEE Trans. Speech Audio Process.*, 15(4):1448–1460, May 2007. [cited at p. 21]

[8] P. Kenny and N. Dehak. A new training regimen for factor analysis of speaker variability. March 2008. [cited at p. 8]

[9] P. Kenny and P. Demouchel. Experiments in speaker verification using factor analysis likelihood ratios. *Odyssey: The speaker and language recognition workshop*, pages 219–226, 2004. [cited at p. 14]

[10] P. Kenny and P. Demouchel. Eigenvoice modeling with sparse training data. *IEEE Trans. Speech Audio Process.*, 13(3):345–354, 2005. [cited at p. 2, 6, 8, 9, 10, 11, 14, 22]

[11] P. Kenny and M. Mihoubi. New map estimators for speaker recognition. *Proc. Eurospeech*, September 2003. [cited at p. 2, 6]

[12] D. Reynolds and T. Quatieri. Speaker verification using adapted gaussian mixture models. *Digital Signal Processing*, 10:19–41, 2000. [cited at p. 2]

[13] C. Vair and D. Colibro. Channel factors compensation in model and feature domain for speaker recognition. *Proc. Odyssey*, pages 1–6, 2006. [cited at p. 23]

[14] R. Vogt and B. Baker. Factor analysis subspace estimation for speaker verification with short utterances. *Interspeech*, 2008. [cited at p. 13]

[15] R. Vogt and S. Sridharan. Experiments in session variability modelling for speaker verification. *Proc. ICASSP*, pages I–987–I–900, 2006. [cited at p. 22]