Feature-based likelihood ratios for speaker recognition from linguistically-constrained formant-based i-vectors

Javier Franco-Pedroso, Joaquin Gonzalez-Rodriguez

ATVS - Biometric Recognition Group
Universidad Autonoma de Madrid, Spain
javier.franco@uam.es

Abstract

In this paper, a probabilistic model is introduced to obtain feature-based likelihood ratios from linguistically-constrained formant-based i-vectors in a NIST SRE task. Linguistically-constrained formant-based i-vectors summarize both the static and dynamic information of formant frequencies in the occurrences of a given linguistic unit in a speech recording. In this work, a two-covariance model is applied to these "higher-level" features in order to obtain likelihood ratios through a probabilistic framework. While the performance of the individual linguistically-constrained systems are not comparable to that of a state-of-the-art cepstral-based system, calibration loss is low enough, providing informative likelihood ratios that can be directly used, for instance, in forensic applications. Furthermore, this procedure avoids the need for further calibration steps, which usually require additional datasets. Finally, the fusion of several linguistically-constrained systems greatly improves the overall performance, achieving very remarkable results for a system solely based on formant features. Testing on the English-only trials of the core condition of the NIST 2006 SRE (and using only NIST SRE 2004 and 2005 data for background and development, respectively), we report equal error rates of 8.47% and 9.88% for male and female speakers respectively, using only formant frequencies as speaker discriminative information.

1. Introduction

Formant frequencies have strong individualization potential [1] and have been used for forensic voice comparison for several decades [2]. However, most of the studies in automatic speaker recognition over the last two decades [3] have been based on higher dimensional representations of the speech signal (i.e. MFCC, PLP, etc.) due to their ability to extract speaker distinguishing information. Although they are based on spectral information, it is difficult to directly relate the physiological traits of an individual with the set of such extracted features [4]. Formant frequencies, on the other hand, are easily interpretable and directly related with anatomical and physiological characteristics [5] [1]. Moreover, interpretable features are helpful in order to correlate with human observations and may lead to find some clues that could be hidden even for very complex cepstral-based systems [6].

In forensic-phonetics, voice comparison is usually performed in the context of linguistic units [1, 7], but reported studies are usually based on limited experimental frameworks (in terms of number of speakers, number of analysed linguistic-units, or both) due to the manual processes involved in order to extract formant frequencies or labelling the analysed units. So, it is of broad interest to analyse the abilities of formant frequencies for speaker recognition following a similar approach but applied on a large-scale experimental framework with the aid of fully automatic systems.

While there have been previous studies on the use of formant frequencies for automatic speaker recognition [8], constraints have been used only in the feature extraction stage but not for speaker modelling. On the other hand, there have been several studies on text-constrained speaker modelling but using mainly cepstral [9] or prosodic features [10]. In both cases, forensic applications have not been addressed in depth. Thus, to some extent this research fill a gap in the literature, and the presented results can give useful insights for the practitioners in the forensic-phonetics field.

In [11], the authors showed that well calibrated likelihood ratios can be obtained per linguistic unit by means of i-vector systems independently developed from linguistically-constrained formant features. However, as using a simple scoring method, an additional calibration step was needed in order to obtain informative likelihood ratios. In this work, a probabilistic framework is applied instead, leading to likelihood ratios that can be directly used avoiding further calibration processes, which usually need additional datasets in order to avoid overoptimistic results. This probabilistic framework is based on a two-covariance generative model similar to that in [12], but with a simpler training step as it has been used in some forensic works [13, 14].

The remainder of the paper is organized as follows. The extraction process of linguistically-constrained formant-based i-vectors is detailed in Section 2. Section 3 introduces the probabilistic model applied to the linguistically-constrained formant-based i-vectors in order to obtain feature-based likelihood ratios. Section 4 describes the experimental framework used for this work, while Section 5 presents the results obtained. Finally, conclusions are drawn in Section 6.

2. Linguistically-constrained formant-based i-vectors

Linguistically-constrained formant-based i-vectors are extracted with the aid of several speech processing tools and attempt to summarize both the static and dynamic information of formant frequencies in the occurrences of a given linguistic unit in a speech recording. First, automatic formant tracking is used in order to obtain the formant frequencies in a given speech file. In order to account for the dynamic information, delta features are also computed and incorporated to the feature vectors. Then, an automatic speech recognition (ASR) system is used to split the stream of feature vectors into different linguistic units.

Finally, for each speech recording, the feature vectors corresponding to the occurrences of a given linguistic unit are used to compute a linguistically-constrained i-vector for that utterance.

2.1. Formant tracking and dynamic information

Automatic formant tracking has been used in order to compute the formant frequencies along a speech recording. Among the free software packages available, Wavesurfer [15] has been selected for this work due to the ease of automate this process for large databases through scripts written in Tcl/Tk [16], as it is developed using the Snack Sound Toolkit library [17].

The Wavesurfer/Snack formant tracker bases its formant-frequency estimates on a linear prediction analysis performed at each frame, and dynamic programming is used to refine the resulting trajectories [18]. It has been used with default parameters for both male and female speakers, except for the number of formant frequencies to be tracked, limited to three for this work (F1-F3).

In order to account for the dynamic information of formant frequencies, the *delta* (Δ) or derivative coefficients have been used. Although delta coefficients cannot summarize the whole formant trajectory along a linguistic segment as other approaches attempt [19, 20, 21], they can characterize the local dynamic information while keeping a frame-by-frame rate and a low dimensionality [11]. Derivative coefficients are finally appended to the instantaneous formant frequencies at each frame (10 ms each), giving rise to our 6-dimensional lower-level feature vectors.

2.2. Region conditioning and types of constraints

Voice comparison in forensic-phonetics is usually performed in the context of linguistic units [1, 7], as formant frequencies present much lower intra-speaker variability and higher inter-speaker variability [22, 7] when these constraints are applied to the features to be compared.

Automatic speech recognition (ASR) systems provide both phonetic content and time interval of speech regions in which the audio stream can be segmented. This phonetic content allow to define a large set of candidate constraints among the different types of linguistic units, showing each of them different characteristics in terms of within-unit formant dynamics, unit-length and frequency of occurrence. Among them, the following have been used for this work:

- Phones: although they are the shortest units and can appear in many different linguistic contexts, their high frequency of occurrence allow to develop more robust constrained systems. For this work, 39 phone units from an English lexicon plus two filled pauses (represented as PUH and PUM) were selected. These linguistic units are represented by the "2-character" ARPABET symbols [23] in the phonetic transcriptions provided by the ASR system [24] used. Table 1 shows the correspondence between Arpabet symbols and the International Phonetic Alphabet (IPA) ones.
- Diphones: defined as every possible combination of phone pairs, the 98 most frequent diphones were selected. Compared with phones, they present longer length but much lower frequency of occurrence. However, they show less contextual variation, which may lead to reduce the intra-speaker variability of formant dynamics between different occurrences of the same diphone.

In this work, the phonetic transcription labels produced by the SRIs Decipher ASR system [24] are used. For this system, trained on English data from telephonic conversations, the Word Error Rate (WER) on native and non-native speakers on the transcribed parts of the Mixer corpus, similar to NIST SRE databases used for this work, was 23.0% and 36.1% respectively.

2.3. I-vector extraction

An i-vector extractor [25] is a factor analysis (FA) based front-end which attempts to summarize the speaker distinguishing information in a given utterance, represented by a set of L feature vectors $\{f_1, f_2, ..., f_L\}$, through a single low-dimensional vector, the so-called identity vector or *i-vector* for short. This i-vector w accounts for the speaker and channel/session information present in a given utterance, representing it in a low-dimensional variability subspace. This is done by converting the speaker- and session-independent supervector (m), usually taken to be the UBM supervector, into the speaker- and session-dependent supervector (M) that represents a given speaker utterance through:

$$M = m + Tw \tag{1}$$

where T is a rectangular matrix of low rank defining the total variability (TV) space that contains the speaker and channel variability.

In order to obtain a linguistically-constrained i-vector (w^C), the i-vector extractor is applied only to the set of feature vectors $\{f_1^C, f_2^C, ...\}$ in the utterance belonging to a particular constraint, C:

$$M^C = m^C + T^C w^C \tag{2}$$

For this purpose, independent UBMs and TV subspaces are trained on the background dataset (see Section 4 for details) for every linguistic constraint under analysis. Both the number of components of the UBM (ranging from 2 to 256) and the number of dimensions of the TV space (ranging from 5 to 50) are optimized on the development dataset (see Section 4 for details) for each linguistic unit/constraint.

Finally, linguistically-constrained i-vectors are centred and whitened on the background dataset, and length-normalized.

3. Probabilistic model

3.1. The generative model

Conversely to [11], were cosine scoring and a further calibration step were used, in this work likelihood ratios are directly derived through a probabilistic framework. For this purpose, a two-covariance model [12] is applied. This is a generative model in which a particular observed i-vector $\mathbf{x}_{ij}$ coming from speaker i is generated through

$$\mathbf{x}_{ij} = \boldsymbol{\theta}_i + \boldsymbol{\psi}_j \tag{3}$$

where $\boldsymbol{\theta}_i$ is a realization of the speaker random variable Θ and $\boldsymbol{\psi}_j$ is a realization of the additive random noise Ψ representing its within-speaker variation. This noisy term is taken to be constant among different speakers and randomly distributed following

$$\Psi \sim \mathcal{N}(0, \mathbf{W}) \tag{4}$$

where $\mathbf{W}$ is the within-speaker covariance matrix. Thus, the conditional distribution of the random variable X_i (from which

$\mathbf{x}_{ij}$ is drawn), given a particular speaker i, follows a normal distribution with mean $\boldsymbol{\theta}_i$ and covariance matrix $\mathbf{W}$

$$\mathrm{X}_{ij}|(\Theta = \boldsymbol{\theta}_i) \sim \mathcal{N}(\boldsymbol{\theta}_i, \mathbf{W}) \tag{5}$$

On the other hand, speakers means are assumed to be normally distributed, following

$$\Theta \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{B}) \tag{6}$$

where $\boldsymbol{\mu}$ and $\mathbf{B}$ are, respectively, the mean vector and the covariance matrix of the between-speaker distribution.

3.2. Model training

Conversely to [12], model *hyperparameters* are directly computed in a single step instead of being iteratively trained to maximize the likelihood of the true partitioning of *m* speakers in the background dataset. This alternative procedure is more commonly used in forensic studies [13, 14], and it is applied in this work in order to avoid overfitting to the limited background dataset (NIST 2004 SRE).

Within-speaker covariance matrix is computed from the background dataset $\mathbf{X}$, comprising N i-vectors coming from m different speakers, through

$$\mathbf{W} = \frac{\mathbf{S}_w}{N - m} \tag{7}$$

being $\mathbf{S}_w$ the within-speaker scatter matrix given by

$$\mathbf{S}_w = \sum_{i=1}^{m} \sum_{j=1}^{n_i} (\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)^{\mathrm{T}} \tag{8}$$

where $\bar{\mathbf{x}}_i$ is the average of the set of n_i i-vectors from speaker i.

On the other hand, the mean vector and the covariance matrix of the between-speaker distribution are respectively computed by

$$\boldsymbol{\mu} = \frac{1}{m} \sum_{i=1}^{m} \bar{\mathbf{x}}_i \tag{9}$$

and

$$\mathbf{B} = \frac{\mathbf{S}_b}{m - 1} - \frac{\mathbf{S}_w}{\bar{n}(N - m)} \tag{10}$$

where $\bar{n}$ is the average number of i-vectors per speaker and the between-speaker scatter matrix, $\mathbf{S}_b$, is given by

$$\mathbf{S}_b = \sum_{i=1}^{m} (\bar{\mathbf{x}}_i - \boldsymbol{\mu})(\bar{\mathbf{x}}_i - \boldsymbol{\mu})^{\mathrm{T}} \tag{11}$$

3.3. Likelihood-ratio computation

Finally, the likelihood ratio between two given linguistically-constrained i-vectors $\mathbf{y}_1$ and $\mathbf{y}_2$ is computed as the ratio between

$$p(\mathbf{y}_1, \mathbf{y}_2) = \int_{\boldsymbol{\theta}} p(\mathbf{y}_1|\boldsymbol{\theta}, \mathbf{W})\, p(\mathbf{y}_2|\boldsymbol{\theta}, \mathbf{W})\, p(\boldsymbol{\theta}|\mathbf{X})\, \mathrm{d}\boldsymbol{\theta} \tag{12}$$

and

$$p(\mathbf{y}_1) \cdot p(\mathbf{y}_2) = \int_{\boldsymbol{\theta}} p(\mathbf{y}_1|\boldsymbol{\theta}, \mathbf{W})\, p(\boldsymbol{\theta}|\mathbf{X})\, \mathrm{d}\boldsymbol{\theta} \times \int_{\boldsymbol{\theta}} p(\mathbf{y}_2|\boldsymbol{\theta}, \mathbf{W})\, p(\boldsymbol{\theta}|\mathbf{X})\, \mathrm{d}\boldsymbol{\theta} \tag{13}$$

where $p(\mathbf{y}_l|\boldsymbol{\theta}, \mathbf{W}) = N(\mathbf{y}_l; \boldsymbol{\theta}, \mathbf{W})$ is the probability of a linguistically-constrained i-vector $\mathbf{y}_l$ given the knowledge of the speaker $\boldsymbol{\theta}$, and $p(\boldsymbol{\theta}|\mathbf{X}) = N(\boldsymbol{\theta}; \boldsymbol{\mu}, \mathbf{B})$ is the between-speaker probability density function obtained from the background dataset $\mathbf{X}$. Closed form expressions for these integrals can be found, for example, in [26].

4. Experimental framework

4.1. Datasets

In order to develop and test the linguistically-constrained systems, we have used the datasets and protocols belonging to the NIST SREs carried out on years 2004 [27], 2005 [28] and 2006 [29]. Among them, only English conversations have been used in order to match the characteristics of the ASR system [24]. No other datasets have been used as the authors have access only to the ASR phonetic labels corresponding to those datasets, kindly provided by SRI.

The composition of these datasets and the purposes they have been devoted to are described below:

- Background: NIST 2004 SRE dataset [27] comprises 2,541 files (1378 5-minutes, 581 30-seconds and 582 10-seconds long) from 125 male speakers and 3,626 files (2022 5-minutes, 802 30-seconds and 802 10-seconds long) from 187 female speakers. It has been used as the background dataset for training UBMs and total variability matrices. It also has been reused in order to train the *hyperparameters* of the probabilistic model.
- Development: NIST 2005 SRE dataset [28] has been used in order to optimize both UBMs components and number of dimensions of the TV subspaces. In [11], this dataset was divided into two halves for additional purposes: one half was used to train the calibration process and the other one to train the fusion rules. Here, as the calibration step is avoided through the introduced probabilistic framework, the whole dataset is used to train the fusion rules. The 1side-1side task of this NIST SRE comprises 11,272 trials from 243 male speakers and 14,793 trials from 342 female speakers.
- Evaluation: English-only trials from the core condition of the NIST 2006 SRE [29] were used for evaluating the proposed approach, consisting of 9,720 male trials for 219 target speakers and 14,293 female trials for 298 target speakers.

4.2. Evaluation metrics

Both the calibration and the discriminative properties of linguistically-constrained systems are analysed in this work. Discriminative properties are mainly evaluated through the equal error rate (EER) [30]. It is also used as the criterion by which the subsets of constraints are selected for the combination of linguistically-constrained systems. However, in accordance to the protocols used [29], the minimum of the C_{Det} (minDCF) is also reported. On the other hand, calibration properties [31] of linguistically-constrained systems are evaluated through the C_{llr} cost function and the calibration loss (C_{llr}^{loss}) [32].

5. Results

5.1. Reference systems

First, we want to compare with the previous approach in [11], were the same linguistically-constrained formant-based i-

vectors were used. In that study, cosine scoring, z-norm and a calibration step were used to obtain LRs per linguistic-unit. The results per constraint for this system can be seen in Table 2 (only the 10 best performing constraints are shown). The best fused system was obtained through a logistic regression fusion of the N-best performing constraints (see Section 5.3 or [11] for more details), trained on the same development dataset used in this work (NIST 2005 SRE). The performance of this fused system on the evaluation dataset (NIST 2006 SRE) is shown in Table 1.

Secondly, we wan to compare with a state-of-the-art cepstral-based system. Our cepstral reference system is based on an i-vector extractor from (unconstrained) MFCC features [25] and a Gaussian PLDA scoring stage [33]. Both gender-dependent 1024-component UBMs and 600-dimensional TV subspaces are trained on the background dataset (NIST SRE 2004), but GPLDA *hyperparameters* are trained on both the background and the development dataset (NIST SRE 2004 and 2005), applying a dimensionality reduction to 200. The performance of this system on the evaluation dataset (NIST SRE 2006) is also shown in Table 1.

	Reference systems			
	Male		Female	
	EER (%)	minDCF	EER (%)	minDCF
Formant-based	9.57	0.0456	12.89	0.0543
Cepstral-based	4.21	0.0232	5.67	0.0303

Table 1: *Results on the evaluation dataset for both formant-based and cepstral-based reference systems.*

5.2. Independent linguistically-constrained systems

Table 3 shows the results for the 10-best performing constraints (in terms of the EER) on the evaluation dataset when the introduced probabilistic framework is applied. As it can be seen, compared to the previous approach (Table 2), the discriminative performance per linguistic-unit (in terms of the EER) is significantly improved ($\sim$15% relative improvement on average for the shared constraints among the 10-best performing ones). Furthermore, although it is slightly increased compared to the previous approach, very low calibration losses are obtained (C_{llr}^{loss} $\sim$0.04 on average for this 10-best performing set) without the need for a specific calibration step.

5.3. Fusion of linguistically-constrained systems

Feature-based likelihood-ratios from different linguistic-constraints can be combined in order to account for the speaker distinguishing information spread among the different units. In this work, two fusion techniques have been used:

- First, a simple fusion rule, consisting on averaging the log-LRs of the subset of *N* constraints to be combined, has been applied through

$$logLR = \frac{1}{N} \sum_{\forall C\, in\, subset} logLR^C \qquad (14)$$

  where $logLR^C$ is the log-LR for a particular constraint *C*.

- Secondly, a linear combination of log-LRs is applied through

$$logLR = \alpha_0 + \sum_{\forall C\, in\, subset} \alpha^C logLR^C \qquad (15)$$

  where the vector of weights $\boldsymbol{\alpha} = [\alpha_0, \alpha^{C_1}, \alpha^{C_2}, ..., \alpha^{C_N}]$ is obtained by *logistic regression* [34] training on the development database, using the FoCal toolkit [35].

The specific subset of *N* constraints to be fused is obtained as follows. First, linguistically-constrained systems are sorted by performance, in terms of the EER, on the development dataset. Then, different fused systems are obtained by combining the first two, three, *etc.*, best performing linguistically-constrained systems, up to the total number of constraints. Among them, the fused system with the best performance on the development dataset, which is obtained by fusing the *N*-best performing constraints, is selected. While this may not be the set of constraints with the best performance on the evaluation dataset, it is expected that, for a large enough value of N, most of the best-performing constraints will be shared among development and evaluation datasets.

Table 4 shows the results obtained for both fusion techniques on the evaluation dataset. As it can be seen, the average rule make use of a much lower number of constraints than the logistic regression technique. This issue was analysed in [11], where it was shown that the performance of the logistic regression technique on the development dataset improved as more constraints were fused. For male trials, while the discriminative capabilities are similar for both techniques in terms of EER and minDCF, calibration properties are significantly better for the logistic regression technique, specially the calibration loss. The latter is also true for female trials, but also the discriminative capabilities are significantly better compared to the average rule.

Regarding our reference systems (Table 1), the logistic regression fusion of the feature-based LRs obtains a relative improvement in performance, in terms of the EER, of 11.5% and 23.3% for male and female trials, respectively, compared to our formant-based reference. While the performance is still far from the cepstral-based reference system, it is a very remarkable result for a system solely based on formant features which, in addition, can be directly applied in different forensic settings.

6. Conclusions

In this paper, we have introduced a probabilistic framework in order to obtain feature-based likelihood ratios from formant-based linguistically-constrained i-vectors. Linguistically-constrained formant-based i-vectors summarize both the static and dynamic information of formant frequencies in the occurrences of a given linguistic unit in a speech recording, and are extracted in a fully automatic way with the aid of several speech processing tools, including automatic formant tracking and automatic speech recognition.

A probabilistic model applied to formant-based i-vectors of a given linguistic constraint allows to provide feature-based likelihood ratios for isolated linguistic units, avoiding further calibration processes which usually need additional datasets. Although the discriminative performance of linguistically-constrained systems is not comparable to that of a cepstral-based state-of-the-art system, informative calibrated LRs can be obtained for voice comparisons without the need for further calibration steps.

Cosine scoring + z-normalization + calibration (log. reg.)									
Male					Female				
Constraint	EER (%)	minDCF	C_{llr}	C_{llr}^{loss}	Constraint	EER (%)	minDCF	C_{llr}	C_{llr}^{loss}
AE	21.21	0.0850	0.6668	0.0143	AY	24.59	0.0841	0.7101	0.0111
AY	21.38	0.0825	0.6580	0.0158	AE	24.59	0.0876	0.7308	0.0131
N	22.26	0.0812	0.6896	0.0168	L	24.68	0.0869	0.7355	0.0127
L	23.24	0.0839	0.7083	0.0133	N	24.77	0.0839	0.7256	0.0112
AX	23.80	0.0844	0.7001	0.0150	R	26.49	0.0932	0.7681	0.0132
AH	23.96	0.0964	0.7286	0.0158	AX	27.15	0.0932	0.7764	0.0100
PUH	24.32	0.0933	0.7296	0.0137	OW	27.79	0.0936	0.7830	0.0098
Y	24.68	0.0915	0.7325	0.0180	DH	27.79	0.0940	0.7876	0.0114
EH	24.83	0.0972	0.7544	0.0140	EH	28.06	0.0990	0.8196	0.0157
R	24.96	0.0937	0.7380	0.0149	AH	28.89	0.0974	0.8185	0.0079

Table 2: *Results on the evaluation dataset for the 10 best-performing constraints obtained with the previous approach in [11].*

Two covariance model									
Male					Female				
Constraint	EER (%)	minDCF	C_{llr}	C_{llr}^{loss}	Constraint	EER (%)	minDCF	C_{llr}	C_{llr}^{loss}
L	18.21	0.0817	0.6074	0.0421	L	20.21	0.0833	0.6542	0.0360
N	18.35	0.0764	0.6373	0.0574	N	21.02	0.0805	0.7455	0.1190
AY	18.43	0.0813	0.6090	0.0318	AE	21.16	0.0849	0.6896	0.0521
AE	19.50	0.0835	0.6411	0.0390	AY	21.58	0.0814	0.6606	0.0261
R	20.61	0.0889	0.6672	0.0275	AX	23.16	0.0898	0.6937	0.0144
Y	21.38	0.0888	0.6918	0.0575	R	23.59	0.0913	0.7147	0.0240
AX	21.50	0.0830	0.7012	0.0460	DH	24.30	0.0925	0.7316	0.0320
IH	21.76	0.0926	0.6885	0.0298	AH	25.05	0.0952	0.7541	0.0209
OW	21.90	0.0899	0.6911	0.0284	Y-AE	25.15	0.0866	0.7976	0.0914
DH	22.32	0.0892	0.6739	0.0231	OW	25.33	0.0906	0.7265	0.0294

Table 3: *Results on the evaluation dataset for the 10 best-performing constraints when the introduced probabilistic framework is applied.*

	***N*-best fusion of linguistically-constrained systems**									
	Male					Female				
	N	EER (%)	minDCF	C_{llr}	C_{llr}^{loss}	*N*	EER (%)	minDCF	C_{llr}	C_{llr}^{loss}
Average rule	19	8.76	0.0444	0.4318	0.1342	15	11.82	0.0565	0.4743	0.0864
Logistic regression	138	**8.47**	**0.0451**	**0.3210**	**0.0269**	139	**9.88**	**0.0512**	**0.3488**	**0.0140**

Table 4: *Results on the evaluation dataset for the fusion of the N-best performing linguistically-constrained systems, for the average rule and the logistic regression fusions.*

Furthermore, feature-based LRs can be successfully combined through different fusion techniques, obtaining great improvements in discriminative performance compared with the independent linguistically-constrained systems by themselves. For a simple average fusion rule, tens of units can be fused at the cost of slightly higher calibration losses. For the logistic regression technique, as being a trained fusion rule, a larger number of units can be fused while keeping very good calibration properties. While the performance is still far from the cepstral-based reference system, it is a very remarkable result for a system solely based on formant features which, in addition, can be directly applied in different forensic settings.

7. Acknowledgements

This work has been supported by the Spanish Ministry of Economy and Competitiveness (project CMC-V2: Caracterizacion, Modelado y Compensacion de Variabilidad en la Señal de Voz, TEC2012-37585-C02-01). Also, the authors would like to thank SRI for providing the Decipher phonetic transcriptions of the NIST 2004, 2005 and 2006 SREs that have allowed to carry out this work.

8. References

[1] Phillip Rose, *Forensic Speaker Identification*, Forensic Science. Taylor and Francis, 2002.

[2] Francis Nolan, *The phonetic bases of speaker recognition*, Cambridge University Press, Cambridge (UK), 1983.

[3] Tomi Kinnunen and Haizhou Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech Communication*, vol. 52, no. 1, pp. 12–40, 2010.

[4] Jonathan Darch, Ben Milner, Xu Shao, Saeed Vaseghi, and Qin Yan, "Predicting formant frequencies from MFCC vectors," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2005), Philadelphia, Pennsylvania, USA, March 18-23, 2005*, 2005, pp. 941–944.

[5] Francis Nolan and Catalin Grigoras, "A case for formant analysis in forensic speaker identification," *International Journal of Speech Language and the Law*, vol. 12, no. 2, 2005.

[6] Joaquin Gonzalez-Rodriguez, Juana Gil, Rubén Pérez, and Javier Franco-Pedroso, "What are we missing with i-vectors? a perceptual analysis of i-vector-based falsely accepted trials," in *Proceedings of Odyssey 2014: The Speaker and Language Recognition Workshop*, 2014, pp. 33–40.

[7] Kirsty McDougall, "Dynamic features of speech and the characterization of speakers: Toward a new approach using formant frequencies," *International Journal of Speech Language and the Law*, vol. 13, no. 1, pp. 89 – 126, 2006.

[8] N. Dehak, P. Dumouchel, and P. Kenny, "Modeling prosodic features with joint factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 7, pp. 2095–2103, Sept 2007.

[9] Tobias Bocklet and Elizabeth Shriberg, "Speaker recognition using syllable-based constraints for cepstral frame selection," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2009), 19-24 April 2009, Taipei, Taiwan*, 2009, pp. 4525–4528.

[10] Elizabeth Shriberg, "Higher-level features in speaker recognition," in *Speaker Classification I. Fundamentals, Features, and Methods*. 2007, vol. 4343 of *Lecture Notes in Computer Science*, pp. 241–259, Springer Berlin Heidelberg.

[11] Javier Franco-Pedroso and Joaquin Gonzalez-Rodriguez, "Linguistically-constrained formant-based i-vectors for automatic speaker recognition," *Speech Communication*, vol. 76, pp. 61 – 81, 2016.

[12] Niko Brümmer and Edward de Villiers, "The speaker partitioning problem," in *Odyssey 2010: The Speaker and Language Recognition Workshop, Brno, Czech Republic, June 28 - July 1, 2010*, 2010, p. 34.

[13] C. G. G. Aitken and D. Lucy, "Evaluation of trace evidence in the form of multivariate data," vol. 53, no. 1, pp. 109–122, Feb. 2004.

[14] Grzegorz Zadora, Agnieszka Martyna, Daniel Ramos, and Colin Aitken, *Statistical Analysis in Forensic Science: Evidential Values of Multivariate Physicochemical Data.*, Wiley, 2014.

[15] Kåre Sjölander and Jonas Beskow, "Wavesurfer - an open source speech tool," in *Proceedings of the Sixth International Conference on Spoken Language Processing (ICSLP 2000 / INTERSPEECH), Beijing, China, October 16-20, 2000*, 2000, pp. 464–467.

[16] "Tcl — Wikipedia, The Free Encyclopedia," 2015.

[17] "Snack Sound Toolkit — Wikipedia, The Free Encyclopedia," 2014.

[18] "Snack v2.2.8 manual," .

[19] Najim Dehak, Patrick Kenny, and 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), Antwerp, Belgium, August 27-31, 2007*, 2007, pp. 1234–1237.

[20] Joaquin Gonzalez-Rodriguez, "Speaker recognition using temporal contours in linguistic units: The case of formant and formant-bandwidth trajectories," in *Proceedings of the 12th Annual Conference of the International Speech Communication Association (INTERSPEECH 2011), Florence, Italy, August 27-31, 2011*, 2011, pp. 133–136.

[21] Javier Franco-Pedroso, Fernando Espinoza-Cuadros, and Joaquin Gonzalez-Rodriguez, "Formant trajectories in linguistic units for text-independent speaker recognition," in *Proceedings of the International Conference on Biometrics (ICB 2013), 4-7 June, 2013, Madrid, Spain*, 2013, pp. 1–6.

[22] Francis Nolan, "The 'telephone effect' on formants: a response," *International Journal of Speech Language and the Law*, vol. 9, no. 1, 2002.

[23] J. E. Shoup, "Phonological aspects of speech recognition," in *Trends in Speech Recognition*, Wayne A. Lea, Ed. 1980, pp. 125–138, Englewood Cliffs: Prentice Hall.

[24] Sachin S. Kajarekar, Nicolas Scheffer, Martin Graciarena, Elizabeth Shriberg, Andreas Stolcke, Luciana Ferrer, and Tobias Bocklet, "The SRI NIST 2008 speaker recognition evaluation system," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2009), 19-24 April 2009, Taipei, Taiwan*, 2009, pp. 4205–4208.

[25] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, May 2011.

[26] Javier Franco-Pedroso, Daniel Ramos, and Joaquin Gonzalez-Rodriguez, "Gaussian mixture models of between-source variation for likelihood ratio computation from multivariate data," *PLoS ONE*, vol. 11, no. 2, pp. 1–25, 02 2016.

[27] "The NIST Year 2004 Speaker Recognition Evaluation Plan," .

[28] "The NIST Year 2005 Speaker Recognition Evaluation Plan," .

[29] "The NIST Year 2006 Speaker Recognition Evaluation Plan," .

[30] Joaquin Gonzalez-Rodriguez, "Evaluating automatic speaker recognition systems: An overview of the nist speaker recognition evaluations (1996-2014)," *Loquens*, vol. 1, no. 1, pp. 1–15, January 2014.

[31] DavidA. van Leeuwen and Niko Brümmer, "An Introduction to Application-Independent Evaluation of Speaker Recognition Systems," in *Speaker Classification I*, Christian Müller, Ed., vol. 4343 of *Lecture Notes in Computer Science*, pp. 330–353. Springer Berlin Heidelberg, 2007.

[32] Niko Brümmer and Johan du Preez, "Application-independent evaluation of speaker detection," in *Computer Speech and Language*, 2006, vol. 20, pp. 230 – 275.

[33] Daniel Garcia-Romero and Carol Y. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems," in *Proceedings of the 12th Annual Conference of the International Speech Communication Association (INTERSPEECH 2011), Florence, Italy, August 27-31, 2011*, 2011, pp. 249–252.

[34] Stéphane Pigeon, Pascal Druyts, and Patrick Verlinde, "Applying logistic regression to the fusion of the nist'99 1-speaker submissions," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 237–248, 2000.

[35] Niko Brümmer, "Toolkit for evaluation, fusion and calibration of statistical pattern recognizers," .

[36] "Arpabet — Wikipedia, The Free Encyclopedia," 2014.

A. Mathematical notation

Column vectors are denoted by bold lower-case letters and matrices by bold upper-case letters, while scalar quantities are denoted by lower-case italic letters. Random variables are denoted by upper-case non-italic letters. $P(\cdot)$ is used to indicate the probability of a certain event, while $p(\cdot)$ denotes a probability density function. We denote a d-dimensional Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ by $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and the corresponding probability density function by $N(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ ($\mathbf{x} \in \mathbb{R}^d$).

B. Phonetic transcription codes

Vowels					
Monophthongs			*Monophthongs*		
Arpabet	IPA	Word examples	Arpabet	IPA	Word examples
AO	ɔ	off; fall; frost	AE	æ	at; fast
AA	ɑ	father; cot	*Diphthongs*		
IY	i	bee; see	Arpabet	IPA	Word examples
UW	u	you; new; food	EY	eɪ	say; eight
EH	ɛ	red; men	AY	aɪ	my; why; ride
IH	ɪ	big; win	OW	oʊ	show; coat
UH	ʊ	should; could	AW	aʊ	how; now
AH	ʌ	but; sun	*R-coloured vowels*		
	ə	sofa; alone	Arpabet	IPA	Word examples
AX		discus	ER	ɜ	her; bird; heart; nurse

Consonants					
Stops			*Affricates*		
Arpabet	IPA	Word examples	Arpabet	IPA	Word examples
P	p	pay	CH	tʃ	chair
B	b	buy	JH	dʒ	just
T	t	take	*Semivowels*		
D	d	day	Arpabet	IPA	Word examples
K	k	key	Y	j	yes
G	g	go	W	w	way
Fricatives			*Liquids*		
Arpabet	IPA	Word examples	Arpabet	IPA	Word examples
F	f	for	L	ɫ	late
V	v	very	R	ɾ *or* ɹ	run
TH	θ	thanks; Thursday	DX	ɾ	wetter
DH	ð	that; the; them	*Nasals*		
S	s	say	Arpabet	IPA	Word examples
Z	z	zoo	M	m	man
SH	ʃ	show	N	n	no
HH	h	house	NG	ŋ	sing

Table 1: *39 phones from the Arpabet phonetic transcription code and their correspondent IPA symbols (extracted from [36]).*