UNIVERSIDAD AUTÓNOMA DE MADRID

ESCUELA POLITÉCNICA SUPERIOR
DEPARTAMENTO DE TECNOLOGÍA Y DE LAS COMUNICACIONES

TEMPORAL CONTOURS IN LIGUISTIC UNITS FOR AUTOMATIC TEXT-INDEPENDENT SPEAKER RECOGNITION

–TRABAJO FIN DE MÁSTER–

CONTORNOS TEMPORALES EN UNIDADES LINGÜÍSTICAS PARA EL RECONOCIMIENTO AUTOMÁTICO DE LOCUTOR INDEPENDIENTE DE TEXTO

Author: Javier Franco Pedroso
(Ingeniero de Telecomunicación,
Universidad Politécnica de Madrid)

A thesis submitted for the degree of:

Máster Oficial en Ingeniería Informática y de Telecomunicación
(Master of Science)

Madrid, June 2013

Departamento: Tecnología Electrónica y de las Comunicaciones
Escuela Politécnica Superior
Universidad Autónoma de Madrid (UAM), SPAIN

Título: Temporal Contours in Linguistic Units for Automatic Text-Independent Speaker Recognition

Autor: **Javier Franco Pedroso**
Ingeniero de Telecomunicación
(Universidad Politécnica de Madrid)

Director: **Prof. Joaquín González Rodríguez**
Doctor Ingeniero de Telecomunicación
(Universidad Politécnica de Madrid)
Universidad Autónoma de Madrid, SPAIN

Fecha: 19 de Junio de 2013

Tribunal: Presidente:
Prof. Joaquín González Rodríguez
Universidad Autónoma de Madrid, SPAIN

Vocal 1:
Prof. Doroteo Torre Toledano
Universidad Autónoma de Madrid, SPAIN

Vocal 2:
Dr. Daniel Ramos Castro
Universidad Autónoma de Madrid, SPAIN

Suplente:
Prof. Julián Fiérrez Aguilar
Universidad Autónoma de Madrid, SPAIN

Calificación:

The research described in this Thesis was carried out within the ATVS – Biometric Recognition Group at the Dept. of Tecnología Electrónica y de las Comunicaciones, Escuela Politécnica Superior, Universidad Autónoma de Madrid. The work was partially funded by the PriorSpeech project (TEC2009-14719-C02-01, Ministerio de Ciencia e Innovación).

Abstract

THIS MASTER THESIS IS FOCUSED ON designing an automatic speaker recognition system that exploits ideas from linguistic/phonetic traditional human-based speaker recognition and current forensic science practice, integrating them into a likelihood-ratio framework for Bayesian inference of identity. In order to do this, feature extraction protocols from linguistic units are exploited with the aid of powerful automatic speech recognition technologies. Speakers have been modeled from these features and tested by means of the most commonly used automatic speaker recognition technologies, evaluating the performance on challenging datasets (NIST SRE 2004, 2005 and 2006) and evaluation protocols (NIST SRE 2006 English-only 1side-1side male trials). Results have been analyzed in terms of both discriminating power and calibration properties, showing to what extent such kind of approach can be used for forensic purposes. Furthermore, the issue of efficiently combining different pieces of information has been addressed, showing that speaker individualizing information scattered among linguistic units can be fused in several useful ways.

A MIS PADRES.
A MIS HERMANOS.
A MI TÍA CÁNDIDA.

A GEMA.

A GRETA.

Acknowledgements

THIS MASTER THESIS summarizes the work carried out in the last two years during my Master studies with the ATVS – Biometric Recognition Group at the Dept. of Tecnología Electrónica y de las Comunicaciones, Escuela Politécnica Superior, Universidad Autónoma de Madrid. The work was partially funded by the PriorSpeech project (TEC2009-14719-C02-01) of the Spanish Ministry of Science and Innovation.

Foremost, I would like to thank my advisor Prof. Joaquín González-Rodríguez for giving me the opportunity of working in this so interesting research line, and especially for his guidance and support over all this time. The confidence he has always shown in me has definitely fostered my motivation. In the framework of the ATVS research group, I have received also the support from Prof. Javier Orgtega-García and from Prof. Dorote T. Toledano, who have given me the opportunity to work on some other stimulating projects. Also I would like to thank two people that have also guided me and helped me to carry out this Master Thesis, and not just for that but especially for their friendship: Dr. Daniel Ramos and Dr. Javier González-Domínguez.

It is mandatory to thank specially my work mate Fernando Espinoza because we have worked side by side, and this Thesis is in part the result of his work.

Of course I also would like to thank to every mate I had at ATVS for sharing good (and not so good) times in daily work: Ignacio López, Rubén Vera, Álvaro Diéguez, Ruifang Wang, Miriam Moreno, Pedro Tomé, Javier Galbally, Ester Sosa, Marta Gómez, Alicia Lozano, Rubén Zazo, Ram Prasad, Julián Fiérrez, María Puertas, Almudena Gilperez, Sara Antequera, Alberto Montero, Lucas Pérez, Manuel Freire, Víctor González, Ismael Mateos, Danilo Spada, Javier Simón and Alejandro Abejón. All of you are excellent people.

Outside the ATVS, I would like to thank my lifelong friends, and of course my parents, brothers and aunt, but very specially to Gema and Greta, because you fill my life of love and happiness.

Javier Franco Pedroso
Madrid, Junio de 2013

Contents

List of Figures

List of Tables

Chapter 1: Introduction

AUTOMATIC SPEAKER RECOGNITION has focused in the last decade on two concurrent problems: the compensation of session variability effects, mainly through high-dimensional supervectors and latent variable analysis [Kenny *et al.*, 2005] [Kenny *et al.*, 2008] [Dehak *et al.*, 2011], and the production of an application-independent calibrated likelihood ratio per speaker recognition trial [Brummer and du Preez, 2007], able to elicit useful speaker identity information to the final user with any given application prior. The results are highly efficient text-independent systems in challenging conditions, as the Speaker Recognition Evaluations (SRE) organized by the US National Institute of Standards and Technologyñ (NIST), where lots of data from hundreds of speakers in similar conditions are available. Thus, all the speech available in every trial is used to produce detection performances difficult to imagine a decade ago.

However, in the presence of strong mismatch (as e.g. in forensic conditions, where acoustic and noise mismatch, apart from highly different emotional contexts, speaker roles or health/intoxication states can be present between the control and questioned speech), those acoustic/spectral systems could be unusable as all our knowledge about the two speech samples is deposited into a single likelihood ratio, obtained from all the available speech in the utterance this LR could be strongly miscalibrated (being then highly misleading) as the system has been developed under severe database mismatch between training and testing data. Moreover, it is difficult (or even impossible) to collect enough data to develop a system robust to every combination of mismatch factors present in actual case data, an important problem in real applications.

Traditional forensic voice comparison is usually based on the analysis of temporal dynamics of higher level features in the context of linguistic units. A usual procedure in forensic laboratories is that a speech expert, typically a linguist/phonetician, can isolate or mark segments of compatible/comparable speech between both samples, segments being from seconds long to just some short phonetic events in given articulatory contexts that corresponds to some linguistic units.

Formant analysis has a long tradition in forensic phonetics, and they are features that linguists and phoneticians are comfortable with when defending in court. Formant frequencies and their dynamics have shown strong individualization potential [McDougall, 2006][Nolan, 1983], and different researchers, mostly linguists and phoneticians following the pioneering steps of Phil Rose [de Castro *et al.*, 2009][Morrison, 2009][Rose, 2002][Zhang *et al.*, 2008], have shown how to report likelihood ratios (LRs) from human-supervised formant trajectories, complying with the requisites of modern forensic science [Rose, 2002][Gonzalez-Rodriguez *et al.*, 2007].

However, as formant frequencies are manually extracted and/or supervised for every linguistic unit of interest, a very limited percentage of the available data can be processed, as huge amount of human work is needed. Also, the number and types of comparable units for analysis is always a case-dependent subject, and therefore flexible strategies for analysis and combination are needed.

1.1 Motivation

The motivations of this Master Thesis come from the issues stated above. Forensic voice comparison and automatic speaker recognition communities have kept apart from each other with the only exception of adopting both the likelihood-ratio as the proper way of presenting the strength of the evidence. Apart from this, the procedures they follow and the features they use are usually completely different. While automatic speaker recognition systems apply brute force processing enormous amounts of data in the same way without incorporating some other knowledge apart from what they learn from very complex modeling techniques, in forensic voice comparison most of the processes are done manually and paying close attention to cues derived from knowledge of the speech production process from a linguistic point of view.

There are two main reasons in order to make them converge. On the one hand, it is important to present speech evidences in court in a coherent way as they are used to forensic voice comparison has been traditionally performed by phoneticians. On the other hand, some of the knowledge that underlies in the linguistic rules of the speech production process can help to better discriminate between speakers.

Furthermore, while there are several studies in the automatic speaker recognition field about temporal contours of prosodic features like pitch or energy, formant trajectories and cepstral contours haven't been investigated in depth, especially when they are constrained to a particular linguistic unit. In the case of formant trajectories, they have been largely investigated in the forensic community, but on very limited datasets due to the manual annotation and semi-supervised feature extraction.

1.2 Goals of the Master Thesis

The main goal of this Master Thesis is to develop a fully automatic framework for performing fine-grained speaker recognition based on temporal dynamics of acoustic features in the context of linguistic units, analyzing the individual performance achieved by using isolated units, which is a common procedure in forensic voice comparison. The performance of these systems will be tested on standard *de facto* databases commonly used in the automatic speaker recognition field.

On the one hand, the use of formant trajectories as features will be analyzed because they have been largely used in forensic phonetics and are understandable cues in order to report forensic evidences in court. On the other hand, we are also interested on cepstral contours as an attempt to merge the interpretability of temporal contours with the discriminating power of MFCC features.

In addition to analyze the individual performance of systems based on isolated linguistic units, several combination methods will be applied to see how all this information scattered over the different units can be merged in a way that leads to a performance improvement.

Finally, as it has been the main purpose of higher level systems for text-independent speaker recognition, formant trajectories and cepstral contour based systems will be fused with state-of-the-art speaker recognition systems.

1.3 Outline of the Dissertation

The chapter structure of the Dissertation is as follows:

- Chapter 1 introduces the issues of applying automatic speaker recognition to forensic voice comparison, the proposed framework, and gives the motivation, outline and contributions of this Master Thesis.
- Chapter 2 presents the state of the art in automatic speaker recognition and summarizes related works, detailing the motivations of the Thesis.
- Chapter 3 describes the proposed approach, introducing the features used and detailing how the available technologies are applied to our purpose.
- Chapter 4 describes the speech databases and protocols used to test the proposed approach, as well as the performance metrics that measures both the accuracy and the reliability of the systems.
- Chapter 5 presents the results obtained on our experimental framework by both reference and proposed systems, analyzing the strengths and weaknesses of our approach.
- Chapter 6 concludes the Dissertation summarizing the main results obtained and outlining future research lines.

1.4 Research contributions

The research work conducted along this Master Thesis has led to the following publications:

J. Gonzalez-Rodriguez, J. Gonzalez-Dominguez, **J. Franco-Pedroso** and D. Ramos. "*A linguistically-motivated speaker recognition front-end through session variability compensated cepstral trajectories in phone units*". Proceedings of the 2012 International Conference on Acoustics, Speech and Signal Processing (ICASSP 2012), pp. 4389-4392. March 2012. Kyoto, Japan.

J. Franco-Pedroso, J. Gonzalez-Rodriguez, J. Gonzalez-Dominguez and D. Ramos. "*Fine-grained automatic speaker recognition using cepstral trajectories in phone units*". Proceedings of the 2012 Annual Conference of the International Association for Forensic Phonetics and Acoustics. 5th-8th August 2012, Santander, Spain.

J. Franco-Pedroso, J. Gonzalez-Rodriguez, J. Gonzalez-Dominguez and D. Ramos. "*Fine-grained automatic speaker recognition using cepstral trajectories in phone units*". Quantitative approaches to problems in linguistics - Studies in honour of Phil Rose (ISBN 978-3-86288-384-4). Cathryn Donohue, Shunichi Ishihara, William Steed (ed.). LINCOM Studies in Phonetics 08. 2012.

J. Franco-Pedroso, F. Espinoza-Cuadros and J. Gonzalez-Rodriguez. "*Cepstral Trajectories in Linguistic Units for Text-Independent Speaker Recognition*". Proceedings of IberSPEECH 2012: "VII Jornadas en Tecnología del Habla" and III Iberian SLTech Workshop. 21-23 November 2012. Madrid, Spain.

J. Franco-Pedroso, F. Espinoza-Cuadros and J. Gonzalez-Rodriguez. "*Formant Trajectories in Linguistic Units for Text-Independent Speaker Recognition*". Proceedings of ICB-2013: The 6th IAPR International Conference on Biometrics. June 4-7, 2013. Madrid, Spain.

Chapter 2: Speaker recognition systems

THIS CHAPTER PRESENTS firstly the main technologies that have contributed to the state of the art of automatic speaker verification systems and that are somehow related to the systems developed in our proposed approach. Then, the main related works are summarized in order to have an idea of what has been done by the research community to analyze what have worked, what have not, and what hasn't been done yet.

2.1 State of the art

In this section, the evolution of automatic speaker recognition systems over the last ten years will be briefly presented, paying particular attention to the technologies in which our proposed approach is based and those that have contributed significantly to the current state-of-the-art. In this sense, some of the technologies less related with the systems involved in this Master Thesis, despite being widely used, have been left apart. This is the case of those based on Support Vector Machines [Campbell *et al.*, 2006].

2.1.1 GMM-UBM

The Gaussian Mixture Model (GMM) – Universal Background Model (UBM) framework [Reynolds *et al.*, 2000] has been the state of the art in text-independent speaker recognition from short-term spectral features (typically MFCC feature vectors) for many years until the emergence of new techniques based on factor analysis [Kenny *et al.*, 2008]. This scheme can be seen as a likelihood ratio detector between a GMM target model and an independent GMM model, the so-called UBM. The UBM model is trained with speech (features 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 obtained for the target and the UBM model.

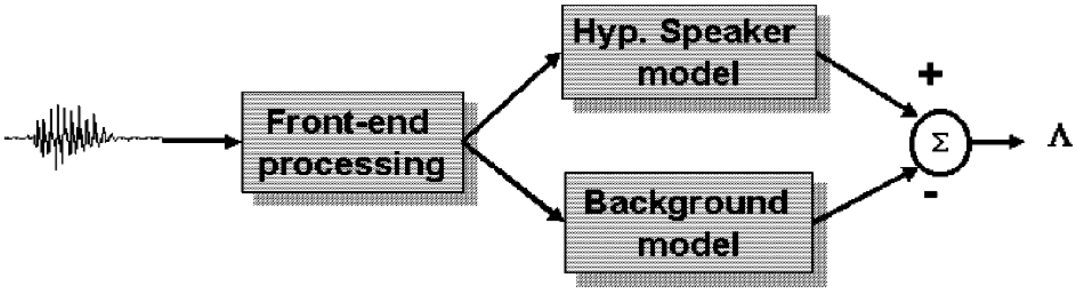


Figure 1. Likelihood ratio-based speaker detection system (extracted from [Reynolds *et al.*, 2000]).

2.1.1.1 GMMs

A GMM (λ) is a likelihood function for F-dimensional feature vectors x_t given by the following mixture density

$$p(x_t|\lambda) = \sum_{c=1}^{C} \omega_c p_c(x_t)$$

where the mixture weights, ω_k, satisfy the constraint $\sum_{c=1}^{C} \omega_c p_c = 1$ and $p_c(x_t)$ is a F-variate normal density parameterized by a mean F×1 vector, $\boldsymbol{\mu_c}$. and a F×F covariance matrix, Σ_c:

$$p_c(x_t) = N(x_t|\mu_c, \Sigma_c) = \frac{1}{\sqrt{(2\pi)^D|\Sigma_c|}} \exp\, (-\frac{1}{2}(x_t - \mu_c)^T \Sigma_c (x_t - \mu_c))$$

For example, Figure 2 shows a 4-component GMM trained from 2-dimensional feature vectors.

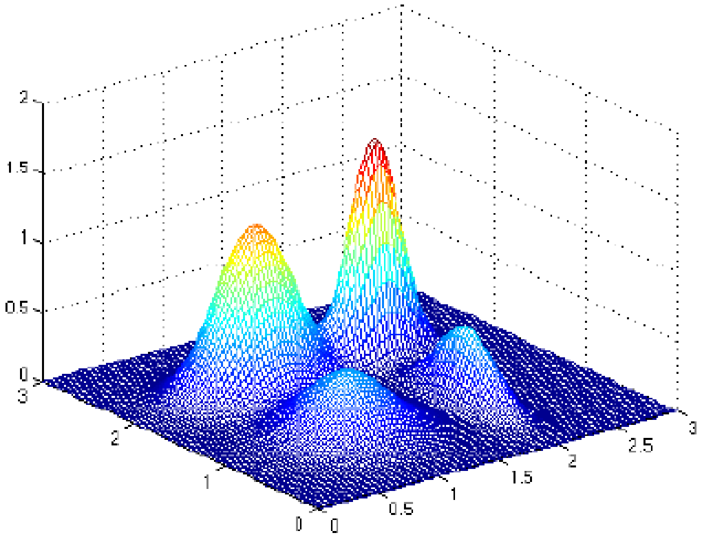


Figure 2. A 4-component GMM trained from 2-dimensional feature vectors

2.1.1.2 UBM training

As it has been previously mentioned, the UBM is a large GMM representing the speaker-independent distribution of features corresponding to the alternative hypothesis (in contrast to the target speaker hypothesis), and so it is trained on speech that is reflective of the expected alternative speech to be encountered during recognition.

Training a GMM consists of estimating the parameters $\lambda = \{\omega_k, \mu_k, \Sigma_k\}_{k=1}^{K}$ from a set of training observations. This is usually done by means of a Maximum Likelihood estimation, through the *Expectation-Maximization* (EM) algorithm [Dempster *et al.*, 1977], after a clustering stage via K-means [Linde *et al.*, 2003] to favor a quick convergence of the EM algorithm.

2.1.1.3 MAP adaptation

Usually, the amount of speech samples available for target speaker modeling are not enough to properly train a GMM from scratch. The basic idea in the adaptation approach is to derive the speaker's model by updating the well-trained parameters in the UBM via adaptation. For this purpose, *Maximum A Posteriori* (MAP) adaptation [Gauvain and Lee, 1994] is used. Although all the parameters of the model $(\omega_c, \mu_c, \Sigma_c)$ can be adapted, usually just the mean vectors are updated.

Given a UBM and training vectors from the hypothesized speaker, X=\{x_1, ...,x_T\} the adapted mean new vectors are derived as a trade-off between the UBM model means, μ_c, and the new data in the form

$$\mu_c' = \alpha_c \frac{1}{n_c} f_c + (1 - \alpha_c)\mu_c$$

where

$$\alpha_c = \frac{n_c}{n_c + \tau}$$

$$n_c = \sum_{t=1}^{T} P_{ct}$$

$$f_c = \sum_{t=1}^{T} P_{ct} x_t$$

$$P_{ct} = \frac{\omega_c p_c(x_t)}{\sum_{c=1}^{C} \omega_c p_c(x_t)}$$

being n_c and f_c the so-called 0th and 1st-order statistics respectively, P_{ct} the Gaussian occupation probability and τ the relevance MAP factor, which controls the importance of training samples and the UBM within the adaptation procedure.

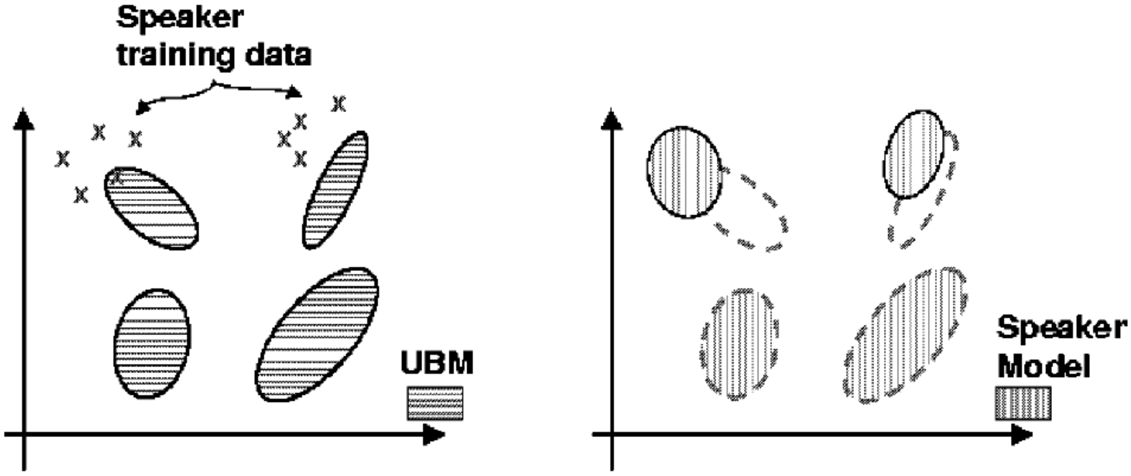


Figure 3. Example of a MAP adaptation in a 2-dimensional feature space (extracted from [Reynolds *et al.*, 2000]).

2.1.1.4 Log-likelihood ratio

In the verification stage, the score for a set of testing observations $X=\{x_1, ..., x_T\}$ given the target speaker model, λ_s, is computed as a likelihood ratio between the speaker model and the background model, λ_{UBM}. Taking logs this takes the form

$$\mathcal{L}(X, \lambda_s, \lambda_{UBM}) = \frac{1}{T}\sum_{t=1}^{T}\{\log p(x_t|\lambda_s) - \log p(x_t|\lambda_{UBM})\}$$

Thus, the difference in likelihood between the target and the background model in generating the observations X are measured, doing comparable the score ranges of different speakers.

2.1.2 Joint Factor Analysis

One of the main problems of acoustic systems for speaker recognition is the variability affecting short-term spectral features when dealing with recordings coming from different transmission channels, captured using different microphones, and so on (the so-called session variability). First successful approaches, like Cepstral Mean Normalization (CMN) [Atal, 1974; Furui, 1981], RASTA filtering [Hermanski and Morgan, 1994; Malayath *et al.*, 2000] and Feature Warping [Pelecanos and Sridharan, 2001], tried to solve this problem acting directly on the features according to a general rule, rather than modeling the specific variability within a given recording. Next approaches, like Speaker Model Synthesis [Teunen *et al.*, 2000] or Feature Mapping [Reynolds, 2003] tried to model channel variability by training or adapting channel-specific models in order to match the channel conditions between the testing recording and the target speaker model, but in a discrete manner. Furthermore, none of these approaches deals with the variability due to the target speaker itself between different recordings due to different phonetic content, speaking rate, mood, etc.

The factor analysis (FA) modeling approaches [Kenny and Dumouchel, 2004] combine the previous techniques of GMM-MAP adaptation [Reynolds *et al.*, 2000], eigenvoice-MAP [Kenny *et al.*, 2003] and eigenchannel-MAP [Kenny *et al.*, 2005], breaking with the established manner of conceiving the variability associated to a speech signal by considering variability as a continuous source rather than discrete and by explicitly modeling both session and inter-speaker variability. Moreover, there is a fundamental hypothesis that is the basis of FA approaches: 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 speaker and session variability subspaces, so that they act as priors in order to disclose the specific variability contained in a given recording.

All these approaches are defined in a supervector space: as in the common GMM-UBM framework the weights and the covariances are usually shared between the UBM and the adapted speaker models, every model can be represented in a common feature space as a vector formed by the concatenation of the model means (as the model means are vectors itself, the vector formed by their concatenation is called supervector).

2.1.2.1 GMM-MAP in the supervector space

Given a GMM-UBM system with C components defined in an F-dimensional feature space, a speaker-dependent means model supervector μ_s (CF×1) for a speaker s is derived by MAP from the UBM means supervector μ as

$$\mu_s = \mu + Dz_s$$

where the term Dz_s represents the shift/offset from the mean μ as a result of the MAP adaptation, and it is formed by the diagonal CF×CF matrix D, and the CF×1 weights vector z_s which is assumed to be distributed with a standard normal prior. By the form in previous equation and assuming the prior of z standard normal distributed, it can be inferred that, in MAP, speaker-dependent means supervectors are considered to be normally distributed with mean μ and covariance $B = D^2$, CF×CF.

2.1.2.2 Eigenvoice-MAP

Eigenvoice-MAP performed an analogous analysis to GMM-MAP but considering that the variance of the distribution is restricted to a subspace of rank R_s within the supervector space, where $R_s << CF$. Note that the implicitly assumption formulated in eigenvoice-MAP is then that the eigen-analysis of covariance B results on a few non-zero eigenvalues, exactly R_s. In matrix form

$$\mu_s = \mu + V y_s$$

where V is a low-rank matrix ($CF \times R_s$) which explains the speaker variance, in this case $B = VV^T$ and y_s the weights which represent the speaker *s* through the speaker variability subspace spanned by V. Note, nevertheless, that by varying y_s, the model μ_s varies across the space spanned by V; that is within a R_s-dimensional linear manifold of the supervector space. The vector y_s ($R_s \times 1$) is usually referred to as the *speaker factors*, since it represents the speaker variability within V, and mathematically responds to the latent factors within a FA modeling framework.

2.1.2.3 Eigenchannel-MAP

It is assumed that the supervector obtained for a speaker *s* given an utterance *h* is a distorted version of the *true* speaker supervector due to the session variability. Eigenchannel-MAP models this distortion modifying the supervector space by and additional term as

$$\mu_{sh} = \mu_s + U x_{sh}$$

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

2.1.2.4 The JFA model

JFA integrates both GMM-MAP and eigenvoice-MAP modeling ideas in order to derive the speaker-dependent component of a mean speaker supervector model. So that

$$\mu_s = \mu + V y_s + D z_s$$

Note that by this form the assumed variance B is now explained by both V and D (B = VV^T + D^2), and as such, it combines the advantages of GMM-MAP and eigenvoice-MAP: first, the variability is supposed to be, to a great extent, constrained in the subspace spanned by V; and second, other speaker variability out of this manifold is also accounted.

Moreover, JFA also adds the session variability effect modeled by eigenchannel-MAP, being the final JFA model represented in matrix form as

$$\mu_{sh} = \mu + Vy_s + Dz_s + Ux_{sh}$$

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 Λ = {μ,V,D,U}, the speaker-dependent factors y_s, z_s and the speaker- and utterance-dependent x_{sh} factors. The speaker-independent 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.

2.1.3 Total variability

Total variability [Dehak *et al.*, 2011] represents a step further in the JFA model where a single subspace is trained to jointly model both session and speaker variability. This subspace, the so-called total variability subspace, aims to constrain in a low dimensional space both the session and the speaker variability. Mathematically, this generative latent variable model can be formulated as

$$\mu_s = \mu + Tw$$

where T is the total variability matrix and w are the latent factors of the model, also called total vectors or i-vectors (for *identity* vector). This model allows representing an utterance by a single vector, as the supervector does, but in a much lower dimensionality space where both session and speaker variability is supposed to be confined. In this reduced dimensional space, classical techniques like LDA (linear discriminant analysis) can be easily applied in order to compensate both session and speaker variability, and simple scoring techniques as cosine distance in order to measure the similarity between speech samples [Dehak *et al.*, 2011].

2.2 Related works

2.2.1 Forensic voice comparison from linguistic units

Although in the last years automatic speaker recognition has begun to be used for forensic purposes, forensic voice comparison has been traditionally done by linguistics/phoneticians. A usual procedure in forensic laboratories is that a speech expert, typically a linguist/phonetician, can isolate or mark segments of compatible/comparable speech between control and questioned samples, segments being from seconds long to just some short phonetic events in given articulatory contexts. It has been argued [Rose, 2002] that forensic-phonetic voice comparison should be done based on linguistic features for three reasons: first, because speakers can differ linguistically; second, because it is the linguistic structure that specifies what is comparable (it is common sense to compare samples with respect to the same sound); and third, because phoneticians are able to focus on and describe the speech sound independently of the voice it is being realized in, while the description of non-linguistic data is not so well advanced. The specific linguistic units used for forensic voice comparison depend, of course, on the language spoken by the individuals involved [Li and Rose, 2012] [Zhang *et al.*, 2008], but also on the acoustic features in which the comparison is based.

Both formant frequencies and its dynamics have been largely used for forensic voice comparison. Formant frequencies, initially used to distinguish between speech sounds, were also found to have strong individualization potential [Nolan, 1983]. Usually, formant centre frequencies are extracted at the temporal midpoint of vowels [Rose and Winter, 2010], reflecting in part certain anatomical dimensions of a speaker as the length and configuration of the vocal tract [Stevens, 1971], but also the mean frequencies over the timecourse of the vowel have been used [Zhang *et al.*, 2008].

Formant dynamics were also proposed [McDougall, 2006] under the assumption of presenting higher inter-speaker variability within linguistic units than the static measures of formant centre frequencies. A 'phonetic target' is assumed to be a range of articulatory or acoustic configurations for a given segment which achieve a satisfactory percept of that segment in connected speech. While speakers demonstrate very similar acoustic properties at moments at which targets are achieved (e. g. formant frequencies at a segment's midpoint time-slice), much larger differences are exhibited in the ways they move between the targets [Nolan, 2002]. Formant dynamics are especially suitable for diphthongs because they exhibit more changes in formant frequencies over time [Morrison *et al.*, 2008].

Temporal dynamics of formant frequencies can be parameterized in order to capture speaker-distinguishing properties in an effective and economical way that enables to compare large number of speakers. Linear regression was initially used to approximate a polynomial function to the temporal evolution of each formant frequency [McDougall, 2006] [Morrison, 2008] for different polynomial degrees, showing that different linguistic units require different complexity of fitted curves. Also, Discrete Cosine Transform (DCT) has been used in order to code formant dynamics [Morrison, 2009], keeping a number of the lower order coefficients for a compact representation.

Similar classification or scoring techniques have been used with both instantaneous measurements and temporal dynamics of formant frequencies. While some works have performed discriminant analysis for speaker classification [McDougall, 2006], the most common approach is to derive a similarity score between control and questioned samples in order to obtained a final likelihood ratio per trial. Most of the studies [Gonzalez-Rodriguez *et al.*, 2007] [Morrison, 2008] [Li and Rose, 2012] have used the MVLR approach [Aitken and Lucy, 2005], but also the GMM-UBM framework [Reynolds *et al.*, 2000] has been used [Rose and Winter, 2010]. Both two methods were theoretically designed with the aim of producing likelihood ratios directly, but in practice a further calibration [Brummer and du Preez, 2006] step is needed after applying the scoring process. In both cases, the temporal contours of a specific linguistic unit in a control sample are compared with the temporal contours of the same linguistic unit in the test sample.

Due to the manual annotation of linguistic units, and the manual measurements or semi-supervised fitting of polynomial curves to formant trajectories, a huge amount of human work is needed, reducing drastically the amount of data that can be processed. Moreover, databases used usually consist of clean speech samples, not reflecting the more challenging real acoustic conditions. Finally, no common evaluation framework is used for testing the approaches, making hard to compare the results of the different approaches. Even though very encouraging results have been obtained with these approaches, further experiments are needed.

2.2.2 Syllable-based prosodic contours for speaker recognition

Some research from the automatic speaker recognition field have been focused on other prosodic features [Adami *et al.*, 2003], mainly on pitch and energy contours and unit duration, but also on more complex features like the so-called SNERFs [Shriberg *et al.*, 2005]. Also, there are some works on modeling formant trajectories [Dehak *et al.*, 2007] or MFCC contours [Kockmann and Burget, 2008a]. However, although the feature extraction is performed over some linguistic unit types, in order to model a specific speaker all features coming from the different unit types are grouped together, performing a unit-independent verification.

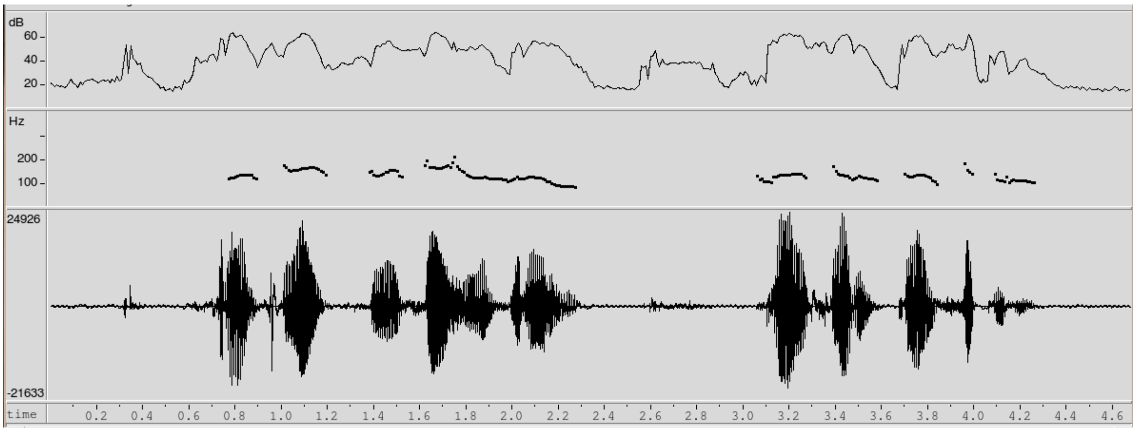


Figure 4. Energy (above) and pitch (between) contour of a speech signal (below) extracted by means of Wavesurfer [Sjolander].

Pitch and energy tracking is usually computed in an automatic way by either the Snack Toolkit/Wavesurfer [Sjolander] or the Praat [Boersma, 2001] software packages, and extracted for syllable-like units [Shriberg *et al.*, 2005] [Kockmann and Burget, 2008a] [Kockmann *et al.*, 2010]. Most of these approaches make use of an automatic speech recognition (ASR) or a Large-Vocabulary-Continuous-Speech-Recognition (LVCSR) system in order to segment the speech signal in to syllables, using human-created rules [Shriberg *et al.*, 2005]. However, other segmentation methods have been explored in [Kockmann *et al.*, 2010] that do not need this kind of systems and so are not language-dependent, although they are not so accurate. For example, syllables can be created from the output of a phone recognizer [Kockmann and Burget, 2008b], or use directly the phone boundaries in order to segment the speech signal. Also, syllable segments can be determined by the Vowel Onset Points (VOP) [Mary and Yegnanarayana, 2008] or estimated from energy valleys [Dehak *et al.*, 2007]. It has been demonstrated [Kockmann *et al.*, 2010] that the higher the quality of the segmentation, the better the performance of the system, but even a simple fixed-length sliding-window scheme can be used with not too much performance degradation.

Once the prosodic features are extracted for syllable-like units, their temporal contours are coded by means of either a Legendre polynomial [Dehak *et al.*, 2007] or the Discrete Cosine Transform (DCT) [Kockmann and Burget, 2008a]. In both methods, the feature segment is modeled by taking the coefficients of an n-th order Legendre polynomial or by taking the n leading coefficients of the DCT, respectively. The final feature vector is made by the concatenation of the pitch and energy contour coefficients, and the duration of the segment is also added.

Regarding the speaker modeling methods, GMM-UBM [Kockmann and Burget, 2008a], JFA [Kockmann *et al.*, 2010a] and total variability [Kockmann *et al.*, 2010b] [Kockmann *et al.*, 2011] frameworks have been used. All these methods require great amounts of data, but due to automatic feature extraction, standard *de facto* speaker recognition databases can be used.

The main interest in these systems for automatic speaker recognition is the improvements they can provide when fusing with acoustic short-term spectral systems, so no analysis has been done on the performance of individual units by itself, which is of interest in the case of forensic applications as it has been stated in the previous section.

2.2.3 ASR dependent speaker recognition systems

Another research line of related works has been focused on what has been called ASR-dependent or text-constrained systems. Similarly to the syllable-based prosodic systems described in the previous section, feature extraction is performed in an ASR- or text-dependent way by means of an LVCSR system, but unlike those systems, text-dependence is kept for the modeling and testing stages. Another difference between these approaches is the type of features used, being the text-dependent systems usually based on MFCC features, although prosodic ones has been used as well [Shriberg and Ferrer, 2007].

Regarding the units or, more generally, the constraints that have been used, first studies [Park and Hanzen, 2002] were focused on phonetic classes instead of isolated phone units, training different models for vowels, fricatives and so on. In [Sturim *et al.*, 2002], complete words were used instead, training, on the one hand, independent models for each word, and on the other hand, independent models for different groups of words. It was shown that the results for the models trained on groups of words perform better than those from the fusion of independent words, and better than using all the speech available, showing that using knowledge about the spoken text could produce low error rates by focusing only on limited acoustic units in the speech.

In [Bocklet and Shriberg, 2009], several syllable-based constraints were used, but instead of modeling specific linguistic units, several aspects shared among different syllables were used; for example, syllable nuclei, onset or codas, but also syllables containing some particular phone or one-syllable words. It was shown that the system based on this latter constraint achieves the best individual result, and the combination of different constraints outperformed the reference system for most of the test conditions.

Similar constraints were used on [Shriberg and Bocklet, 2011] but adding phone-based new ones in order to compare between the usual language-dependent LVCSR system used and a language-independent phone recognizer. Despite considerable differences in constraint region alignments between the language-dependent and language-independent versions, most of the constraints from the language-independent system show surprisingly little degradation from their language-dependent counterparts, being some cases in which the language-independent version outperformed the language-dependent version.

In [Sanchez *et al.*, 2011], new constraints based on prosodic or acoustic (pitch values, voicing frames, etc.) were added, as well as some based on turn-taking or discourse-related and speaking rate. Most of the study is also focused on finding the best system configuration for each constraint instead of using the same configuration for all of them.

Regarding the technologies used, as this kind of systems are based on MFCC features, same modeling techniques as in unconstrained systems have been used, from GMM-UBM [Park and Hanzen, 2002] [Sturim *et al.*, 2002] [Bocklet and Shriberg, 2009] to JFA [Shriberg and Bocklet, 2011] [Sanchez *et al.*, 2011]. For those based on prosodic features, also SVMs have been used [Shriberg and Ferrer, 2007].

All these studies have shown that setting constraints to the speech processed by the systems leads to better performances by taking advantage of focusing on specific sounds when looking for speaker distinguishing information, in part due to the reduction of intra-speaker variability that appears when processing all the speech in an utterance. However, none of these approaches has focused on broader linguistic-constraints, and no analysis from the forensic point of view has been done.

Chapter 3: Exploiting temporal contours in linguistic units

THIS CHAPTER DETAILS the proposed approach in order to analyze, in a fully automatic way, the speaker verification performance that can be obtained, from temporal acoustic features, for isolated linguistic units.

As it has been seen in Chapter 2, some studies from the forensic field have worked with formant trajectories on isolated linguistic units, but these analyses are quite limited due to the manual data processing, focusing on just one or two linguistic units in a database of at most twenty speakers. Furthermore, no analysis of the combination between units has been done.

On the other hand, in the automatic speaker recognition field, when temporal contours have been extracted, linguistic information has been used only for feature extraction purposes, and paying no attention to it in the modeling stage. Conversely, when the speaker modeling process has been done in a unit-dependent way, static acoustic features have been used. Moreover, no analysis has been done from the forensic point of view.

So, the aim of the proposed approach is to combine traditional acoustic-phonetic features and procedures from forensic voice comparison with the power of automatic speaker recognition systems.

3.1 Feature extraction

As already stated, the proposed system is based on temporal contours in linguistic units. These contours have been extracted from both formant frequencies and MFCC feature vectors. Prior to bound linguistic units for temporal contour modeling, feature extraction is performed over the entire speech signal.

3.1.1 Formant trajectories

Formants were initially defined [Gunnar Fant, 1960] as “the spectral peaks of the sound spectrum”. However, in science and phonetics this term is also used to mean the acoustic resonance of the human vocal tract because they are very close to the corresponding maximum in spectrum of the complete sound, making them interchangeable. The formant with the lowest frequency is called *f*1, the second *f*2, etc. Formant frequencies have been used in order to identify speech sounds; for example, most often the two first formants are enough to distinguish between vowels. Also, formant frequencies and their dynamics have shown strong individualization potential for discrimination of speakers [Nolan, 1983] [Rose, 2002] [McDouglas, 2006]. In order to extract these formant frequencies in a fully automatic way, the Wavesurfer formant tracker tool has been used.

Wavesurfer [Sjolander and Beskow, 2000] is a free software audio editor widely used for studies of acoustic phonetics that provides an interactive display for waveform, spectrograms, pitch tracks or transcriptions visualization, therefore being a graphical user-oriented tool. However, it's written in Tcl/Tk [Tcl] using the Snack audio library [Sjolander], what makes it scriptable in order to automatically process a large number of audio files.

Wavesurfer formant tracker estimates speech formant trajectories through dynamic programming, used to optimize trajectory estimates by imposing frequency continuity constraints. The formant frequencies are selected from candidates proposed by solving for the roots of the linear predictor polynomial computed periodically. The local costs of all possible mappings of the complex roots to formant frequencies are computed at each frame based on the frequencies and bandwidths of the component formants for each mapping. The cost of connecting each of these mappings with each of the mappings in the previous frame is then minimized using a modified Viterbi algorithm.

Although Wavesurfer can estimate formant frequencies from *f*1 to *f*4 and their corresponding bandwidths, only the three first formant frequencies have been used for this work, being extracted with a 10 ms time resolution. Figure 5 shows a speech signal in both time (below) and frequency (above) domains, being highlighted on different colors the first 3 formant frequencies extracted by the Wavesurfer formant tracker.

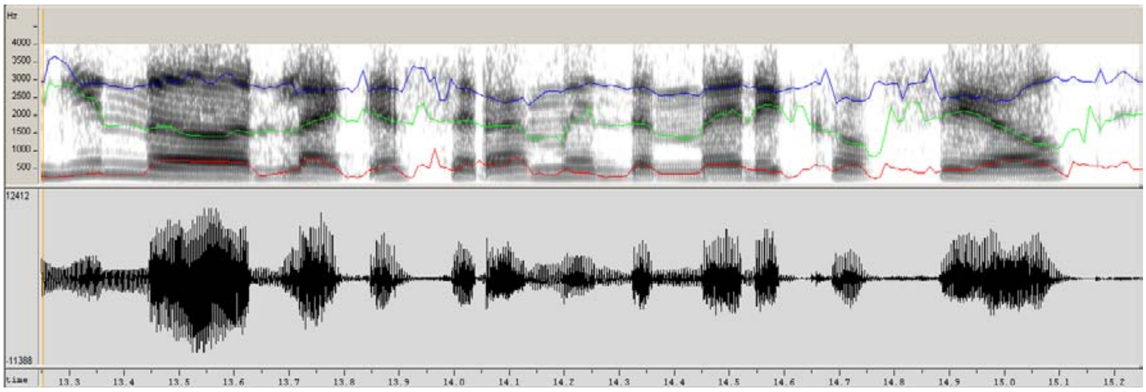


Figure 5. A speech signal (below) and its spectrogram (above) with the first three formant frequencies highlighted: *f*1 (red), *f*2 (green) and *f*3 (blue).

3.1.2 Cepstral contours

Mel Frequency Cepstral Coefficients (MFCC) [Davis and Mermelstein, 1980] are short-term spectral features that were originally defined for automatic speech recognition purposes and then adopted in automatic speaker recognition, being nowadays the standard *de facto* features in state of the art speaker recognition systems.

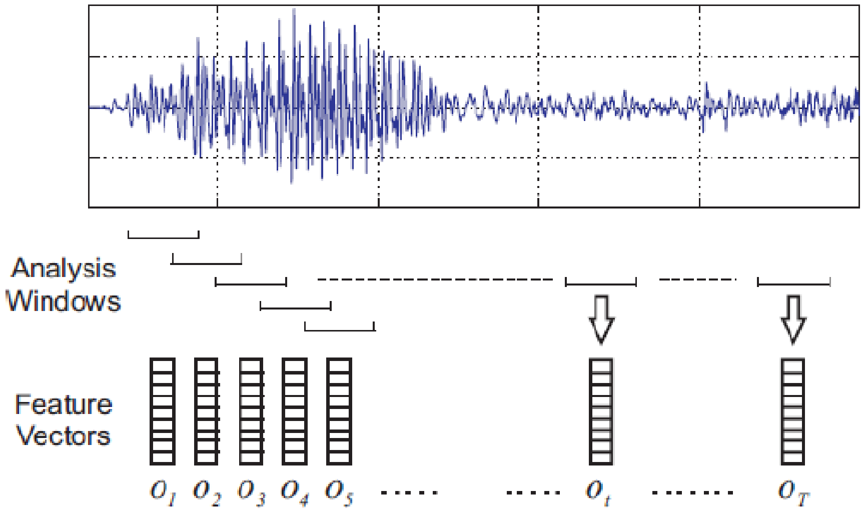

Analysis Windows
Feature Vectors
o_1 o_2 o_3 o_4 o_5 .... o_t .... o_T

Figure 6. Feature vectors extraction from overlapped analysis windows.

Speech is a signal that continuously changes due to articulary movements. In order to obtain MFCC vectors, it is firstly divided into short frames of about 20-30 ms in duration that are processed individually (it is assumed that the signal remains stationary within this interval), obtaining a final feature vector from each of these frames. Consecutive frames are slightly overlapped in order to avoid any loss of information due to the further processing. Speech frames can be pre-emphasized prior to further steps to boosts the higher frequencies. Then, a windowing function is applied because of the finite-length effects of the discrete Fourier Transform (DFT), attenuating the signal near the frame limits. There are several types of windowing function, while Hamming is the most usual. The fast Fourier transform (FFT), a fast implementation of the DFT, decomposes the signal into its frequency components. Only the magnitude spectrum is retained, having found out that its envelope contains information about the resonance properties of the vocal tract and is the most informative part of the spectrum for speaker recognition. The shape of this envelope is modeled by a set of psycho-acoustically motivated filter bank, obtaining the energy at different frequency ranges in the Mel scale [Stevens *et al.*, 1937]. Finally, logarithmic compression and Discrete Cosine Transform (DCT) are applied, retaining the lowest DCT coefficients as the MFCC feature vector.

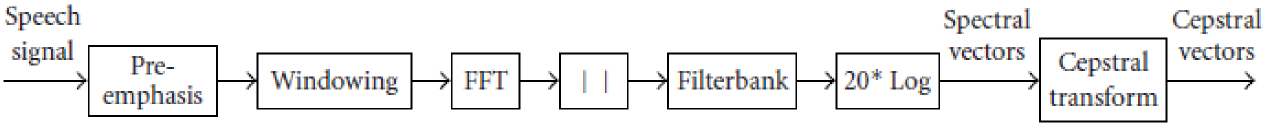


Figure 7. Modular representation of MFCC feature extraction (extracted from [Bimbot *et al.*, 2004]).

In this work, 19-coefficients cepstral vectors have been extracted from a 20 filter Mel bank, using a 20 ms Hamming window with a 50% overlap between consecutive windows (producing then one vector each 10 ms). These feature vectors have been further processed, in order to remove additive noise and other transmission channel effects, applying Cepstral Mean Normalization (CMN) [Atal, 1974; Furui, 1981], RASTA filtering [Hermanski and Morgan, 1994; Malayath *et al.*, 2000] and feature Warping [Pelecanos and Sridharan, 2001]. Cepstral contours are defined in this work as the temporal evolution of each dimension in MFCC feature vectors.

3.2 ASR region conditioning

In order to have a fully automatic speaker recognition system capable to work at a linguistic unit level, automatic speech recognition (ASR) is needed for defining both phonetic content and time interval of speech regions containing the units to be segmented. For this purpose, the phonetic transcription labels produced by the SRI's Decipher ASR system [Kajarekar *et al*, 2009] were used. For this system, trained on English data from telephonic conversations, the Word Error Rate (WER) of native and nonnative speakers on transcribed parts of the Mixer corpus, similar to NIST SRE databases used for this work, was 23.0% and 36.1% respectively.

Looking for multiple separate contributions to the speaker identity in a speech file, linguistic units are the natural and straightforward group of segments to work with. Several groups of units can be explored, showing each of them different characteristics in terms of speaker identification from their formant trajectories specificities:

- Phones: showing the biggest frequencies of occurrence among the six groups (from 20 to over 100 per conversation), they are highly dependent on their contexts. Additionally, within-phone formant and bandwidth trajectories show limited excursions (except in the case of diphthongs).
- Diphones: they show on average richer contours than phones but poorer than triphones, presenting good enough occurrence frequencies (from units to 20-30).
- Triphones: they show the richer contours on average, but their frequency of occurrence drops dramatically. Additionally, some of them have high

number of articulation targets resulting in complex contours to be modeled with a small number of parameters.

- Center phone in triphones: we extract the contours just from the central phone in a given triphone, limiting context variability. Being attractive, they share the same low frequency as triphones and the low number of articulation targets as phones.
- Syllables: they show both high frequency of appearance and rich contours, both of them desireable properties. They share some of the units with phones, diphones and triphones but as a group show less contextual variation.
- Words: only a few of them are frequent enough to perform well (function words as “but”, backchannels as "yeah", fillers like "uh", discourse markers like "so", etc.) but they can be idiosyncratic for speakers. They are also often surrounded on one or both sides by a pause, which helps reduce contextual variation.

This study focuses on phones and diphones: 39 phone units from an English lexicon plus two filled pauses (PUH, PUM) were used, represented by the Arpabet phonetic transcription code [Arpabet]. Table 1 shows the symbols used to represent all these units, and their correspondence with the International Phonetic Alphabet (IPA) for English.

Arpabet	IPA	Arpabet	IPA
AO	ɔ	CH	tʃ
AA	ɑ	JH	dʒ
IY	i	F	f
UW	u	V	v
EH	ɛ	TH	θ
IH	ɪ	DH	ð
UH	ʊ	S	s
AH	ʌ/ə	Z	z
AX	ə	SH	ʃ
AE	æ	HH	h
EY	eɪ	M	m
AY	aɪ	N	n
OW	oʊ	NG	ŋ
AW	aʊ	L	l
ER	ɝ /ɚ	R	r *or* ɹ
P	p	DX	ɾ
B	b	Y	j
T	t	W	w
D	d	PUH	-
K	k	PUM	-
G	ɡ		

Table 1. Arpabet symbols and their correspondence with those from the IPA.

Diphone units were defined by the combination of any two consecutive phone units, although only a subset of 98 diphones of all the possible combinations was used (those presenting higher frequency of occurrence).

3.3 Temporal contours coding

Once the feature vectors have been extracted and the phonetic transcriptions are available, linguistic units under analysis can be bounded in order to retain only the features belonging to them. In the case of formant trajectories, the speech segment belonging to a unit is represented by a feature matrix of 3 frequency values x #frames/unit, while in the case of MFCCs this feature matrix has 19 cepstral coefficients x #frames/unit. This variable-length segment (due to the different number of frames between units) is duration equalized to a number of frames equivalent to 250 ms (25 frames), following results in previous studies [de Castro *et al.*, 2009] [Morrison, 2009]. Then, the temporal evolution of each feature within the unit can be coded.

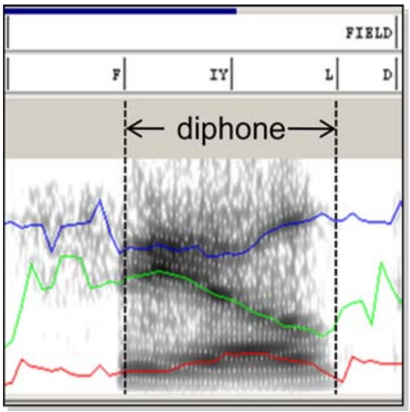


Figure 8. Temporal contour of the first three formant frequencies in a diphone unit.

As it has been mentioned in Section 2.2.2, two schemes have been mainly used for contour modeling: one of them is to fit the temporal trajectory to a polynomial curve (usually a Legendre polynomial expansion); and the other one is to model it by a Discrete Cosine Transform (DCT). Both methods capture the characteristics of the curve, like mean, slope and finer details. The grade of detail can be controlled by in-/decreasing the polynomial order or using more/less DCT coefficients, respectively.

In this work, the DCT has been used for temporal contours coding, retaining the first 5 coefficients. This 5 coefficients per temporal contour are concatenated for all the features, yielding a single final vector per unit of 15 dimensions (3 formant trajectories x 5 DCT coefficients) in the case of formant frequencies and of 95 dimensions (19 cepstral contours x 5 DCT coefficients) in the case of MFCCs. In this way, each linguistic unit can be represented by a single feature vector that will be processed by a different system.

Figure 9 summarizes the purpose of the whole feature extraction process, obtaining one feature vector from temporal contours of linguistic units of different lengths.

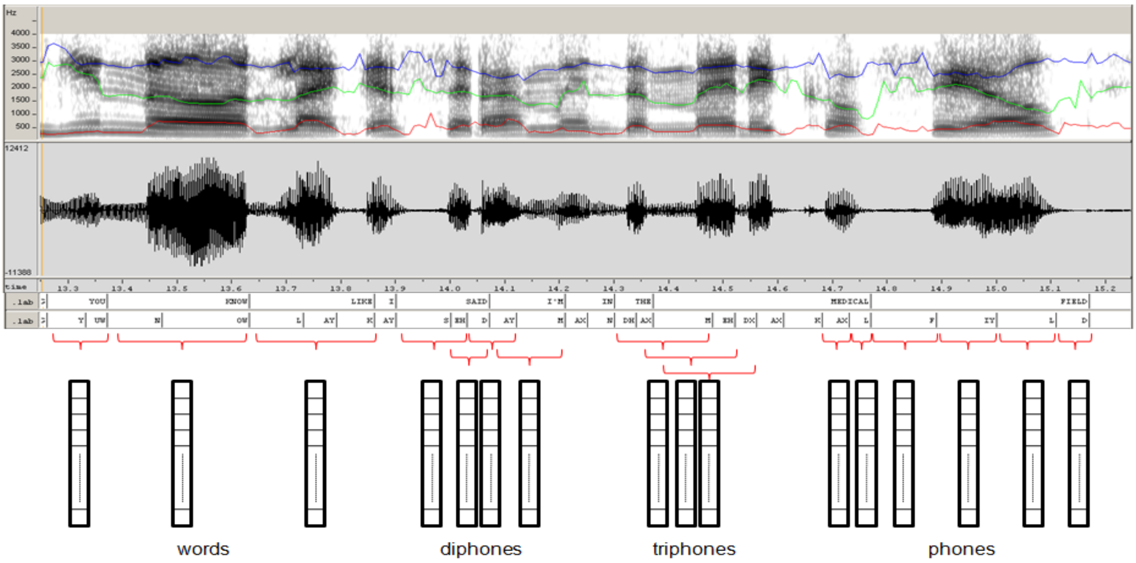


Figure 9. Constant-length feature vectors extraction from variable-length linguistic units.

3.4 Unit-dependent speaker recognition systems

Proposed systems are based on the well known GMM-UBM framework [Reynolds *et al.*, 2000], using duration-equalized DCT-coded temporal trajectories per linguistic unit as feature vectors. The GMM-UBM systems have been the state-of-the-art in the text-independent speaker recognition field for many years until the emergence of JFA [Kenny *et al.*, 2008] and total variability [Dehak *et al.*, 2011] techniques, which have outperformed the former ones through accurately modeling the existing variability in the supervector feature space. For this work, GMM-UBM systems have been chosen for two main reasons: i) as we are using a new type of features, we need first to find the optimal configuration for this GMM-UBM new framework, which is the basis of supervector-based systems; and ii) because we aim to model speakers in a unit-dependent way, a much smaller amount of data is available for training purposes, so probably not enough data would be available to capture the existing variability in each unit domain (also having into account that we only have ASR labels from the 2004, 2005 and 2006 NIST SREs).

Three different unit-dependent GMM-UBM configurations were tested for phone units previously to perform experiments reported in this work:

1. UBM and speaker models trained on unit-independent data; evaluation trials performed on unit-dependent test data (as we did in our first approach [Franco-Pedroso *et al.*, 2012]).
2. UBM trained on unit-independent data; speaker models adapted from unit dependent training data; evaluation trials performed on unit-dependent test data.
3. UBM and speaker models trained on unit-dependent data; evaluation trials performed on unit-dependent test data (fully unit-dependent).

For each configuration, different numbers of mixtures were tested, ranging from 2 up to 1024 mixtures increasing in powers of 2. It was found out that the best results were obtained for the fully unit-dependent configuration, so this is the configuration used to obtain the individual linguistic unit results reported in this work: UBM and maximum-a-posteriori (MAP) adapted speaker models were trained and tested on unit-dependent data (using every unit segment available in both training and testing utterances), yielding an independent GMM-UBM system for each linguistic unit. This procedure yields N scores per trial (N = #units) which can be used either as individual speaker recognition systems or, additionally, combined in a single fused system. None of these individual systems include any type of score normalization. Figure 10 shows a diagram block of the GMM-UBM system for a particular phone unit.

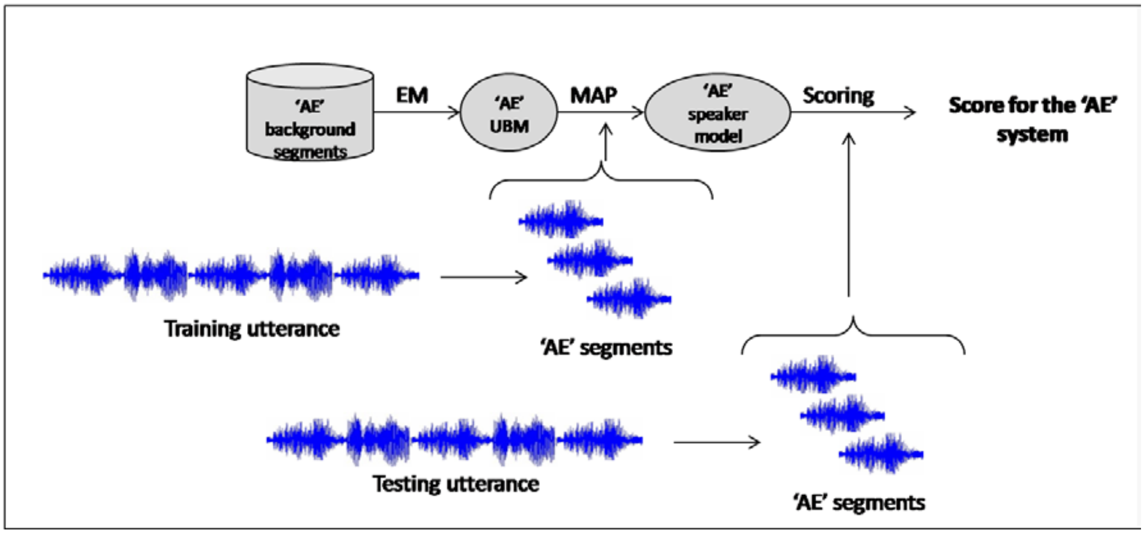


Figure 10. GMM-UBM system for the 'AE' phone unit.

In the case of formant trajectories, GMMs of 32 components were used for phone units and 16 components in the case of diphone units, while in the case of cepstral contours, 8 components were sued for phone units and 4 components for diphone units.

3.5 System combination methods

Both individual unit performance and different unit combinations have been analyzed in this work. Individual linguistic-unit systems allow us to report informative likelihood ratios for very short speech samples, as it is the case of forensic applications where a speech expert, typically a linguist/phonetician, can isolate or mark segments of compatible/comparable sounds between speech samples (typically, several segments belonging to some linguistic unit). Aditionally, when different types of information can be used, individual units are combined to achieve better discriminative capabilities.

Individual systems have been combined in both intra- (different phones between them and different diphones between them) and inter-unit (pooling phones and diphones together) manners. Two different fusion techniques were used: sum fusion and logistic regression fusion. The former one was performed after linear logistic regression calibration per unit, while the latter one was performed in a single calibration/fusion step.

Another issue is what units should be selected for fusion. Two strategies have been used in this work. The first of them is to select the n-best performing units by setting a threshold for the Equal Error Rate (EER) of the units to be fused, leaving out those performing worse. However, this procedure does not guarantee that the best fused system will be achieved because some units with lower performance by itself could contribute to the fused system if its LR's are sufficiently low correlated with those produced by the other units to be fused. On the other hand, testing all of the possible combinations would be an exhaustive task, so we performed a Sequential Forward Selection (SFS) with the EER as the evaluation criterion (similar to that used in [de Castro *et al.*, 2009]) based on the following steps:

1) Take the best performing unit in terms of EER as the initial units set.
2) Take the next best performing unit and fuse with the previous set. If the fusion improves the performance of the previous set, this unit is added to the units set, otherwise rejected.
3) The previous step is repeated for all the units in increasing EER order.

This procedure allows us to find complementarities between units that otherwise would not have been revealed, but avoiding the complex task of testing each possible combination.

Chapter 4: Experimental framework

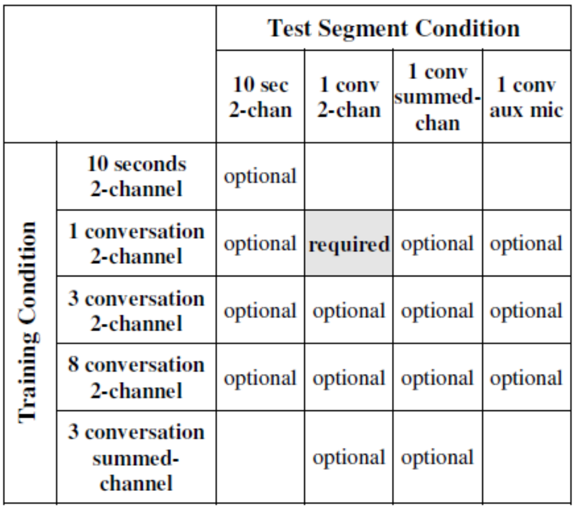
THIS CHAPTER PRESENTS, foremost, the databases and protocols used in order to test the proposed approach. The goodness of the results obtained for a particular technique or a specific technology depends largely on the processes followed in order to test it. As it has been mentioned, most of the related works in the forensic field have been tested for just some few linguistic units using databases of at most ten or twenty speakers, leading to just some hundred of voice comparisons, and using clean speech in most of the cases. This way it is hard to extrapolate these results to real life applications where hundreds of speakers can be presented to the system and different acoustic conditions may affect speech recordings.

Later, different performance metrics are clearly defined for specific purposes depending on the aspect of the system in which we are interested (discriminating power or calibration properties).

4.1 NIST SRE databases and protocols

The US National Institute of Standards and Technology (NIST) has been conducting Speaker Recognition Evaluations (SRE) from 1997 in order to measure the state-of-the-art and to find the most promising algorithmic approaches in text-independent speaker recognition. These evaluations define datasets and protocols to measure system performance in an objective way, so that the results obtained by systems based on very different technologies can be compared in a common framework.

For each SRE, NIST provides an evaluation dataset consisting of two subsets: a training dataset containing excerpts of target speakers to be modeled, and a test dataset containing test segments from unknown individuals to be compared with target speaker models. Several conditions in duration (10 sec., 1 conversation about 5 min., 3 conversations, etc.) and audio types (separated sides of a telephonic conversation, both sides mixed on a single channel, microphonic recordings) for training and test segments are established, so that different combinations between them define different tasks to be faced. As an example, Table 2 shows the different tasks proposed in NIST SRE 2006.



		Test Segment Condition			
		10 sec 2-chan	1 conv 2-chan	1 conv summed-chan	1 conv aux mic
Training Condition	10 seconds 2-channel	optional			
	1 conversation 2-channel	optional	**required**	optional	optional
	3 conversation 2-channel	optional	optional	optional	optional
	8 conversation 2-channel	optional	optional	optional	optional
	3 conversation summed-channel		optional	optional	

Table 2. Matrix of training and test segment conditions. The shaded entry is the required core test condition (extracted from [The NIST Year 2006 Speaker Recognition Evaluation Plan]).

For this work, only the required task (1 conversation 2-channel vs 1 conversation 2-channel) has been evaluated, in which each trial consists of a 2-channel 5 minutes telephonic conversation (of approximately 2.5 minutes of net speech per conversation side) for training and another 2-channel conversation for testing, being identified the target channel. Telephonic conversations provided involve several transmission channels (landline, GSM, CDMA) and different handsets (carbon-button, electrect, head-mounted, cordless, etc.), what makes it a very challenging task due to the high acoustic variability.

Each trial must be independently judged as “true” (the model speaker speaks in the test segment) or “false” (the model speaker does not speak in the test segment) in order to compute the detection cost, C_{Det}, and a confidence score given reflecting the system’s estimate of the probability that the test segment contains speech from the target speaker, in order to produce DET curves. While it is not mandatory, providing estimated LR values as scores is suggested.

Among all SREs conducted so far, only datasets from 2004, 2005 and 2006 has been used due to the need of having phonetic transcriptions for linguistic units labeling – as we haven’t a robust ASR system, phonetic transcriptions provided by SRI International [SRI] have been used, and only those corresponding to these years were available. Furthermore, only data from English speakers were used because linguistic units are defined on an English lexicon, as it has been shown in Section 3.2. Also, it should be mentioned that our approach has been tested only on male speakers, due to the high number of systems that are involved (one per linguistic unit).

4.2 Datasets partitioning

Dataset partitioning is a main issue when designing any pattern recognition system. The performance of a classifier depends on how datasets are managed. The greater the amount (and variety) of data for testing the system, the more reliable will be the results. Also, the greater the amount (and variety) of data for training the system, the more robust it will be, given that a greater number of different conditions can be modeled. Moreover, training and testing datasets should be disjoint sets in order to avoid overfitting when training the parameters of the system. However, datasets are limited, so the amount dedicated to each of the two purposes must be balanced. In order to maximize the amount of tests performed while keeping a sufficient amount of data for training the model parameters, some techniques such as cross-validation [] are often used, preserving different parts of the data for optimization and evaluation based on rotated subsets.

As previously stated, the NIST SRE datasets available for this work are limited to those from 2004, 2005 and 2006 years. Taking into account the necessity of large amounts of data for UBM training, the larger partition has been used for training, consisting of SRE04 and SRE05 datasets, keeping the SRE06 dataset for testing purposes. The specific datasets (with details of their composition) and the purposes to which are devoted are listed below.

- **Background dataset: SRE04 and SRE05**
    - Content: 367 male English speakers / 1,808 conversations
    - Purpose: training UBMs for unit-dependent systems
- **Development dataset: SRE05**
    - Content: 243 male English speakers / 11,272 trials
    - Purposes:
        - optimizing the number of mixtures for the GMM-UBM systems
        - training the calibration and fusion rules
- **Testing dataset: SRE06**
    - Content: 219 male English speakers / 9,720 trials
    - Purpose: evaluate the performance of the system

4.3 Performance evaluation

4.3.1 Score-based metrics

Usually, the output of an automatic speaker recognition system is a similarity measure between the controlled and questioned samples, usually called *score*. In order to measure how good a technique is, discriminating power of a set of scores has been used for many years as the main performance measure for automatic speaker recognition systems [Przybocki *et al.*, 2007][van Leeuwen *et al.*, 2006], being this measure associated with correctly discriminating same-source and different-source trials.

Trials can be classified as target (when same-source samples are involved) or non-target (when different-source samples are involved). Once a threshold is set within the range of scores, each trial is either considered by the system as being target (if the score is above the threshold) or non-target (if the score is below the threshold); using verification terms, the user is either *accepted* or *rejected* by the system. As automatic systems are not perfect, this leads to two types of errors:

- False Rejection (FR) error: when a trial considered as non-target actually was a same-source comparison.
- False Acceptance (FA) error: when a trial considered as target actually was a different-source comparison.

When measured for a set of scores, these errors are normalized by the total number of trials involved in the experiment, giving the False Rejection error Rate (FRR) and False Acceptance error Rate (FAR) as percentages. If the threshold is changed, the tradeoff between FRR and FAR is then changed as well (also called the *operating point* of the system). Detection Error Tradeoff (DET) curves [Martin *et al.,* 1997] graph the performance for all the possible operating points, plotting the FAR as a function of the FRR on a logarithmic scale. The Equal Error Rate (EER) is the operating point in which FRR = FAR, and it is usually taken as a global measure of the discriminating power of the system.

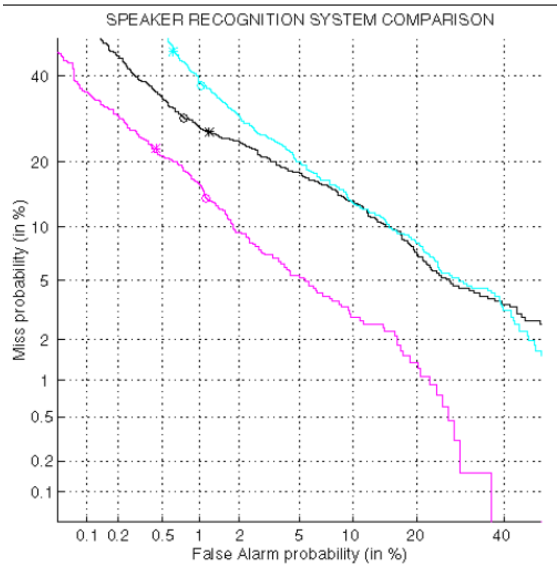


Figure 11. Example of DET curves (extracted from [Martin *et al.,* 1997]). The closer to the origin of coordinates is the curve, the better the system.

Another performance metric also used along this Thesis is the detection cost proposed by the NIST in order to measure the performance of speaker recognition systems participating in their Speaker Recognition Evaluations. This metric sets a fixed cost to FA and FR errors as well as a priori probability for target and non-target individuals. This metric, also known as detection cost function (DCF) is defined as

$$C_{Det} = C_{FR} \cdot P_{FR|Target} \cdot P_{Target} + C_{FA} \cdot P_{FA|NonTarget} \cdot (1 - P_{Target})$$

where C_{FR} and C_{FA} are the relative costs of FR and FA detection errors respectively; $P_{FR|Target}$ (the probability of false reject given a target speaker) is the FRR of the system; $P_{FA|NonTarget}$ (the probability of false acceptance given a non-target speaker) is the FAR; and P_{Target} the *a priori* probability of having target trial.

In the NIST SREs testing protocols used in this work (2005 and 2006), the costs and the target probability are defined as:

- $C_{FR} = 10$
- $C_{FA} = 1$
- $P_{Target} = 0.01$

Inasmuch as C_{Det} can have different values depending on the operating point set for the system (tradeoff between FAR and FRR), the minimum value of this metric for any operating point (minDCF) is used in this work as a global measure of the discriminating power of the system

4.3.2 Likelihood ratio-based metrics and calibration issues

As we have seen in the previous section, it depends on the threshold that a trial is considered as target or non-target based on the score. That is, the score is meaningless itself as the threshold can be changed depending on the application. For forensic purposes, however, a meaningful output is needed to be obtained from the speaker verification system that assists in the decision making process. Moreover, setting a threshold leads to make hard decisions, while reporting forensic evidence should comply with the requisites of modern forensic science [Gonzalez-Rodriguez *et al.*, 2007]. For this purpose, likelihood ratios are used within this Thesis, as our aim is to apply automatic speaker recognition systems for forensic purposes.

In contrast to a score, a likelihood ratio (LR) has a meaning itself, defined as the ratio between the probabilities of observing the evidence under the prosecution and the defense hypothesis:

$$LR = \frac{P(E|\theta_p, I)}{P(E|\theta_d, I)}$$

where:

- E is the evidence, and includes a recovered sample for an unknown origin and a control sample whose origin is known.
- θ_p is the prosecution hypothesis: the recovered sample comes from the suspect.
- θ_d is the defense hypothesis: the recovered sample does not come from the suspect.
- and I is refers to other information relevant for the case.

So, a LR greater than 1 points towards the prosecution hypothesis, while a LR smaller than 1 points towards the defense hypothesis. It is important to see that in any case, no hard decision is taken because the LR is interpreted as the odds given the evidence. Then, taking also into account other prior information coming from other different evidences it is possible to compute the posteriors odds:

$$\frac{P(\theta_p|E, I)}{P(\theta_d|E, I)} = \frac{P(E|\theta_p, I)}{P(E|\theta_d, I)} \cdot \frac{P(\theta_p|I)}{P(\theta_d|I)} = LR \cdot \frac{P(\theta_p|I)}{P(\theta_d|I)}$$

This way, the role of the scientific expert is limited to compute and report the LR, which complements other information of the case in order to make a final decision by the judge or the jury.

In order to measure the accuracy of an automatic system that outputs LRs, the log-likelihood ratio cost (C_{llr}) was defined in [Brummer and du Preez, 2006]:

$$C_{llr} = \frac{1}{2 \cdot N_p} \sum_{i=1}^{N_p} \log_2(1 + \frac{1}{LR_i}) + \frac{1}{2 \cdot N_d} \sum_{j=1}^{N_d} \log_2(1 + LR_j)$$

where N_p and N_d are, respectively, the number target and non-target trials in the set. This metric is application independent in the sense that C_{llr} is the expected cost for any value of C_{FR} and C_{FA} averaged over a set of LR values, assuming $P_{Target} = P_{NonTarget} = 0.5$. If the system is accurate, it should have a C_{llr} value below 1, ideally a lot below 1.

The process of converting scores to likelihood ratios is referred as calibration and it is a key task in the application of automatic speaker recognition systems to forensic scenarios. Among the different proposed methods to calibrate systems, a widely adopted is a linear transformation of scores as performed in [Brummer and du Preez, 2006] via logistic regression, being the one that has been used in this Thesis. This transformation is trained from a set of background scores to minimize the C_{llr} with the constraint of preserving the discriminating power of the scores set.

C_{llr} can be decomposed on two terms: one due to the discriminating power of the system itself ($minC_{llr}$), and one due to the calibration process that represents to what extent the calibrated scores can be interpreted as LRs. So, the difference between C_{llr} and $minC_{llr}$ is known as *calibration loss*

Another way of analyze the calibration properties of a set of LRs (and other properties of the system) commonly used in forensic voice comparison is to represent the cumulative distribution of LRs from same-source comparisons and different-source comparisons in what is called a Tippet plot, showing the proportion of LRs observed from same-source or different-source comparisons equal to or bigger than a given LR value.

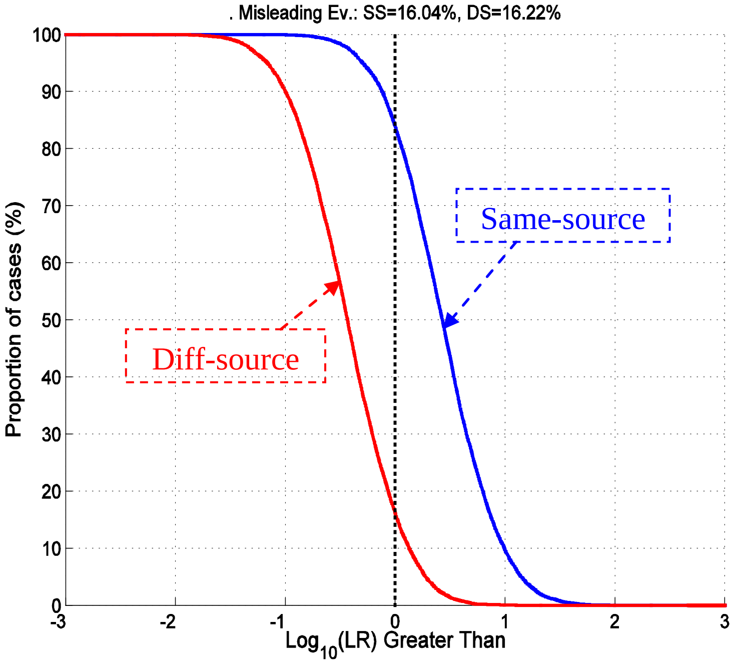


Figure 12. Example of a Tippet plot.

For same-source comparisons, a well calibrated system should have a high proportion (ideally 100%) of cases in which the log-LR is greater than 0. Similarly, for different-source comparisons it should have a very low proportion (ideally 0%) cases in which the log-LR is greater than 0. On the other hand, a well calibrated system with not too high discriminating power (small difference between same-source and different-source curves for log-LR greater than 0) shouldn't have neither a high proportion of same-source comparisons in which the log-LR is very high, nor a high proportion of different-source comparison in which the log-LR is very low. This is because a system that hasn't high discriminating power should not provide so very confident LRs.

Chapter 5: Results

THIS CHAPTER PRESENTS the results achieved in our experimental framework by both reference and proposed systems. Proposed approach is firstly analyzed in a unit-dependent way in terms of discriminating power and calibration properties. Then, different ways of combining information from linguistic units are presented, involving several fusion techniques and unit selection schemes. Finally, higher level systems based on formant trajectories are combined with state-of-the-art MFCC-based systems.

5.1 Reference systems

In order to have a baseline to compare with, two non-linguistic reference systems have been tested on the same task (1side-1side, English-only male speakers) of the 2006 NIST Speaker Recognition Evaluation used to test our proposed approach.

One of them is a 1024-mixture GMM-UBM system trained using the same data partitioning used for the unit-dependent systems. This system is based on static (as opposed to our temporal contours) MFCC features (19 coefficients plus first derivatives). The purpose of using this system is to compare our unit-dependent systems with a system using the same technology but based on classical acoustic features.

The other one is an i-vector system, also based on static MFCC features, build from the same UBM used in the GMM-UBM system. Both total variability (400 dimensions) and LDA (200 dimensions) matrices, as well as the Within Class Covariance Normalization [Dehak *et al.*, 2011], were trained on SRE04 and a half of SRE05, leaving apart the other half in order to obtain scores for calibration purposes and to train the logistic regression fusion. The purpose of using this system is to compare our unit-dependent systems with a state-of-the-art system.

Unlike our unit-dependent systems, these reference systems make use of the whole speech samples provided for both training and testing, which are about 150 seconds long in average (net speech). None score normalization technique has been applied to those systems, since it hasn't been applied to our unit-dependent systems.

Table 3 shows the results for both systems in our test set.

System	EER (%)	minDCF
GMM-UBM, MFCC-based	10.26	0.0457
i-vector, MFCC-based	8.86	0.0407

Table 3. EER (%) and minDCF for MFCC-based systems in the NIST SRE 2006 English-only male 1side-1side task.

5.2 Unit-dependent systems

Due to the large number of units processed by our unit-dependent systems (41 phone units plus 98 diphone units), only those of them performing best are shown in this section. If the reader wants to see the results for a particular unit not shown here, they can be seen in the Appendix.

5.2.1 Phones

Table 4 shows the individual performance of the ten best performing phone units in terms of EER (ordered by performance) for the NIST SRE 2006 English-only male 1side-1side task.

In the case of formant trajectories, although the performance of these phone-dependent systems is far from that of our reference systems, it is actually a remarkable result taking into account the amount of speech used by each system (no more than 6.7 seconds per utterance in average for any of the units; see the Appendix) and the type of features used (high-level features). Moreover, all of them have good calibration properties: small calibration loss (difference between C_{llr} and minC_{llr} metrics) and C_{llr}'s below 1. This allows us to obtain informative calibrated likelihood ratios from very short speech samples (just the speech segments belonging to that unit present in the utterance), as we can see in the Tippett plot in Figure 13 for the best performing phone unit ('AY'), where there is a very small proportion of misleading LRs smaller/greater than 10. Also, given that the discriminating power of the system is limited, correct LRs are restricted as well, being most of them smaller than 100. So, this kind of systems can provide useful information about the strength of the evidence in forensic applications using similar procedures to those followed by forensic phoneticians.

Furthermore, it has to be taken into account that the feature extraction process is affected by errors both in the formant tracking and in the ASR system (time alignment and phone decoding errors). These errors can be avoided in a real forensic scenario, where a speech expert can mark or segment manually the linguistic units to be analyzed.

Formant trajectories					Cepstral contours				
Phone unit	EER (%)	minDCF	C_{llr}	minC_{llr}	Phone unit	EER (%)	minDCF	C_{llr}	minC_{llr}
AY	21.67	0.0907	0.6949	0.6593	N	15.92	0.0713	0.5520	0.5082
L	23.74	0.0966	0.7490	0.7173	AE	18.98	0.0813	0.6087	0.5832
AE	24.92	0.0922	0.7466	0.7161	AY	21.68	0.0869	0.6822	0.6428
R	25.47	0.0957	0.7672	0.7430	M	22.28	0.0857	0.6824	0.6583
Y	26.03	0.0948	0.7916	0.7615	IY	23.32	0.0923	0.7453	0.7002
N	26.27	0.0942	0.7790	0.7554	Y	24.00	0.0906	0.7313	0.7062
AX	26.54	0.0990	0.8080	0.7750	PUH	24.18	0.0908	0.7359	0.7149
PUH	27.36	0.0931	0.7925	0.7689	R	24.65	0.0887	0.7295	0.7116
OW	27.78	0.0944	0.8088	0.7898	OW	24.65	0.0987	0.7917	0.7396
IH	28.92	0.0990	0.8172	0.7889	UW	24.79	0.0898	0.7391	0.7198

Table 4. Performance metrics for the 10 best performing phones of both formant trajectories and cepstral contours based systems in the NIST SRE 2006 English-only male 1side-1side task.

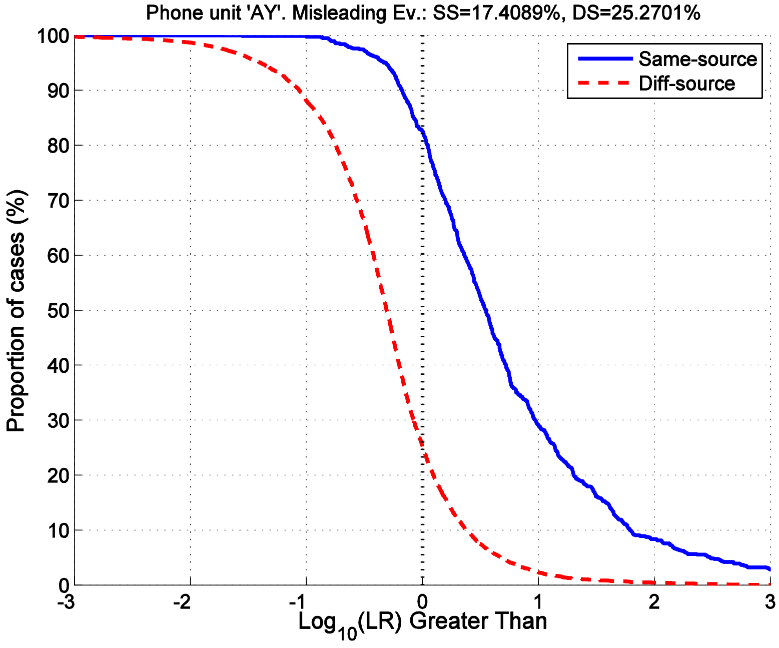


Figure 13. Tippett plot for the best performing phone unit ('AY') for formant trajectories in the NIST SRE 2006 English-only male 1side-1side task.

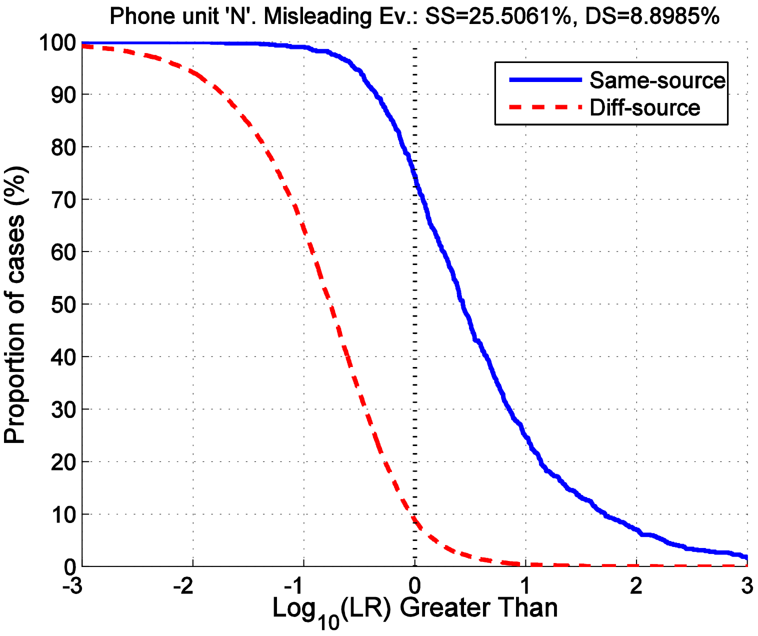


Figure 14. Tippett plot for the best performing phone unit ('AY') for cepstral contours in the NIST SRE 2006 English-only male 1side-1side task.

In the case of cepstral contours, much better discriminating power can be obtained for the unit-dependent systems. For the phone unit 'N', an EER as low as 15.92% is obtained from just 6.5 seconds of speech per utterance in average (see the Appendix). Also, calibration properties are as good as in the case of formant trajectories, as it can be seen in Tippet plot in Figure 14 for the best performing unit 'N'.

It is interesting to see the units that best perform for formant trajectories and for cepstral units. Although both types of features are somehow related as cepstral features come from frequency measurements, the former ones are subjected to further processing (Mel filtering, logarithm, DCT) making them completely different. However, it can be seen in Table 1 that some of the best performing units are shared between both types of features ('AY', 'AE', 'N', 'Y', 'R', 'PUH' and 'OW'), so it could be supposed that the discrimination capabilities rely on the temporal dynamics constrained to the linguistic information. However, there is also a close relationship between the average amount of speech per unit available in an utterance (see the Appendix) and the performance of that unit-dependent system inasmuch as we are using a generative modeling approach, so further analysis on this assumption should be done.

It is also worth noting that for both types of features one of the best performing units is not a phone in fact, but the filled pause 'PUH' (corresponding to the sustained sound 'UH' but not in the context of a word).

5.2.2 Diphones

Table 5 shows the individual performance of the ten best performing diphone units in terms of EER (ordered by performance) for the NIST SRE 2006 English-only male 1side-1side task.

As it can be seen, diphone units have in average much lower performance than phone units for both types of features. This is a consequence of the feature extraction process and the generative modeling technique used; a particular two phone combination (diphone unit) has a fewer number of tokens in a speech sample than its constituent phones, and because we are coding each linguistic unit in a single feature vector, much less feature vectors are available to train the GMM of that unit. However, some diphones reach better performance that some phone units for both types of features, and the calibration properties are still good enough to provide useful information for forensic purposes: although C_{llr}'s are greater due to the discriminating power loss, the calibration loss is still very small. As for the phone units, unit-dependent systems based on cepstral contours perform better than those based on formant trajectories in the case of diphone units.

Formant trajectories					Cepstral contours				
Diphone unit	EER (%)	minDCF	C_{llr}	minC_{llr}	Diphone unit	EER (%)	minDCF	C_{llr}	minC_{llr}
Y-AE	29.65	0.0964	0.8269	0.8043	AX-N	23.84	0.0899	0.7583	0.7097
Y-UW	29.78	0.0993	0.8439	0.8240	N-D	24.92	0.0876	0.7563	0.7037
L-AY	30.46	0.0969	0.8343	0.8089	Y-UW	27.18	0.0960	0.8223	0.7812
DH-AE	31.13	0.0980	0.8668	0.8413	L-AY	29.11	0.0972	0.8156	0.7955
AX-N	31.54	0.0992	0.8760	0.8528	Y-AE	29.78	0.0976	0.8383	0.8094
UW-N	31.67	0.0957	0.8634	0.8421	AE-N	30.72	0.0993	0.8479	0.8230
N-OW	32.92	0.0996	0.8738	0.8594	N-OW	30.86	0.0995	0.8455	0.8185
AE-N	34.86	0.1000	0.9024	0.8767	AE-T	31.89	0.0969	0.8720	0.8526
N-D	35.05	0.0995	0.9065	0.8884	UW-N	32.20	0.0953	0.8417	0.8188
L-IY	35.58	0.0995	0.9002	0.8822	AY-K	32.45	0.0970	0.8494	0.8356

Table 5. Performance metrics for the 10 best performing diphones of both formant trajectories and cepstral contours based systems in the NIST SRE 2006 English-only male 1side-1side task.

Again, it is interesting to see how most of the best performing units are shared between both types of features ('Y-AE', 'Y-UW', 'AX-N', 'N-D', 'L-AY', 'AE-N', 'N-OW' and 'UW-N'). In this case, the relatioship between average amount of speech per utterance and performance of the system is stronger than in the case of phone units due to the much smaller amount of training and testing vectors, as previously explained.

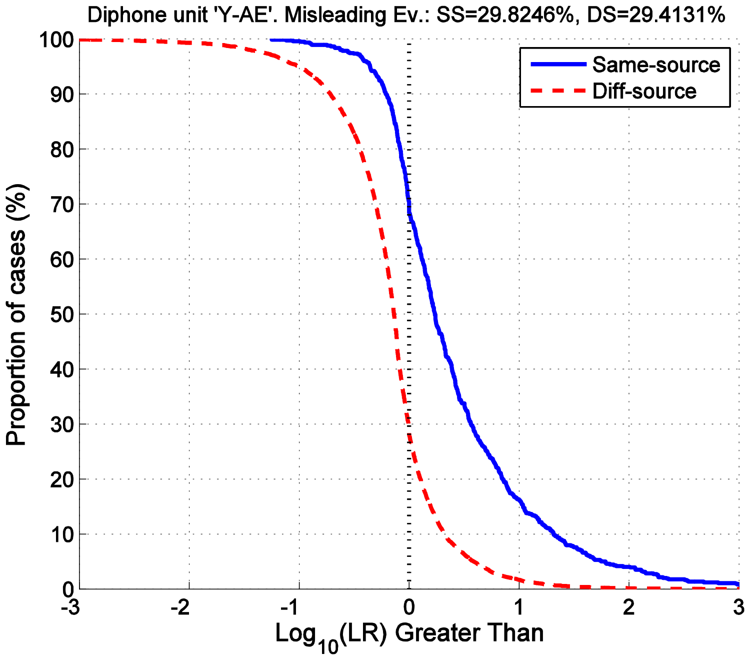


Figure 15. Tippett plot for the best performing diphone unit ('Y-AE') for formant trajectories in the NIST SRE 2006 English-only male 1side-1side task.

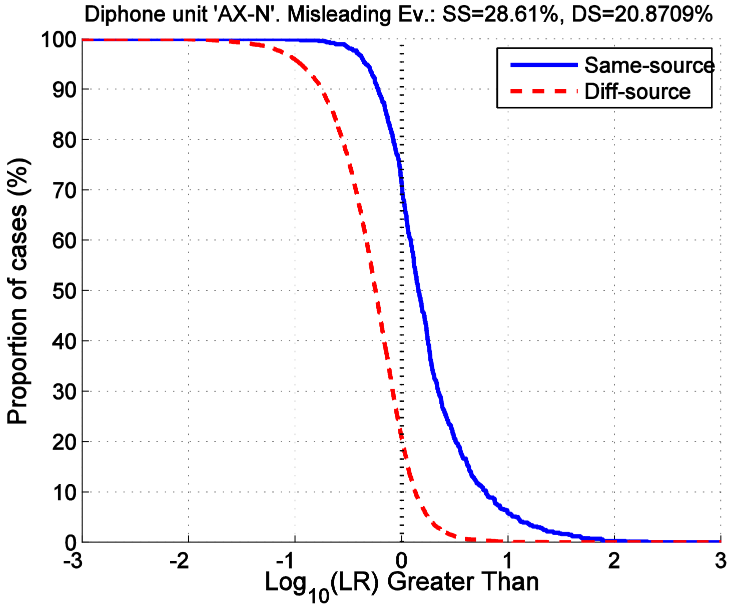


Figure 16. Tippett plot for the best performing diphone unit ('AX-N') for cepstral contours in the NIST SRE 2006 English-only male 1side-1side task.

5.3 Intra-unit fusions

Now we have seen the performance of individual linguistic units, the issue of combining different pieces of information scattered among different units can be addressed. In this section, combinations between units of the same type (phones or diphones between them) are analyzed.

5.3.1 Phones

Figure 17 shows, for the systems based on formant trajectories, the EER of the fused system as a function of the number of phones combined by means of the sum rule and the logistic regression techniques using the two types of unit selection schemes defined in Section 3.5. Solid-line curves represent fusion results for different thresholds set for the EER, while circles represent the result for the unit selection algorithm. The performance of the GMM-UBM MFCC-based reference system is also shown as a red dashed line. For the case of fusing phones performing better than a certain threshold, it can be seen that, for both type of fusion techniques, the EER of the fused system converge for a number of fused phones greater than 12, being this EER lower for the logistic regression technique (13%) than for the sum fusion rule (14%). In both cases, the performance of the fused system is greatly improved with respect to any of the individual phone systems, and quite close to that of the reference system (10.62%) using a much smaller amount of speech data (about 10% of the whole utterance for the case of fusing 12 phones). Moreover, it is worth noting that the unit selection algorithm used can achieve better fusion results (12.23%) than simply setting a threshold for the EER of the units to be fused in the case of the sum fusion rule.

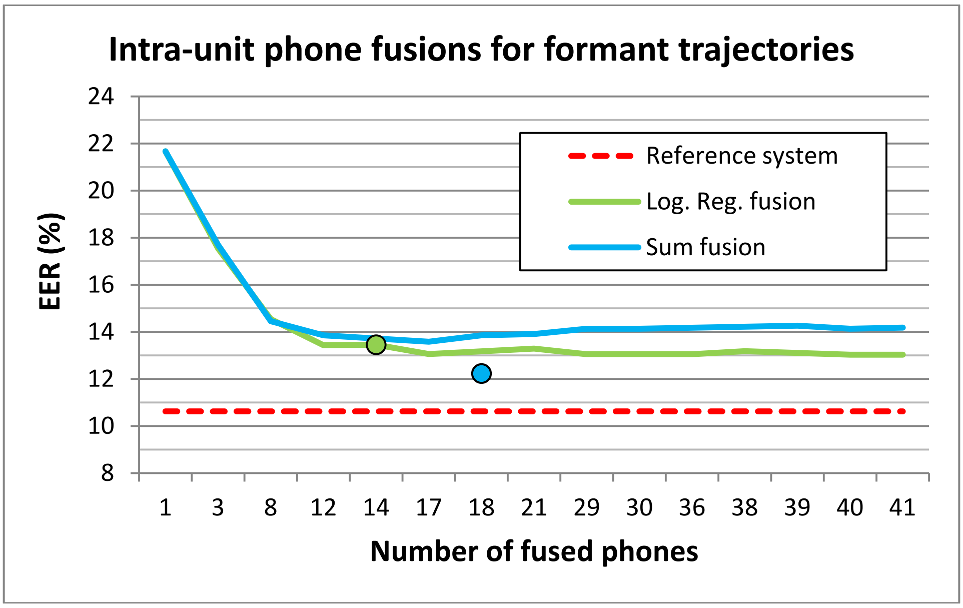


Figure 17. Intra-unit phone fusions for formant trajectories.

Best intra-unit phone fusion	EER (%)	# fused units
Sequential Forward Selection, sum fusion	12.29	18

Table 6. Best intra-unit phone fusion for formant trajectories.

Figure 18 shows the same results for the systems based on cepstral contours. It can be seen that also in this case the EER of the fused system tends to converge for a number of fused phones greater than 10. As happened for the individual unit systems, fused systems are also better when they are based on cepstral contours. It is very remarkable that the performance of the reference system can be achieved fusing just 4 phones, and a great performance over that can be achieved adding 6 more phones. Again, the sum fusion of the units selected by the Sequential Forward Selection algorithm achieves the best result, outperforming even the results of the i-vector MFCC-based system on this task.

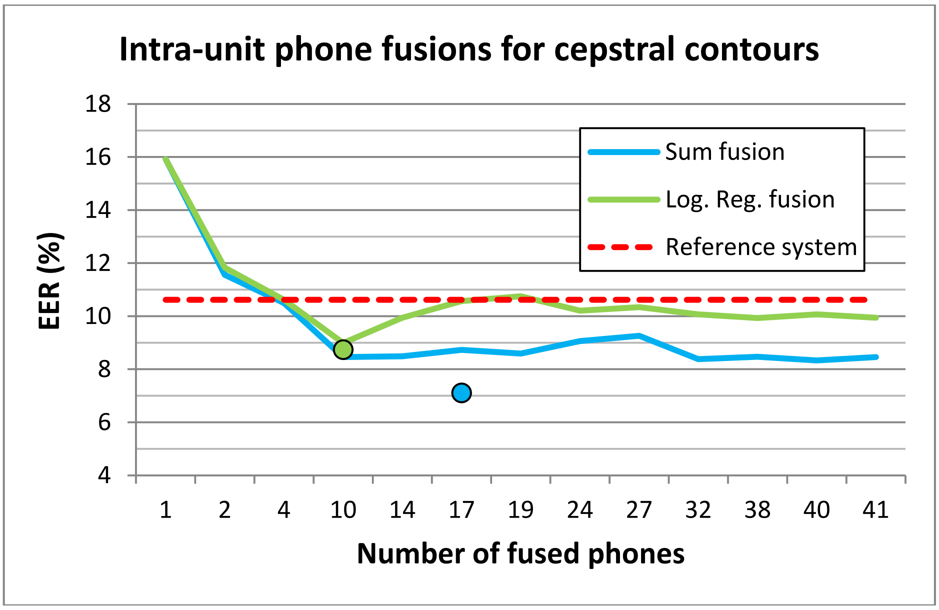


Figure 18. Intra-unit phone fusion for cepstral contours.

Best intra-unit phone fusion	EER (%)	# fused units
Sequential Forward Selection, sum fusion	7.11	17

Figure 19. Best intra-unit phone fusion for cepstral contours.

5.3.2 Diphones

Figure 19 shows, for the systems based on formant trajectories, the results of the same experiments presented in previous section but carried out on diphone units. In this case, the EER of the fused system converges for a higher number of fused units, and this EER is higher for logistic regression (17%) than for the sum rule (14.5%). Again, the unit selection algorithm achieves the better result for the sum fusion rule (13.7%).

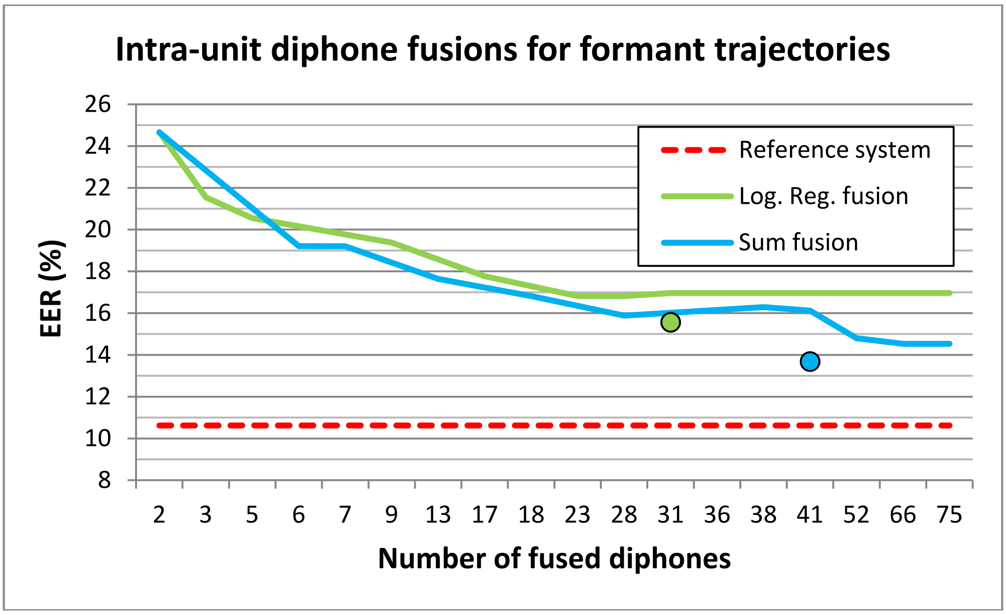


Figure 20. Intra-unit diphone fusion for formant trajectories.

Best intra-unit diphone fusion	EER (%)	# fused units
Sequential Forward Selection, sum fusion	13.68	41

Table 7. Best intra-unit diphone fusion for formant trajectories.

In the case of cepstral contours (Figure 21), very similar results are obtained by both the sum rule and the logistic regression fusions, achieving a performance better than that of the reference GMM-UBM MFCC-based system by combining less than 20 diphone units. Again for cepstral contours, the best fusion outperforms even the results of the i-vector MFCC-based system.

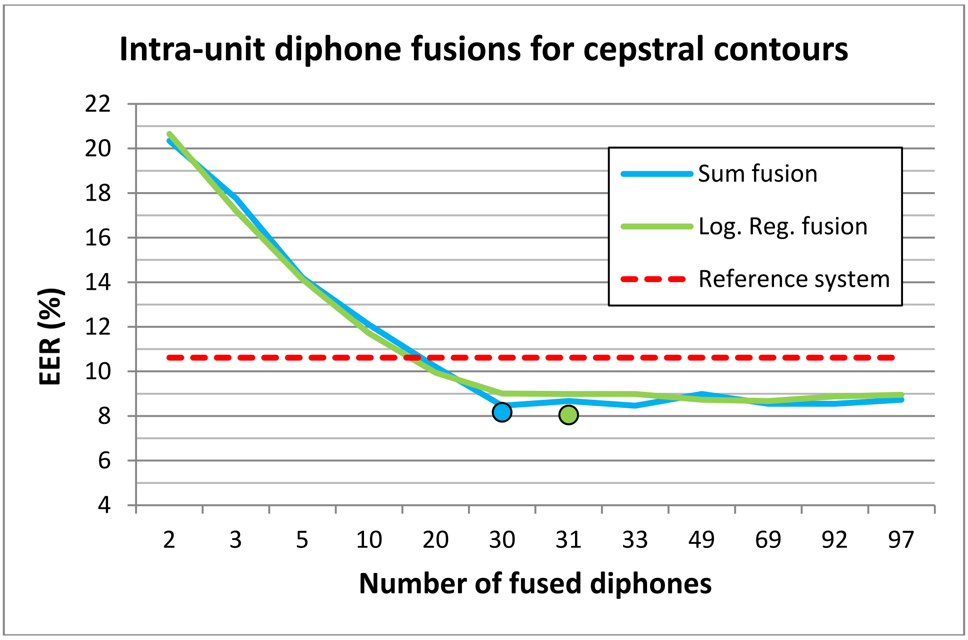


Figure 21. Intra-unit diphone fusions for cepstral contours.

Best intra-unit diphone fusion	EER (%)	# fused units
Sequential Forward Selection, log. reg. fusion	8.05	31

Table 8. Best intra-unit diphone fusion for cepstral contours.

5.4 Inter-unit fusions

In the previous section we have seen how well combine different units from each type (i.e., different phones between them and different diphones between them), but it is also interesting to see how can be combined units from different types between them. For this purpose, the same fusion techniques and combination schemes have been used putting together both phones and diphones, yielding results show in Figure 22 for the systems based on formant trajectories.

It can be seen that better results can be achieved by combining phones and diphones units than working in an intra-unit manner, taking advantage of different linguistic levels. In the case of formant trajectories, it is possible to achieve a 11.97% EER for the logistic regression fusion technique combining a high number of linguistic units (90). For the sum fusion rule, although the EER converges to a higher value, the unit selection algorithm can achieve again a better result (12.18%) with a reduced number of fused units (17).

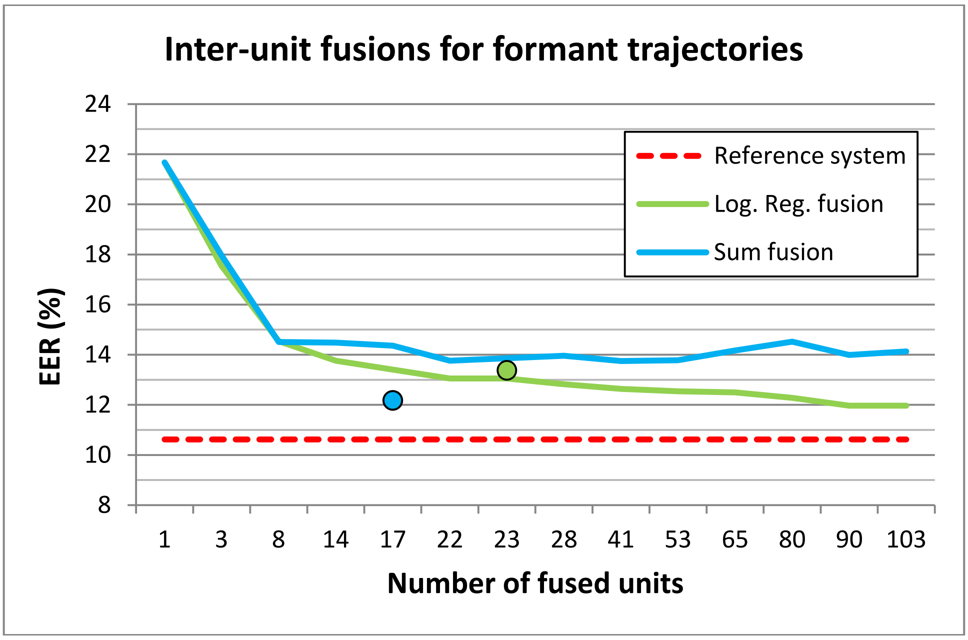


Figure 22. Inter-unit fusions for formant trajectories.

Best inter-unit fusion	EER (%)	# fused units
N-best, log. reg. fusion	11.97	90

Table 9. Best inter-unit fusion for formant trajectories.

Also in the case of cepstral contours (Figure 23), the combination of the two linguistic levels leads to greater improvements of discriminating power, reaching an EER as low as 6.57% when fusing by means of the sum rule just 22 units selected by the SFS algorithm.

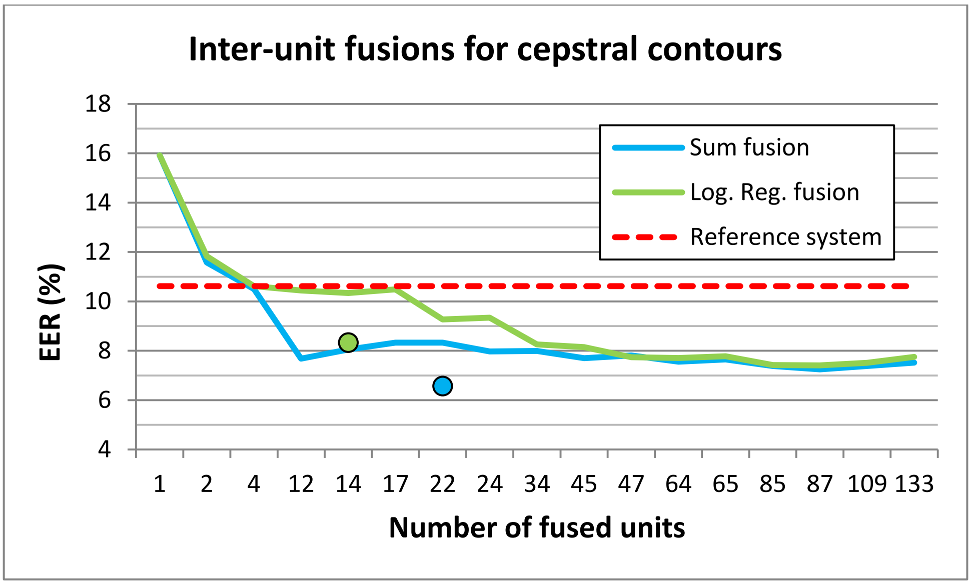


Figure 23. Inter-unit fusions for cepstral contours.

Best inter-unit fusion	EER (%)	# fused units
Sequential Forward Selection, sum fusion	6.57	22

Table 10. Best inter-unit fusion for cepstral contours.

Finally, best fusions of unit-dependent systems based on cepstral contours are summarized in Table 11 in order to compare with state of the art systems based on static MFCCs. As it can be seen, in the case of formant trajectories, all fused systems performs worse than the reference ones. In contrast, for cepstral contours their best fusions at every level (intra-unit phone, intra-unit diphone and inter-unit) outperform results obtained by any of our MFCC-based reference systems.

	System	# fused units	EER (%)
Formant	Diphones – best fused system (sum)	41	13.68
	Phones – best fused system (sum)	18	12.29
	Phones+diphones – best fused system (sum)	90	11.97
	GMM-UBM MFCC-based	-	10.26
	i-vector MFCC-based	-	8.86
Cepsta	Diphones – best fused system (log. reg.)	31	8.05
	Phones – best fused system (sum)	17	7.11
	Phones+diphones – best fused system (sum)	22	6.57

Table 11. Comparison between state-of-the-art MFCC-based systems and the best fusions of linguistic systems.

5.5 Fusion with non-linguistic systems

Apart from being more interpretable, one of the advantages of systems based on higher-level features is the potential for combination with short-term spectral systems, due to the different nature and time span of features involved. In order to analyze this complementarity, the fusion of the systems based on linguistic units and classical MFCC-based systems has been carried out.

Table 12 shows the individual performance for the best fusion of systems based on formant trajectories, for our two short-term reference systems, and for the fusion of both of them with the higher-level system. Although the short-term spectral systems differ in almost 2% EER, the fused system achieves very similar results in both cases, showing that very complementary information is being provided by the higher-level system. Moreover, the performance is greatly improved in both cases, being highly remarkable the 17% relative improvement obtained in the case of the fusion with the i-vector system.

System	EER (%)	minDCF
1) GMM-UBM MFCC	10.26	0.0457
2) i-vector MFCC	8.86	0.0407
3) Formant trajectories - best fusion	11.97	0.0636
Sum fusion of 1 and 3	7.51	0.0437
Sum fusion of 2 and 3	7.33	0.0356

Table 12. Performance of the best fusion of formant trajectories and MFCC-based systems in the NIST SRE 2006 English-only male 1side-1side task.

Table 13 shows the same as Table 12 but for the best fusion of higher level systems based on cepstral contours. It can be seen that, while the performance of the fused system doesn't change significantly when fusing with the GMM-UBM reference system, a significant improvement can be achieved when fusing with the i-vector system.

System	EER (%)	minDCF
1) GMM-UBM MFCC	10.26	0.0457
2) i-vector MFCC	8.86	0.0407
3) Cepstral contours - best fusion	6.57	0.0367
Sum fusion of 1 and 3	6.97	0.0329
Sum fusion of 2 and 3	5.53	0.0266

Table 13. Performance of the best fusion of cepstral contours and MFCC-based systems in the NIST SRE 2006 English-only male 1side-1side task.

Chapter 6: Conclusions and future work

THIS CHAPTER SUMMARIZES the main results and contributions of this Master Thesis and outlines future research lines to work on.

6.1 Conclusions

In this Master Thesis we have presented an analysis of the contributions of individual linguistic units to automatic speaker recognition by means of their temporal contours, both from formant frequencies and from MFCC features. In this way, some elements of traditional phonetic-acoustic and automatic approaches have been combined to face the forensic speaker recognition problem.

It has been shown that useful information can be obtained from isolated linguistic units due to their good calibration properties, with the advantage of being understandable cues easy to present in court for forensic purposes. The ratio between performance and amount of speech processed is very good for some of them, taken into account that for most of the units less than 7 seconds in average per utterance are available for training and testing stages. Interestingly, most of the best performing units are shared between the two types of dynamic features used. However, it should be clarified by further research whether this coincidence is due to the dynamic properties of acoustic features linked to linguistic information or to the fact that more data is available for some of these units to train and test our generative approach.

Moreover, it has been shown that speaker distinguishing information scattered among different units can be efficiently combined at different levels (intra- and inter-unit) by means of several techniques, reaching significant performances in the case of formant frequencies and outperforming state-of-the-art MFCC-based systems in the case of cepstral contours. It has been found that the best results are obtained when taking into account information from different types of linguistic units although there can be some overlap between speech segments from where they are extracted.

Finally, the best fusions of unit-dependent systems have been combined as well with state-of-the-art MFCC-based systems, leading to great improvements due to the different nature and time span of the features from each system.

6.2 Future work

As this is our first approach to the problem, lots of research can be done. First of all, the number of components of GMM-UBM unit-dependent systems has been set for convenience to the same value for every unit within a same unit-type (phones or diphones), in order to ease carrying experiments. However, it is likely that the best configuration for a particular linguistic unit require different number of components than any other, similarly to what has been observed for cepstral-constrained systems [Sanchez *et al.*, 2011]. This can be also applied to DCT coefficients retained in order to code the temporal contour, since longer linguistic units usually present richer contours. Also, other linguistic units (syllables, triphones, etc.) can be analyzed into the same likelihood ratio framework.

One important issue of further research is to apply newer automatic speaker recognition technologies such as JFA or total variability for obtaining unit-dependent speakers models. However, some difficulty arises for us in doing this. As we are coding each linguistic unit in a single feature vector, the amount of training samples is largely reduced in comparison with the use of short-term spectral features. Moreover, this fact is reinforced due to the unit-dependent processing of utterances, which leads to having just some feature vectors per utterance for speaker modeling. Furthermore, JFA and total variability needs lots of data in order to properly train variability matrices, and we are limited by the necessity of ASR transcription for bounding linguistic units, as we only have those of 2004, 2005 and 2006 NIST SREs. This pushes us to look for other variants in order to segment the speech signal, like those based on language independent phone tokenizers [Shriberg and Bocklet, 2011].

Also, although a clear likelihood-ratio framework has been used for obtaining well calibrated LRs, it would be of great interest to compare it with other methods that computes LR values directly from the features, as for example the MVLR technique [Aitken and Lucy, 2005].

References

[Adami *et al.*, 2003]

Adami, A.G., Mihaescu, R., Reynolds, D.A. and Godfrey, J.J. Modeling prosodic dynamics for speaker recognition. In *Proceedings of the 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing -ICASSP*, pp. 788-791, vol.4, Hong Kong, China, April 6-10, 2003.

[Aitken and Lucy, 2005]

C. G. G. Aitken and D. Lucy. Evaluation of trace evidence in the form of multivariate data. *Applied Statistics* 53, pp. 109-122, with corrigendum pp. 665-666, 2005

[Arpabet]

Wikipedia contributors. Arpabet. *Wikipedia, The Free Encyclopedia.*

[Atal, 1974]

B. Atal. Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verificaction. *Journal of the Acoustic Society of America*, 55 (6), 1304-1312, 1974.

[Bimbot *et al.*, 2004]

Frederic Bimbot, Jean-François Bonastre, Corinne Fredouille, Guillaume Gravier, Ivan Magrin-Chagnolleau, Sylvain Meignier, Teva Merlin, Javier Ortega-Garcia, Dijana Petrovska-Delacretaz and Douglas A. Reynolds. A Tutorial on Text-Independent Speaker Verification. *EURASIP Journal on Applied Signal Processing*, Vol. 4 (2004), pp. 430-451.

[Bocklet and Shriberg, 2009]

Tobias Bocklet and Elizabeth Shriberg. Speaker recognition using syllable-based constraints for cepstral frame selection. In *Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4525-4528, Taipei, Taiwan, April 19-24, 2009. Print ISBN: 978-1-4244-2353-8.

[Boersma, 2001]

Boersma, P. Praat, a system for doing phonetics by computer. *Glot International*, Vol. 5, No. 9/10, 2001, pp. 341-345. http://www.praat.org/

[Brummer and du Preez, 2006]

N. Brümer and J. du Preez. Application-independent evaluation of speaker detection. *Computer Speech and Language*, 20(2-3):230-275, 2006.

[Campbell *et al.*, 2006]

W. Campbell, J. Campbell, D. Reynolds, E. Singer, P. Torres-Carrasquillo. Support vector machines for speaker and language recognition. *Computer Speech and Language*, 20 (2-3), 210-229, 2006.

[de Castro *et al.*, 2009]

A. d. Castro, D. Ramos, and J. Gonzalez-Rodriguez. Forensic speaker recognition using traditional features comparing automatic and human-in-the-loop formant tracking. In *Proceedings of the 10th Annual Conference of the International Speech Communication Association (Interspeech'09)*, pp. 2343-2346, Brighton, United Kingdom, September 6-10, 2009.

[Dehak *et al.*, 2007]

Najim Dehak, Patrick Kenny, Pierre Dumouchel. Continuous prosodic features and formant modeling with joint factor analysis for speaker verification. In *Proceedings of the 8th Annual Conference of the International Speech Communication Association (INTERSPEECH 2007)*, pp. 1234-1237, Antwerp, Belgium, August 27-31, 2007.

[Dehak *et al.*, 2011]

Dehak, N., Kenny, P., Dehak, R. and Dumouchel, P. Front-End Factor Analysis for Speaker Verification. *IEEE Transactions on Audio, Speech and Language Processing*, 19(4), pp. 788–798 (2011).

[Dempster *et al.*, 1997]

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. 22, 48.

[Franco-Pedroso *et al.*, 2012]

J. Franco-Pedroso, J. Gonzalez-Rodriguez, J. Gonzalez-Dominguez, and D. Ramos. Fine-grained automatic speaker recognition using cepstral trajectories in phone units. *Quantitative approaches to problems in linguistics – Studies in honor of Phil Rose.* Cathryn Donohue, Shunichi Ishihara, William Steed (editors). ISBN 9783862883844. LINCOM Studies in Phonetics 08, pp. 185-196, 2012.

[Furui, 1981]

S. Furui. Cepstral analysis technique for automatic speaker verification. *IEEE Transactions on Acoustics, Speech Signal Processing*, 29 (2), pp. 254-272, 1981.

[Gauvain and Lee, 1994]

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. 23, 51.

[Gonzalez-Rodriguez *et al.*, 2007]

Gonzalez-Rodriguez, Joaquin, Phil Rose, Daniel Ramos, Doroteo T. Toledano & Javier Ortega-Garcia. 2007. Emulating DNA: Rigorous Quantification of Evidential Weight in Transparent and Testable Forensic Speaker Recognition. *IEEE Transactions on Audio, Speech and Language Processing*, 15(7), pp. 2104—2115.

[Gunnar Fant, 1960]

Fant, G. (1960). *Acoustic Theory of Speech Production.* Mouton & Co, The Hague, Netherlands.

[Hermanski and Morgan, 1994]

H. Hermansky and N. Morgan. RASTA processing of speech. *IEEE Transactions on Speech and Audio Processsing*, 2 (4), pp. 578-589.

[Kajarekar *et al.*, 2009]

S. Kajarekar, N. Scheffer, M. Graciarena, E. Shriberg, A. Stolcke and L. Ferrer. The SRI NIST 2008 Speaker Recognition Evaluation System. In *Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4205-4209, Taipei, Taiwan, April 19-24, 2009.

[Kenny *et al.*, 2003]

P. Kenny, M. Mihoubi, and P. Dumouchel. New MAP Estimators for speaker recognition. In *Proceedings of the 8th European Conference on Speech Communication and Technology (Eurospeech 2003)*, pp. 2961-2964, Geneva, Switzerland, Sept. 2003.

[Kenny and Dumouchel, 2004]

P. Kenny and P. Dumouchel. Experiments in Speaker Verification Using Factor Analysis Likelihood Ratios. In *Proceedings of Odyssey 2004: The Speaker and Language Recognition Workshop*, pp. 219–226, Toledo, Spain, May 31 - June 3, 2004.

[Kenny *et al.*, 2005]

P. Kenny, G. Boulianne, and P. Dumouchel. Eigenvoice modeling with sparse training data. *IEEE Transactions on Speech and Audio Processing*, 13(3):345–354, 2005.

[Kenny *et al.*, 2008]

Kenny, P., Ouellet, P., Dehak, N. and Gupta, V. A Study of Inter-speaker Variability in Speaker Verification. *IEEE Transactions on Audio, Speech and Language Processing.* 16(5), pp. 980–988 (2008).

[Kockmann and Burget, 2008]

Kockmann, M. and Burget, L. Syllable based Feature-Contours for Speaker Recognition. In *Proceedings of the 14th International Workshop on Advances in Speech Technology*, Maribor, SI, 2008, p. 4.

[Kockmann and Burget, 2008b]

Kockmann, M. and Burget, L. Contour modeling of prosodic and acoustic features for speaker recognition. In *Proceedings of 2008 IEEE Workshop on Spoken Language Technology*, pp. 45-48, Goa, India, December 15-18, 2008. ISBN 978-1-4244-3472-5.

[Kockmann *et al.*, 2010a]

Marcel Kockmann, Lukas Burget, Jan Cernocky. Investigations into prosodic syllable contour features for speaker recognition. In *Proceedings of the 2010 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, pp. 4418-4421, Dallas, Texas, US, March 15-19, 2010.

[Kockmann *et al.*, 2010b]

Kockmann, M., Burget, L., Glembek, O., Ferrer, L. and Cernocky, J. Prosodic Speaker Verification using Subspace Multinomial Models with Intersession Compensation. In *Proceedings of the 11th Annual Conference of the International Speech Communication*

Association (INTERSPEECH 2010), pp. 1061-1064, Makuhari, Chiba, Japan, September 26-30, 2010. ISBN 978-1-61782-123-3, ISSN 1990-9772.

[**Kockmann *et al.*, 2011**]

Marcel Kockmann, Luciana Ferrer, Lukas Burget and Jan Cernocky. iVector Fusion of Prosodic and Cepstral Features for Speaker Verification. In *Proceedings of the 12th Annual Conference of the International Speech Communication Association (INTERSPEECH 2011)*, pp. 265-268, Florence, Italy, August 27-31, 2011.

[**Li and Rose, 2012**]

Jingwen Li and Phil Rose. Likelihood Ratio-based Forensic Voice Comparison with F-pattern and Tonal F0 from the Cantonese /əy/ Diphthong. In *Proceedings of the 14th Australasian International Conference on Speech Science and Technology (SST)*, pp. 201-204, Sydney, Australia, December 3-6, 2012.

[**Linde *et al.*, 2003**]

Y. Linde, A. Buzo, and R. Gray. An Algorithm for Vector Quantizer Design. Communications. *IEEE Transactions on Communications*, 28(1):84–95, Jan. 2003. URL http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1094577. 23.

[**Malayath *et al.*, 2000**]

N. Malayath, H. Hermansky, S. Kajarekar and B. Yegnanarayana. Data-driven temporal filters and alternatives to GMM in speaker verification. *Digital Signal Processing*, 10 (1-3), pp. 55-74, 2000.

[**Martin *et al.,* 1997**]

A. Martin, G. Doddington, T. Kamm, M. Ordowski and M. Przybocki. The DET curve in assessment of decision task performance. In *Proceedings of EuroSpeech97, 5th European Conference on Speech Communication and Technilogy*, pp. 1895-1898, Rhodes, Greece, September 22-25, 1997.

[**Mary and Yegnanarayana, 2008**]

Mary, L. and Yegnanarayana, B. "Extraction and representation of prosodic features for language and speaker recognition". In Speech Communications Volume 50, Issue 10, October 2008, pp. 782-796.

[**McDougall, 2006**]

K. McDougall. Dynamic features of speech and the characterization of speakers: towards a new approach using formant frequencies. *International Journal of Speech, Language and the Law* 13(1), pp. 89-126, 2006.

[**Morrison, 2008**]

Morrison, G. S. Forensic voice comparison using likelihood ratios based on polynomial curves fitted to the formant trajectories of Australian English /aI/. *International Journal of Speech, Language and the Law*, 15, pp. 249-266.

[Morrison *et al.*, 2008]

Geoffrey Stewart Morrison, Phil Rose and Yuko Kinoshita. Extraction of likelihood-ratio forensic evidence from the formant trajectories of diphthongs. *Journal of the Acoustical Society of America*, Volume 123, Issue 5, pp. 3877-3877 (1 page), June 2008. DOI:10.1121/1.2935780.

[Morrison, 2009]

G. S. Morrison. Likelihood-ratio-based forensic speaker comparison using parametric representations of vowel formant trajectories. *Journal of the Acoustical Society of America*, 125, pp. 2387–2397 (2009).

[Nolan, 1983]

F. Nolan. *The Phonetic bases of speaker recognition.* Cambridge University Press, Cambridge (UK), 1983.

[Nolan, 2002]

Nolan F. The 'telephone effect' on formants: a response. *Forensic Linguistics*, 9(1):74-82, 2002.

[Park and Hanzen, 2002]

Akex Park and Timothy J. Hansen. ASR Dependent Techniques For Speaker Identification. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP 2002)*, Denver, Colorado, USA, September 2002, pp. 1337–1340.

[Pelecanos and Sridharan, 2001]

J. Pelecanos y S. Sridharan. Feature warping for robust speaker verification. In *Proceedings of Speaker Odyssey: the Speaker Recognition Workshop (Odyssey 2001)*, Creta, Grecia, June 2001, pp. 213-218.

[Przybocki *et al.*, 2007]

Przybocki, M.A., Martin, A.F. and Le, A.N. NIST Speaker Recognition Evaluations Utilizing the Mixer Corpora—2004, 2005, 2006. *IEEE Transactions on Audio, Speech, and Language Processing*, (Volume: 15, Issue: 7), pp. 1951-1959. September 2007.

[Reynolds *et al.*, 2000]

Reynolds, D.A., Quatieri, T.F., Dunn, R.B. Speaker verification using adapted Gaussian mixture models. *Digital Signal Processing* 10, pp. 19–41 (2000).

[Reynolds, 2003]

D. Reynolds. Channel robust speaker verification via feature mapping. In *Proceedings of the 2003 International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2003)*, Vol. 2, pp. 53-56, Hong Kong, China, April 2003.

[Rose, 2002]

Rose, P. *Forensic Speaker Identification.* Taylor&Francis, 2002.

[Rose and Winter, 2010]

Phil Rose and Elaine Winter. Traditional Forensic Voice Comparison with Female Formants: Gaussian mixture model and multivariate likelihood ratio analyses. In

Proceedings of the 13th Australasian International Conference on Speech Science and Technology (SST), Melbourne, Australia, December 14-16, 2010, pp 42-45, ISBN 978-0-9581946-3-1.

[Sanchez *et al.*, 2011]

M. Sanchez, L. Ferrer, E. Shriberg and A. Stolcke. Constrained cepstral speaker recognition using matched UBM and JFA training. In *Proceedings of the 12th Annual Conference of the International Speech Communication Association (INTERSPEECH 2011)*, pp. 141-144, Florence, Italy, August 27-31, 2011.

[Shriberg *et al.*, 2005]

E. Shriberg, L. Ferrer, S. Kajarekar, A. Venkataraman and A. Stolcke. Modeling Prosodic Feature Sequences for Speaker Recognition. *Speech Communications* 46, July 2005, pp. 455-472.

[Shriberg and Ferrer, 2007]

Elizabeth Shriberg and Luciana Ferrer. A Text-Constrained Prosodic System for Speaker Verification. In Proceedings of the *8th Annual Conference of the International Speech Communication Association (Interspeech 2007)*, pp. 1226-1229, Antwerp, Belgium, August 27-31, 2007.

[Sjolander and Beskow, 2000]

K. Sjolander and J. Beskow. Wavesurfer – an open source speech tool. In *Proceedings of the 6th International Conference on Spoken Language Processing (ICSLP 2000)*, pp. 464-467, Beijing, China, 2000.

[Sjolander]

Sjolander K. The Snack Sound Toolkit. Online on: http://www.speech.kth.se/snack

[SRI]

SRI International: http://www.sri.com/

[Stevens *et al.*, 1937]

Stevens, Stanley Smith; Volkman; John; & Newman, Edwin B. (1937). A scale for the measurement of the psychological magnitude pitch. *Journal of the Acoustical Society of America* 8 (3):185-190.

[Stevens, 1971]

Stevens, K. N. Sources of inter- and intra-speaker variability in the acoustic properties of speech sounds. In *Proceedings of the 7th International Congress of Phonetic Science*, pp. 22-28, August, 1971, Montreal, Mouton, 206-32.

[Sturim *et al.*, 2002]

Sturim, D.E., Reynolds, D.A., Dunn, R.B. and Quatieri, T.F. Speaker verification using text-constrained Gaussian Mixture Models. In *Proceedings of the 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2002)*, pp. 677-680, Volume 1, May 13-17 2002, Orlando, Florida, USA. Print ISBN: 0-7803-7402-9.

[Shriberg and Bocklet, 2011]

Elizabeth Shriberg and Tobias Bocklet. Language-independent constrained cepstral features for speaker recognition. In *Proceedings of the 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2011)*, pp. 5296-5299, May 22-27, 2011, Prague, Check Republic. Print ISBN: 978-1-4577-0538-0.

[Tcl]

Tool Command Language: http://en.wikipedia.org/wiki/Tcl.

[Teunen *et al.*, 2000]

R. Teunen, B. Shahshahani, and L. Heck. A Model-based Transformational Approach to Robust Speaker Recognition. In *Proceedings of Sixth International Conference on Spoken Language Processing, ICSLP 2000 / INTERSPEECH 2000*, pp. 495-498, Beijing, China, October 16-20, 2000.

[van Leeuwen *et al.*, 2006]

D. van Leeuwen, A. Martin, M. Przybocki and J. Bouten. The NIST 2004 and TNO/NFI speaker recognition evaluations. *Compute Speech and Language*, 20(2-3):128-158, 2006.

[Zhang *et al.*, 2008]

Cuiling Zhang, Geoffrey Stewart Morrison, Philip Rose. Forensic speaker recognition in Chinese: a multivariate likelihood ratio discrimination on /i/ and /y/. In *Proceedings of INTERSPEECH, 9th Annual Conference of the International Speech Communication Association*, pp. 22-26, Brisbane, Australia, September 22-26, 2008.

Appendix

- Individual unit-dependent results for formant trajectories:

Phone unit	EER (%)	minDCF	C_{llr}	minC_{llr}	Avg. length per utterance (s)
AA	34.10	0.0985	0.9005	0.8820	1.7
AE	24.92	0.0922	0.7466	0.7161	6.0
AH	30.46	0.0983	0.8435	0.8274	2.6
AO	31.82	0.0984	0.8655	0.8501	2.1
AW	36.66	0.0988	0.9259	0.9098	1.1
AX	26.54	0.0990	0.8080	0.7750	6.7
AY	21.67	0.0907	0.6949	0.6593	6.7
B	35.58	0.0985	0.9111	0.8964	1.9
CH	43.68	0.1000	0.9846	0.9755	0.7
D	36.05	0.0998	0.9222	0.9069	2.9
DH	29.38	0.0950	0.8160	0.7889	2.8
DX	40.98	0.1000	0.9711	0.9568	0.5
EH	29.11	0.0980	0.8262	0.8009	2.7
ER	35.58	0.0998	0.9140	0.8972	2.0
EY	32.21	0.0981	0.8593	0.8422	2.9
F	41.57	0.0997	0.9672	0.9572	2.0
G	38.42	0.1000	0.9509	0.9340	1.1
HH	36.80	0.0986	0.9386	0.9187	2.2
IH	28.92	0.0990	0.8172	0.7889	3.6
IY	30.59	0.0974	0.8453	0.8285	4.9
JH	40.64	0.0997	0.9665	0.9517	0.7
K	35.66	0.0997	0.9114	0.8993	4.3
L	23.74	0.0966	0.7490	0.7173	4.2
M	31.13	0.0966	0.8258	0.8091	3.4
N	26.27	0.0942	0.7790	0.7554	6.5
NG	40.41	0.0999	0.9562	0.9401	1.3
OW	27.78	0.0944	0.8088	0.7898	5.9
P	41.85	0.0999	0.9788	0.9639	1.8
PUH	27.36	0.0931	0.7925	0.7689	6.2
PUM	46.35	0.0998	0.9724	0.9603	1.4
R	25.47	0.0957	0.7672	0.7430	3.8
S	35.64	0.0992	0.9240	0.9068	6.0
SH	40.31	0.1000	0.9743	0.9515	0.8
T	33.27	0.0988	0.8910	0.8755	6.5
TH	41.53	0.0999	0.9683	0.9576	0.8
UH	42.22	0.0999	0.9754	0.9610	0.5
UW	33.70	0.0988	0.8762	0.8630	3.2
V	41.13	0.0997	0.9604	0.9499	1.5
W	33.57	0.0976	0.8791	0.8638	2.5
Y	26.03	0.0948	0.7916	0.7615	4.6
Z	35.33	0.0986	0.9105	0.8980	2.5

Table 14. Performance of phone-dependent units for formant trajectories in the NIST SRE 2006 English-only male 1side-1side task.

Diphone unit	EER (%)	minDCF	C_{llr}	minC_{llr}	Avg. length per utterance (s)
AA-R	38.48	0.1000	0.9630	0.9400	0.7
AA-T	37.08	0.0993	0.9311	0.9138	0.7
AE-N	30.73	0.0993	0.8479	0.8230	1.9
AE-T	31.90	0.0969	0.8720	0.8526	1.4
AE-V	38.82	0.0998	0.9380	0.9230	0.6
AH-M	33.02	0.0996	0.8985	0.8774	0.9
AH-N	36.94	0.0993	0.9444	0.9193	0.6
AH-T	35.82	0.0997	0.9215	0.9050	0.9
AO-L	43.68	0.1000	0.9813	0.9684	0.8
AO-R	34.04	0.0976	0.8868	0.8720	1.4
AW-T	41.06	0.0996	0.9598	0.9452	0.6
AX-B	42.88	0.1000	0.9797	0.9637	0.6
AX-D	44.24	0.0997	0.9782	0.9671	0.4
AX-G	42.52	0.0996	0.9740	0.9641	0.4
AX-K	34.78	0.0994	0.9213	0.9028	1.0
AX-L	37.48	0.0999	0.9403	0.9195	1.4
AX-M	40.72	0.1000	0.9680	0.9558	0.5
AX-N	23.84	0.0899	0.7583	0.7097	2.7
AX-NG	34.91	0.0964	1.0099	0.8752	1.1
AX-S	37.92	0.0991	0.9601	0.9317	1.2
AX-T	35.05	0.0993	0.9138	0.8952	1.4
AX-V	44.76	0.1000	0.9865	0.9737	0.5
AX-Z	43.29	0.1000	0.9742	0.9635	0.6
AY-D	42.53	0.1000	0.9777	0.9676	0.7
AY-K	32.45	0.0970	0.8494	0.8356	1.6
AY-M	33.22	0.0985	0.9014	0.8764	1.5
AY-N	40.45	0.1000	0.9528	0.9392	0.7
AY-T	36.00	0.0978	0.9192	0.8955	0.9
B-AH	41.85	0.1000	0.9643	0.9498	0.6
B-AX	44.37	0.1000	0.9779	0.9672	0.6
B-IY	41.75	0.1000	0.9733	0.9599	0.7
D-AX	43.28	0.1000	0.9746	0.9658	0.5
D-DH	45.31	0.1000	0.9839	0.9697	0.3
DH-AE	34.06	0.0993	0.8885	0.8729	1.4
DH-AX	38.82	0.1000	0.9541	0.9367	1.1
DH-EH	41.78	0.0999	0.9757	0.9593	0.8
DH-EY	40.33	0.1000	0.9725	0.9557	0.9
D-IH	46.83	0.1000	0.9892	0.9807	0.4
D-OW	42.74	0.0998	0.9572	0.9441	0.4
D-UW	45.86	0.1000	0.9871	0.9716	0.7
DX-AX	44.63	0.1000	0.9884	0.9773	0.4
DX-IY	41.54	0.0997	0.9699	0.9543	0.5
EH-L	42.21	0.0997	0.9716	0.9545	0.7
EH-N	36.84	0.0997	0.9205	0.9045	1.0
EH-R	36.43	0.0993	0.9107	0.8962	1.2
HH-AE	38.56	0.0995	0.9364	0.9178	0.9
HH-W	46.63	0.0995	0.9858	0.9779	0.5
I-HN	36.00	0.1000	0.9532	0.9126	1.2
IH-NG	32.89	0.0961	0.8822	0.8536	0.8
IH-T	36.20	0.1000	0.9253	0.9111	1.3
IY-AX	43.15	0.0996	0.9762	0.9671	0.4
IY-N	38.42	0.0990	0.9474	0.9251	0.7
IY-P	43.98	0.0996	0.9751	0.9600	0.5

JH-AX	42.47	0.0991	0.9639	0.9553	0.4
K-AH	42.12	0.0999	0.9658	0.9551	0.5
K-AX	40.17	0.0998	0.9680	0.9572	0.9
K-S	42.94	0.1000	0.9741	0.9632	0.4
L-AX	44.52	0.0995	0.9877	0.9730	0.5
L-AY	29.11	0.0972	0.8156	0.7955	1.6
L-IY	33.57	0.1000	0.9023	0.8779	1.2
M-AX	41.39	0.0999	0.9603	0.9485	0.4
M-AY	44.36	0.0999	0.9744	0.9655	0.6
M-IY	38.69	0.0984	0.9441	0.9103	0.8
N-AA	38.29	0.0986	0.9335	0.9208	0.6
N-AX	40.31	0.1000	0.9703	0.9541	0.6
N-D	24.92	0.0876	0.7563	0.7037	1.9
N-DH	39.90	0.0989	1.0018	0.9286	0.4
NG-K	38.46	0.1000	0.9603	0.9331	0.5
N-IY	38.69	0.0993	0.9427	0.9254	0.6
N-OW	30.86	0.0995	0.8455	0.8185	2.2
N-S	43.28	0.0994	0.9821	0.9676	0.5
N-T	35.72	0.0999	0.9330	0.9056	1.3
OW-N	39.99	0.0997	0.9654	0.9425	0.7
P-AX	40.04	0.1000	0.9646	0.9494	0.5
R-AX	41.53	0.0993	0.9643	0.9504	0.6
R-AY	38.09	0.0984	0.9274	0.9057	1.1
R-IY	34.97	0.0994	0.9165	0.8996	1.1
S-AH	43.96	0.1000	0.9810	0.9658	0.6
S-AX	41.21	0.0997	0.9681	0.9553	0.9
S-OW	36.00	0.1000	0.9288	0.9123	2.2
S-T	38.82	0.0997	0.9596	0.9333	1.5
T-AX	40.39	0.0994	0.9535	0.9348	1.4
T-AY	40.85	0.0998	0.9701	0.9553	0.8
T-DH	43.41	0.1000	0.9755	0.9620	0.5
TH-IH	37.88	0.0988	0.9444	0.9164	0.7
T-R	44.17	0.1000	0.9638	0.9521	0.6
T-S	33.42	0.0993	0.8945	0.8760	1.8
T-UW	35.87	0.0995	0.9264	0.9080	1.2
T-W	48.13	0.0997	0.9930	0.9847	0.4
UH-D	43.15	0.0997	0.9809	0.9709	0.5
UW-N	32.21	0.0953	0.8417	0.8188	0.8
V-AX	42.74	0.1000	0.9743	0.9615	0.4
W-AH	40.85	0.0995	0.9694	0.9487	0.9
W-AX	46.93	0.1000	0.9833	0.9735	0.3
W-EH	44.06	0.1000	0.9812	0.9720	0.8
Y-AE	29.78	0.0976	0.8383	0.8094	4.5
Y-UW	27.19	0.0960	0.8223	0.7812	3.2
Z-AX	39.64	0.0997	0.9704	0.9425	0.7

Table 15. Performance of diphone-dependent units for formant trajectories in the NIST SRE 2006 English-only male 1side-1side task.

- Individual unit-dependent results for cepstral contours:

Phone unit	EER (%)	minDCF	C_{llr}	minC_{llr}	Avg. length per utterance (s)
AA	32.20	0.0983	0.8633	0.8452	1.7
AE	18.98	0.0813	0.6087	0.5832	6.0
AH	29.39	0.0969	0.8235	0.7967	2.6
AO	34.36	0.0992	0.9065	0.8838	2.1
AW	36.99	0.0991	0.9241	0.9111	1.1
AX	27.08	0.0947	0.7882	0.7512	6.7
AY	21.68	0.0869	0.6822	0.6428	6.7
B	34.50	0.0986	0.8922	0.8778	1.9
CH	42.59	0.1000	0.9686	0.9538	0.7
D	32.07	0.0965	0.8661	0.8500	2.9
DH	28.43	0.0934	0.8403	0.7857	2.8
DX	40.44	0.0998	0.9670	0.9484	0.5
EH	31.69	0.0975	0.8574	0.8283	2.7
ER	35.18	0.0987	0.9107	0.8901	2.0
EY	26.40	0.0925	0.7713	0.7515	2.9
F	39.63	0.0993	0.9561	0.9397	2.0
G	35.71	0.1000	0.9291	0.9040	1.1
HH	39.80	0.0992	0.9527	0.9414	2.2
IH	26.95	0.0948	0.7964	0.7495	3.6
IY	23.32	0.0923	0.7453	0.7002	4.9
JH	39.69	0.0997	0.9487	0.9339	0.7
K	27.76	0.0961	0.8219	0.7832	4.3
L	26.51	0.0935	0.7789	0.7451	4.2
M	22.28	0.0857	0.6824	0.6583	3.4
N	15.92	0.0713	0.5520	0.5082	6.5
NG	29.37	0.0934	0.9977	0.7958	1.3
OW	24.65	0.0987	0.7917	0.7396	5.9
P	39.50	0.0988	0.9466	0.9335	1.8
PUH	24.18	0.0908	0.7359	0.7149	6.2
PUM	34.15	0.0953	0.8644	0.8419	1.4
R	24.65	0.0887	0.7295	0.7116	3.8
S	30.04	0.0973	0.8451	0.8059	6.0
SH	39.36	0.0996	1.0546	0.9294	0.8
T	27.89	0.0921	0.8256	0.7647	6.5
TH	38.37	0.1000	1.1207	0.9298	0.8
UH	41.53	0.1000	0.9717	0.9593	0.5
UW	24.79	0.0898	0.7391	0.7198	3.2
V	35.86	0.0990	0.9093	0.8932	1.5
W	35.82	0.0993	0.9167	0.8966	2.5
Y	24.00	0.0906	0.7313	0.7062	4.6
Z	32.07	0.0968	0.8487	0.8312	2.5

Table 16. Performance of phone-dependent units for cepstral contours in the NIST SRE 2006 English-only male 1side-1side task.

Diphone unit	EER (%)	minDCF	C_{llr}	minC_{llr}	Avg. length per utterance (s)
AA-R	38.48	0.1000	0.9630	0.9400	0.7
AA-T	37.08	0.0993	0.9311	0.9138	0.7
AE-N	30.73	0.0993	0.8479	0.8230	1.9
AE-T	31.90	0.0969	0.8720	0.8526	1.4
AE-V	38.82	0.0998	0.9380	0.9230	0.6
AH-M	33.02	0.0996	0.8985	0.8774	0.9
AH-N	36.94	0.0993	0.9444	0.9193	0.6
AH-T	35.82	0.0997	0.9215	0.9050	0.9
AO-L	43.68	0.1000	0.9813	0.9684	0.8
AO-R	34.04	0.0976	0.8868	0.8720	1.4
AW-T	41.06	0.0996	0.9598	0.9452	0.6
AX-B	42.88	0.1000	0.9797	0.9637	0.6
AX-D	44.24	0.0997	0.9782	0.9671	0.4
AX-G	42.52	0.0996	0.9740	0.9641	0.4
AX-K	34.78	0.0994	0.9213	0.9028	1.0
AX-L	37.48	0.0999	0.9403	0.9195	1.4
AX-M	40.72	0.1000	0.9680	0.9558	0.5
AX-N	23.84	0.0899	0.7583	0.7097	2.7
AX-NG	34.91	0.0964	1.0099	0.8752	1.1
AX-S	37.92	0.0991	0.9601	0.9317	1.2
AX-T	35.05	0.0993	0.9138	0.8952	1.4
AX-V	44.76	0.1000	0.9865	0.9737	0.5
AX-Z	43.29	0.1000	0.9742	0.9635	0.6
AY-D	42.53	0.1000	0.9777	0.9676	0.7
AY-K	32.45	0.0970	0.8494	0.8356	1.6
AY-M	33.22	0.0985	0.9014	0.8764	1.5
AY-N	40.45	0.1000	0.9528	0.9392	0.7
AY-T	36.00	0.0978	0.9192	0.8955	0.9
B-AH	41.85	0.1000	0.9643	0.9498	0.6
B-AX	44.37	0.1000	0.9779	0.9672	0.6
B-IY	41.75	0.1000	0.9733	0.9599	0.7
D-AX	43.28	0.1000	0.9746	0.9658	0.5
D-DH	45.31	0.1000	0.9839	0.9697	0.3
DH-AE	34.06	0.0993	0.8885	0.8729	1.4
DH-AX	38.82	0.1000	0.9541	0.9367	1.1
DH-EH	41.78	0.0999	0.9757	0.9593	0.8
DH-EY	40.33	0.1000	0.9725	0.9557	0.9
D-IH	46.83	0.1000	0.9892	0.9807	0.4
D-OW	42.74	0.0998	0.9572	0.9441	0.4
D-UW	45.86	0.1000	0.9871	0.9716	0.7
DX-AX	44.63	0.1000	0.9884	0.9773	0.4
DX-IY	41.54	0.0997	0.9699	0.9543	0.5
EH-L	42.21	0.0997	0.9716	0.9545	0.7
EH-N	36.84	0.0997	0.9205	0.9045	1.0
EH-R	36.43	0.0993	0.9107	0.8962	1.2
HH-AE	38.56	0.0995	0.9364	0.9178	0.9
HH-W	46.63	0.0995	0.9858	0.9779	0.5
I-HN	36.00	0.1000	0.9532	0.9126	1.2
IH-NG	32.89	0.0961	0.8822	0.8536	0.8
IH-T	36.20	0.1000	0.9253	0.9111	1.3
IY-AX	43.15	0.0996	0.9762	0.9671	0.4
IY-N	38.42	0.0990	0.9474	0.9251	0.7
IY-P	43.98	0.0996	0.9751	0.9600	0.5
JH-AX	42.47	0.0991	0.9639	0.9553	0.4
K-AH	42.12	0.0999	0.9658	0.9551	0.5
K-AX	40.17	0.0998	0.9680	0.9572	0.9
K-S	42.94	0.1000	0.9741	0.9632	0.4
L-AX	44.52	0.0995	0.9877	0.9730	0.5
L-AY	29.11	0.0972	0.8156	0.7955	1.6

L-IY	33.57	0.1000	0.9023	0.8779	1.2
M-AX	41.39	0.0999	0.9603	0.9485	0.4
M-AY	44.36	0.0999	0.9744	0.9655	0.6
M-IY	38.69	0.0984	0.9441	0.9103	0.8
N-AA	38.29	0.0986	0.9335	0.9208	0.6
N-AX	40.31	0.1000	0.9703	0.9541	0.6
N-D	24.92	0.0876	0.7563	0.7037	1.9
N-DH	39.90	0.0989	1.0018	0.9286	0.4
NG-K	38.46	0.1000	0.9603	0.9331	0.5
N-IY	38.69	0.0993	0.9427	0.9254	0.6
N-OW	30.86	0.0995	0.8455	0.8185	2.2
N-S	43.28	0.0994	0.9821	0.9676	0.5
N-T	35.72	0.0999	0.9330	0.9056	1.3
OW-N	39.99	0.0997	0.9654	0.9425	0.7
P-AX	40.04	0.1000	0.9646	0.9494	0.5
R-AX	41.53	0.0993	0.9643	0.9504	0.6
R-AY	38.09	0.0984	0.9274	0.9057	1.1
R-IY	34.97	0.0994	0.9165	0.8996	1.1
S-AH	43.96	0.1000	0.9810	0.9658	0.6
S-AX	41.21	0.0997	0.9681	0.9553	0.9
S-OW	36.00	0.1000	0.9288	0.9123	2.2
S-T	38.82	0.0997	0.9596	0.9333	1.5
T-AX	40.39	0.0994	0.9535	0.9348	1.4
T-AY	40.85	0.0998	0.9701	0.9553	0.8
T-DH	43.41	0.1000	0.9755	0.9620	0.5
TH-IH	37.88	0.0988	0.9444	0.9164	0.7
T-R	44.17	0.1000	0.9638	0.9521	0.6
T-S	33.42	0.0993	0.8945	0.8760	1.8
T-UW	35.87	0.0995	0.9264	0.9080	1.2
T-W	48.13	0.0997	0.9930	0.9847	0.4
UH-D	43.15	0.0997	0.9809	0.9709	0.5
UW-N	32.21	0.0953	0.8417	0.8188	0.8
V-AX	42.74	0.1000	0.9743	0.9615	0.4
W-AH	40.85	0.0995	0.9694	0.9487	0.9
W-AX	46.93	0.1000	0.9833	0.9735	0.3
W-EH	44.06	0.1000	0.9812	0.9720	0.8
Y-AE	29.78	0.0976	0.8383	0.8094	4.5
Y-UW	27.19	0.0960	0.8223	0.7812	3.2
Z-AX	39.64	0.0997	0.9704	0.9425	0.7

Table 17. . Performance of diphone-dependent units for cepstral contours in the NIST SRE 2006 English-only male 1side-1side task.