



Repositorio Institucional de la Universidad Autónoma de Madrid

<https://repositorio.uam.es>

Esta es la **versión de autor** del artículo publicado en:

This is an **author produced version** of a paper published in:

Speech Communication 76 (2016): 61 – 81

DOI: <http://dx.doi.org/10.1016/j.specom.2015.11.002>

Copyright: © 2016 Elsevier

El acceso a la versión del editor puede requerir la suscripción del recurso

Access to the published version may require subscription

Linguistically-constrained formant-based i-vectors for automatic speaker recognition[☆]

Javier Franco-Pedroso*, Joaquin Gonzalez-Rodriguez

ATVS - Biometric Recognition Group, EPS, Universidad Autonoma de Madrid, c/Francisco Tomas y Valiente 11, 28049 Madrid, Spain

Abstract

This paper presents a large-scale study of the discriminative abilities of formant frequencies for automatic speaker recognition. Exploiting both the static and dynamic information in formant frequencies, we present linguistically-constrained formant-based i-vector systems providing well calibrated likelihood ratios per comparison of the occurrences of the same isolated linguistic units in two given utterances. As a first result, the reported analysis on the discriminative and calibration properties of the different linguistic units provide useful insights, for instance, to forensic phonetic practitioners. Furthermore, it is shown that the set of units which are more discriminative for every speaker vary from speaker to speaker. Secondly, linguistically-constrained systems are combined at score-level through average and logistic regression speaker-independent fusion rules exploiting the different speaker-distinguishing information spread among the different linguistic units. Testing on the English-only trials of the core condition of the NIST 2006 SRE (24,000 voice comparisons of 5 minutes telephone conversations from 517 speakers -219 male and 298 female-), we report equal error rates of 9.57% and 12.89% for male and female speakers respectively, using only formant frequencies as speaker discriminative information. Additionally, when the formant-based system is fused with a cepstral i-vector system, we obtain relative improvements of $\sim 6\%$ in EER (from 6.54% to 6.13%) and $\sim 15\%$ in minDCF (from 0.0327 to 0.0279), compared to the cepstral system alone.

Keywords: automatic speaker recognition; formant frequencies; formant dynamics; linguistically-constrained systems

1. Introduction

Most of the studies in automatic speaker recognition over the last two decades have been based on compact representations of the speech signal in short analysis windows (i.e.

[☆]Non-standard abbreviations: NIST: US National Institute of Standards and Technology. SRE: Speaker Recognition Evaluation. ASR: Automatic Speech Recognition.

*Corresponding author.

Email addresses: javier.franco@uam.es (Javier Franco-Pedroso), joaquin.gonzalez@uam.es (Joaquin Gonzalez-Rodriguez)

MFCC, RASTA-PLP, etc.) [1]. Although they are based on spectral representations of the speech signal, it is difficult to directly relate the physiological traits of an individual with the set of such extracted features due to the additional transformations to which they are subjected (inverse FFT, DCT, etc.) [2]. Moreover, it is hard to interpret such kind of coefficients inasmuch as they do not correspond to any physical magnitude but to mathematical abstractions (the so-called cepstral domain). Formant frequencies, on the other hand, represent the resonant frequencies of the vocal tract of an individual, being easily interpretable and directly related with anatomical and physiological characteristics [3] [4]. This makes them specially suitable for forensic purposes [5] [6], where formant measurements have been used for forensic voice comparison for several decades [7] [4].

Voice comparison is usually performed in the context of linguistic units in forensic-phonetics [8], 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. In this way, the presented results can give useful insights for the practitioners in that field.

Furthermore, 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 [9]. Such kind of interpretable features, or the systems that make use of them, are usually classified as *higher-level* [10], and sometimes involve some kind of *constraints* [11] that are applied either in the feature extraction process (in order to define the feature itself), in the speaker modelling process (in order to reduce the intra-speaker variability), or both of them [10]. *Higher-level* systems provide very useful and complementary information that usually leads to performance improvements when they are combined with short-term acoustic systems [12] [13] [14].

With the objective of using interpretable features as formant frequencies but being able to evaluate them in the same challenging conditions of the state-of-the-art systems (e.g. the NIST Speaker Recognition Evaluations framework), we present in this paper a speaker verification system based on formant frequencies through the combination of different linguistically-constrained i-vector systems. While previous approaches [12] [15] [16] [17] extract the speaker distinguishing information from formant frequency dynamics through trajectories coding in the context of some linguistic units (phones, diphones, syllables or pseudo-syllables), in this work we address this issue by means of the classical derivative coefficients [18] [19], also known as *delta* (Δ) features, widely used in speech processing [20] in order to account for the dynamic information in the cepstral domain. This approach has the advantage of not reducing each linguistic segment (e.g. phone, diphone, *etc.*) to a single observation vector, relaxing the previous requirements of training data derived from extracting one single feature vector per linguistic segment.

The rest of the paper is organized as follows. Section 2 presents a brief overview of how formant frequencies have been used for speaker recognition, while Section 3 describes the automatic feature extraction process followed in the proposed approach. Section 4 details

how linguistically-constrained i-vector systems are built from formant features with the aid of automatically-generated phonetic labels. Section 5 describes the constraint-selection rules and fusion techniques used in order to combine the linguistically-constrained systems for text-independent speaker recognition. The experimental framework and evaluation metrics are presented in Section 6, including a description of our reference cepstral-based speaker recognition system. Results are shown in Section 7 for both independent linguistically-constrained systems and for several constraint combinations, as well as for the combination of formant and cepstral-based systems. Finally, conclusions are drawn in Section 8 and extended results are reported in a final appendix.

2. Formant frequencies for speaker recognition

Formant frequencies have strong individualization potential [7] and have been used for forensic voice comparison for several decades [4]. Usually, formant centre frequencies are extracted at the temporal midpoint of vowels [21] reflecting in part certain anatomical dimensions of a speaker as the length and configuration of the vocal tract. Also, the mean frequencies over the time-course of the vowel [22] have been used.

In order to obtain richer representations, frame-by-frame formant-frequency distributions have been modelled through either long-term formant distributions (LTFs) [3] or multivariate Gaussian mixture models (GMMs) [5]. It is also common to incorporate formant bandwidth measurements in order to complement the information provided by instantaneous formant frequency values [5] [16], as they are also related to vocal tract conditions.

Formant dynamics were also proposed for speaker recognition [8] under the assumption of presenting higher inter-speaker variability within linguistic units than the static measurements of formant frequencies: while speakers seems to show very similar acoustic properties at moments at which 'phonetic targets' [8] are achieved (e. g. formant frequencies at a segment's temporal midpoint), much larger differences are exhibited in the ways they move between consecutive targets [23].

This transitional information is omitted by statistical distributions obtained from frame-by-frame formant frequencies. In order to capture this dynamic information, two main approaches have been used: polynomial fitting [8] [12] and Discrete Cosine Transform (DCT) [24] [17] of formant trajectories over linguistic units. Both approaches compute a fixed number of polynomial or DCT coefficients per trajectory and concatenates the coefficients from the different formant trajectories, yielding a single feature vector that captures the dynamic information of the different formants in a given linguistic unit. In order to define the speech region where formant trajectories are computed, both manual segmentations (mainly in the forensic field) [24] [8] and automatic speech recognition (ASR) systems [12] [17] have been used. Using coded trajectories as feature vectors, speakers have been modelled through multivariate kernel distributions (MVK) [24] [16] or GMM's [17] in a linguistic unit-dependent manner, or by means of joint factor analysis (JFA), compensating for intersession variability, by pooling together trajectories from different units [12].

Similarly, the approach proposed in this paper is based on formant frequencies, but extracts the dynamic information through derivative coefficients [18] [19] regardless of the lin-

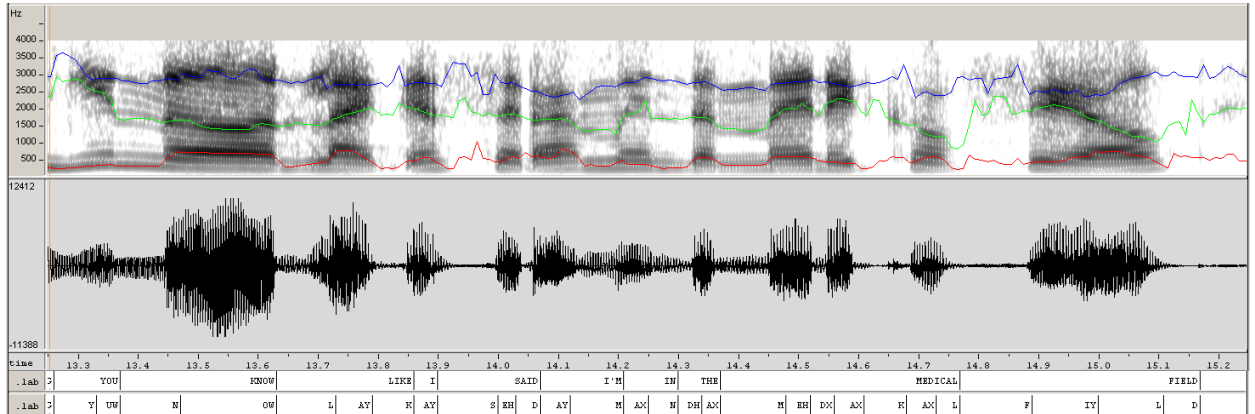


Figure 1: *Formant tracking for a speech sample in Wavesurfer[®], and its corresponding automatically extracted word and phone labels.*

guistic content. These coefficients are also extracted at a frame-by-frame rate and combined with the static information of instantaneous formant frequency values. Then, linguistic units are used as constraints applied to feature vectors in order to develop separate i-vector systems for each linguistic unit, allowing to independently analyse their speaker-distinguishing abilities.

3. Feature extraction

3.1. Formant tracking

Several methods and algorithms have been proposed for formant tracking [20], but only some of them have been implemented and made available within free software packages, as for example Wavesurfer [25], Praat [26] or WinSnoori [27]. Among them, the first one was selected for this work because it allows to easily automate this process for large databases. Wavesurfer is a general-purpose 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 developed using the Snack Sound Toolkit library [28], so scripts for automatic processing of large databases can be written in Tcl/Tk [29].

The 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 [30]. It was used with default parameters for both male and female speakers, except for the number of formant frequencies to be tracked. Most formant tracking estimators focus on formants F_1 - F_3 due to the fact that higher formants are progressively weaker in intensity [20]. Moreover, the average frequency position of F_4 is 3500 Hz, which is close to the cut-off frequency of the telephone-line band-pass filter. As in this work we are dealing with telephone-line speech, formant frequencies have then been extracted for the first three formants, with a 10 ms time resolution.

For the sake of simplicity in the feature-extraction phase, and due to the large number of speakers and linguistic units present in our experimental framework, no specific settings were

used for different speakers or units but a common one. For similar reasons, no exhaustive analysis was made regarding the suitability of the settings used, but just some shallower checks against typical formant values for the measurements obtained. As an automatic system, it will present errors that the following stages have to deal with.

3.2. Dynamic-information

While frame-by-frame formant frequencies can be estimated regardless of the linguistic content present in the speech signal, formant trajectories, as they have been used so far in speaker verification [12] [16] [17], can only be defined by using phonetic segmentations in order to delimit the speech region on which they are going to be coded. Working with automatic systems, both formant tracking errors and misalignment of phone label from the ASR will be observed, leading to erroneous coded trajectories in those cases.

An example is shown in Figure 2, where the phonetic transcription is correct but not properly aligned with the beginning of the acoustic signal and, therefore, spurious formant values computed at the beginning of the segment give rise to an artificial trajectory. If, for example, polynomial fitting is used in order to code the trajectory, the artificial spiky trajectory will require larger values in the higher order coefficients. Thus, the single feature vector corresponding to the whole linguistic unit will provide misleading information. Also, the same problem appears if some isolated spurious formant values arise within a well aligned phonetic transcription. On the contrary, if a frame-by-frame feature-extraction scheme is followed as will be used here, isolated spurious formant values only affect to the feature vectors extracted in these frames instead of the whole linguistic segment.

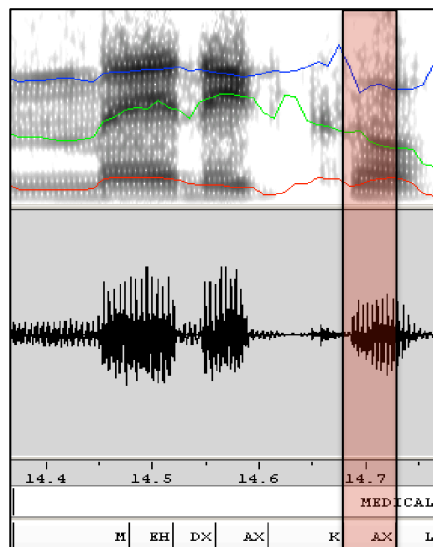


Figure 2: *Example of label temporal misalignment.*

Moreover, although the one-vector-per-linguistic-segment coding approach achieves a highly compact representation of formant trajectories in linguistic units, it greatly reduces the amount of data that can be used to train the parameters of the system, specially for

linguistic units with low frequency of occurrence. This problem is aggravated when linguistic constraints are applied for speaker modelling and comparison. For example, the diphone units analysed in this work present, on average, a frequency of occurrence of 10 times per conversation. Thus, if the trajectories of the first three formant frequencies are coded with the first 5 coefficients of its DCT, and the coefficients concatenated in a single feature vector, only ten 15-dimensional (3 formant trajectories \times 5 coefficients/trajectory) feature vectors will be available per conversation. Thus, sufficient statistics for the speaker modelling process have to be computed from a reduced number of observations (even lower than the number of features per observation). However, if a frame-by-frame feature extraction scheme is followed instead, the larger length of diphones will provide enough number of feature vectors in order to extract reliable sufficient statistics even with a low number of occurrences per conversation. In the previous example, assuming an average number of samples per diphone equal to 10, one hundred 3-dimensional samples will be available.

For these reasons, the *delta* (Δ) or derivative coefficients have been used to account for the dynamic information of formant frequencies instead of trajectory coding. Although delta coefficients cannot include the whole formant trajectory along the linguistic segment, they can characterize the local dynamic information while keeping a frame-by-frame feature extraction scheme. Delta coefficients were originally introduced for cepstrum coefficients [18] [19] in order to characterize the spectral transitional information, and are part of typical state-of-the-art speaker recognition systems. Applied to formant frequencies, this time derivative, approximated by a finite difference, has the following form

$$\frac{\delta F_m(t)}{\delta t} \approx \Delta F_m(t) = \frac{\sum_{k=-K}^K k h_k F_m(t+k)}{\sum_{k=-K}^K h_k k^2} \quad (1)$$

where $F_m(t)$ is the m -th formant frequency at time t and h_k is a window of length $2K + 1$ frames. In this study, a rectangular window ($h_k = 1$) is used with $K = 2$.

Finally, derivative coefficients are appended to instantaneous formant frequencies for each frame, giving rise to our 6-dimensional feature vectors at frame resolution (10 ms), $f(t)$.

$$f(t) = [F_1(t), F_2(t), F_3(t), \Delta F_1(t), \Delta F_2(t), \Delta F_3(t)] \quad (2)$$

While additional dynamic information could be added in a similar setting through the *delta-delta* (or *acceleration*) coefficients [20], this option has been discarded for practical reasons. As it will be shown in the following Section, independent speaker recognition systems are developed based on the different linguistic constraints. Thus, the number of feature vectors available for developing each independent system is highly reduced due to the region-conditioning process. If the dimensionality of the feature vectors is further increased, the ratio between the number of training samples and the complexity of the models is further reduced. As a trade-off between the amount of information and the complexity of the models, only delta features have been included in order to account for the dynamic information of formant frequencies.

Similarly, formants bandwidth information, while also used in forensic voice comparison, has been discarded based on preliminary experiments where including both formant frequencies and bandwidths did not improve the average performance across the different constraints, and have not been considered for further experiments in this work.

4. Linguistically-constrained speaker verification

Linguistically-constrained systems make use of an automatic speech recognition (ASR) system in order to condition the speech regions to be processed. ASR conditioning has been applied in automatic speaker recognition systems based on both cepstral [11] and higher-level [32] features. For cepstral-based systems, ASR conditioning is applied after the feature extraction process, defining the *constraints* to be applied by each subsystem to the features that can be used in speaker modelling and comparison stages. In this way, the intra-speaker variability due to the different lexical content between training and testing utterances is reduced. In the case of higher-level features, constraints are needed in order to define the feature itself, as they usually attempt to capture the dynamic behaviour of a specific measurement (pitch, energy, *etc.*) over several speech frames [10]. This is also the case of formant trajectories coding in the context of linguistic units. However, for systems based on prosodic information, once the features have been extracted, features belonging to different linguistic units are usually pooled together [12] [13] for the speaker modelling and comparison stages.

In this work, although ASR conditioning is avoided in the feature extraction process, constraints are applied in the speaker modelling and comparison stages. In this way, we aim not only to reduce the intra-speaker variability but also to test the discriminative abilities of formant frequencies within each linguistic unit independently, which can provide useful insights, specially to practitioners in forensic phonetics. Moreover, this allows to adopt a flexible approach to automatic speaker recognition where the linguistic specificities of particular speakers can be taken into account by using speaker-dependent constraints.

With this objective, we have developed independent i-vector systems [33] for each of the linguistic constraints under analysis, running in parallel for each speaker comparison (or *trial* in NIST SREs nomenclature) over the set of features belonging to its corresponding constraint. Additionally, calibrated likelihood-ratios (LRs) from a given subset of constraints can be combined in order to provide a single LR per trial.

4.1. Region conditioning

For the purpose of automatic region conditioning, we use the labels provided by an automatic speech recognition (ASR) system that produces transcriptions defining both phonetic content and time interval of speech regions in which the audio stream can be segmented. In this work, the phonetic transcription labels produced by the SRIs Decipher state-of-the-art ASR system [34] 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

Vowels					
<i>Monophthongs</i>			<i>Monophthongs</i>		
Arpabet	IPA	Word examples	Arpabet	IPA	Word examples
AO	ɔ	off; fall; frost	AE	æ	at; fast
AA	ɑ	father; cot	<i>Diphthongs</i>		
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	<i>R-coloured vowels</i>		
	ə	sofa; alone	Arpabet	IPA	Word examples
AX		discus	ER	ɜ	her; bird; heart; nurse

Consonants					
<i>Stops</i>			<i>Affricates</i>		
Arpabet	IPA	Word examples	Arpabet	IPA	Word examples
P	p	pay	CH	tʃ	chair
B	b	buy	JH	dʒ	just
T	t	take	<i>Semivowels</i>		
D	d	day	Arpabet	IPA	Word examples
K	k	key	Y	j	yes
G	g	go	W	w	way
<i>Fricatives</i>			<i>Liquids</i>		
Arpabet	IPA	Word examples	Arpabet	IPA	Word examples
F	f	for	L	ɫ	late
V	v	very	R	r or ɹ	run
TH	θ	thanks; Thursday	DX	ɾ	wetter
DH	ð	that; the; them	<i>Nasals</i>		
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 [31]).

36.1% respectively. While these results are equivalent to those obtained by other state-of-the-art systems on similar databases [35], transcription errors will be non negligible and will produce that, in order to compute the i-vector for a particular linguistic unit, some frames belonging to a different one will be taken into account, degrading the performance of the system based on that unit. In this work, no exhaustive analysis has been done regarding whether the errors occurred are associated with particular units or speakers, as we have no transcriptions available for the datasets used.

An analysis of such kind can be found in [36], where it is shown that errors are related with "*extreme prosodic characteristics, words occurring turn-initially, as discourse makers or preceding disfluent interruption points, and acoustically similar words that also have similar language model probabilities*". Thus, errors seem not to be associated with specific units but influenced by several aspects. It is also highlighted that "*speaker differences cause enormous variance in error rates*", and this seems to be "*not fully explained by differences in word choice, fluency, or prosodic characteristics*". Thus, a plausible cause can be the acoustic specificities of different speakers.

Regarding the results reported in this work, on one hand, a variable ASR performance across units would affect the relative performance among systems based on them. Thus, if a particular unit present worse speaker recognition performance than other, this can be due not only to a less discriminative ability of its formant frequencies but also to the fact that more ASR errors may occur for that particular unit. On the other hand, a variable ASR performance across speakers will reflect, in fact, the particularities of the different speakers, which will be combined with the different discriminative abilities of formant frequencies.

4.2. Types of constraints

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. ASR labels allow to define a large set of candidate constraints from linguistic units [16], showing each of them different characteristics in terms of within-unit formant dynamics, unit-length and frequency of occurrence. Among them, the following were used:

- **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 [37] in the phonetic transcriptions provided by the ASR system [34]. Table 1 shows the correspondence between Arpabet symbols and the International Phonetic Alphabet (IPA) ones, while Figure 3 shows an example of region conditioning for a particular phone unit.
- **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.

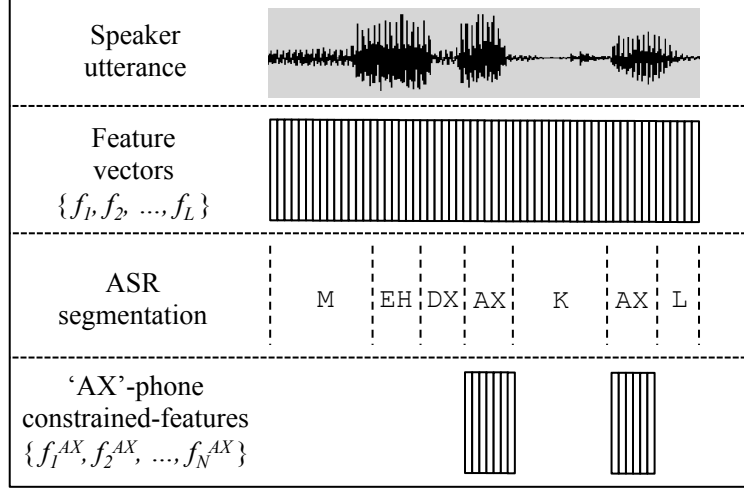


Figure 3: Example of region conditioning for a particular phone unit ('AX').

Although formants obtained from consonants are not regularly analysed within phonetics, we have not restricted the analysis to some specific units (e.g. only vowels, or vowels and some voiced consonants) for two main reasons. First, as the authors are neither linguists nor phoneticians but engineers, the only restriction applied regarding the linguistic units to be analysed is that they present enough frequency of occurrence. And secondly, working with a wide range of linguistic units illustrates the power of using automatic systems, providing a thorough analysis of their individualization potential.

4.3. Linguistically-constrained i-vector systems

An i-vector system [33] is a factor analysis (FA) based front-end for speaker verification which attempts to summarize the speaker distinguishing information in a given utterance, represented by a set of L feature vectors $\{f_1, f_2, \dots, f_L\}$, 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 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:

$$M = m + Tw \quad (3)$$

where T is a rectangular matrix of low rank defining the total variability (TV) space that contains the speaker and channel variability. For the purpose of developing linguistically-constrained systems, this FA model is applied in this work for every given constraint, C :

$$M^C = m^C + T^C w^C \quad (4)$$

Thus, independent UBMs and TV subspaces are trained on the background dataset (see Section 6 for details) from every linguistically-constrained set of feature vectors $\{f_1^C, f_2^C, \dots\}$,

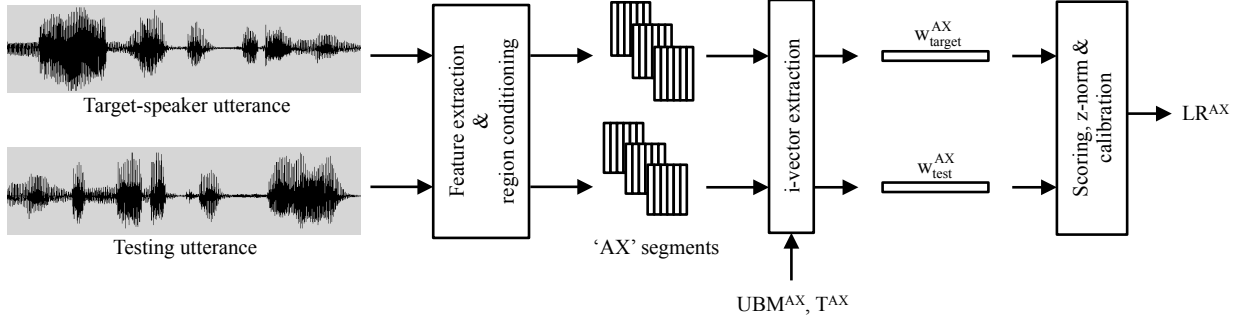


Figure 4: *Linguistically-constrained speaker verification system for a particular phone unit ('AX').*

allowing to obtain a constraint-dependent i-vector (w^C) from the occurrences of a given unit within an utterance (see Figure 4). 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 6 for details) for each linguistic-unit/constraint. Extracted i-vectors are length-normalized and whitened [38] previously to the scoring stage. Then, for a given constraint C , the similarity measure (score) between the target speaker (w_{target}^C) and the testing utterance (w_{test}^C) i-vectors is given by the cosine distance between them:

$$score(w_{target}^C, w_{test}^C) = \frac{\langle w_{target}^C, w_{test}^C \rangle}{\|w_{target}^C\| \|w_{test}^C\|} \quad (5)$$

Finally, constraint-dependent scores are z-normalized [39] and calibrated in an application-independent way [40] through *logistic regression* trained on the development dataset, using the FoCal toolkit [41].

5. Combination of linguistically-constrained systems

For a given speaker comparison, the final likelihood ratio can be either any of the constraint-dependent ones or a combination of a subset of them. In this Section, several strategies have been followed, regarding different aspects, in order to tackle the issue of how to combine the different linguistic constraints. First, the type of linguistic constraints taken into account has to be set. Then, some rule must be followed in order to select the particular constraints to be fused, according to some criterion. Finally, a specific fusion technique must be used in order to combine the likelihood ratios corresponding to different constraints.

5.1. Linguistic-constraint types

In a first stage, constraint combinations have been analysed separately for phone and diphone units. As diphone units are defined as two-phone combinations, they share the same information as phones, but spread over different diphones. However, dynamic information of the transition between two specific phone units is only modelled by diphone units, which may provide significant discrimination ability between speakers. Finally, constraint combinations

will be analysed when pooling together both types of linguistic units in order to test if the transitional information provided by diphone units provide additional discrimination ability to phone units.

5.2. Constraint-selection rules

We address the issue of constraints selection to be fused as a feature selection process [42], testing two constraint-selection schemes as in [17]:

- N-best performing units: for this method, constraints are sequentially fused in decreasing performance order on the development dataset. Once the EER is known for every number of constraints to be fused (see Figure 12a), the subset of constraints with the best performance on the development dataset is selected and applied in the evaluation dataset.
- Sequential Forward Selection (SFS): similarly to the previous method, constraints are sequentially fused in decreasing performance order on the development dataset. However, instead of keeping every subsequent constraint, they are included into the fusion subset only if the performance of the fused system increases. This procedure can be summarized in the following steps:
 1. Take the best-performing constraint as the initial subset.
 2. Take the next best-performing constraint and fuse with the previous subset. If the performance of the fused system is increased with respect to that of the previous step, keep the constraint; otherwise, reject it.
 3. Repeat the previous step until the worst performing constraint is reached.

5.3. Fusion techniques

Two different fusion techniques have been analysed in this work. First, a simple fusion rule consisting on averaging the log-LRs of the subset of N constraints to be combined has been applied through

$$\log LR = \frac{1}{N} \sum_{\forall C \text{ in subset}} \log LR^C \quad (6)$$

where $(\log LR^C)$ is the log-LR for a particular constraint C . While this technique do not take into account the different performance of the different constraints, it has the advantage of not requiring additional training data.

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

$$\log LR = \alpha_0 + \sum_{\forall C \text{ in subset}} \alpha^C \log LR^C \quad (7)$$

where the vector of weights $\alpha = [\alpha_0, \alpha^{C_1}, \alpha^{C_2}, \dots, \alpha^{C_N}]$ is obtained by *logistic regression* [43] training on the development database, using the FoCal toolkit [41].

For both fusion techniques, missing trials are handled in the same way as in [44]. Missing trials may appear when the corresponding constraint is not present in either target-speaker or testing utterances. In such cases, the corresponding sub-system cannot contribute a log-LR for that trial. However, as every linguistically-constrained system is independently calibrated, log-LRs of zero are inserted for missing trials in order to have valid log-LRs for every sub-system to train the fusion rule.

6. Experimental framework

One of the main goals of this work is to quantify the discriminative power of formant frequencies and their dynamics on the experimental frameworks used by the automatic speaker recognition community. NIST SREs have become a *de facto* standard for testing automatic speaker recognition systems, providing since 1997 [45] increasingly challenging datasets and protocols.

In order to develop and test the proposed speaker verification systems, we have used the datasets and protocols belonging to the NIST SREs carried out on years 2004 [46], 2005 [47] and 2006 [48], mainly those corresponding to the *core conditions*, which are composed of 5-minutes length telephone-line recordings of conversational speech. Among them, only English conversations have been used in order to match the characteristics of the ASR system [34].

Two are the main reasons for using only those years NIST SREs. First, the authors have access only to the ASR labels corresponding to those datasets, kindly provided by SRI. And second, the core condition of the NIST 2006 SRE is the main evaluation benchmark where a high number of comparative results are available from different high-level systems [10] [49] [12] [15] [13].

6.1. Performance evaluation metrics

The main evaluation metric used along this work to measure the discriminative performance is the equal error rate (EER) [45]. It is also used as the criterion by which the subsets of constraints are selected for the combination of systems. However, in accordance to the protocols used [48], the minimum of the C_{Det} (minDCF) is also shown. Finally, the C_{llr} cost function and the calibration loss (C_{llr}^{loss}) [40] are included as well in order to evaluate the calibration properties [50] of the different constraints and fusion schemes.

6.2. Background, development and evaluation datasets

The experimental protocol has been carefully designed in order to avoid obtaining overoptimistic results due to any overlap between datasets belonging to different development stages. With this aim, different datasets have been devoted to different purposes.

- Background: NIST 2004 SRE dataset [46] has been used as the background dataset for training UBMs and total variability matrices. This dataset 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. Also, speakers cohorts for Z-normalization were extracted from this dataset, using one 5 minute recording per speaker.

- Development: NIST 2005 SRE dataset [47] has been devoted to perform parameter optimization of the systems. Target speakers from the 1side-1side task were divided into two halves in order to have two different testing frameworks: *sre05-cal* and *sre05-val*, consisting both of them in $\sim 5,500$ male trials from ~ 120 target speakers and $\sim 7,400$ female trials from 171 target speakers. The number of both UBM components and dimensions of the TV subspace were optimized by minimizing the EER bias and variance over these two testing frameworks. Once the parameters of the system for each constraint were set, scores from *sre05-cal* were used to train the calibration process (logistic regression) and scores from *sre05-val* to train the fusion schemes.
- Evaluation: English-only trials from the core condition of the NIST 2006 SRE [48] 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.

6.3. Reference system

Our cepstral-based reference system is also an i-vector system developed by using the same experimental framework as the linguistically-constrained formant-based systems. It is based on mean-normalized, RASTA-filtered and gaussianized MFCC features (19 coefficients plus deltas). 1024-component UBMs and 600-dimensional TV subspaces were trained for each gender. Unlike for the formant-based system, LDA (trained on the background dataset) was applied in order to compensate for the intersession variability [33]. Thus, the similarity measure (score) between a target speaker (w_{target}) and a testing utterance (w_{test}) i-vectors is given by

$$score(w_{target}, w_{test}) = \frac{(A^t w_{target})(A^t w_{test})}{\sqrt{(A^t w_{target})(A^t w_{target})} \sqrt{(A^t w_{test})(A^t w_{test})}} \quad (8)$$

being A the LDA matrix. Finally, scores are z-normalized and calibrated in the same way as the linguistically-constrained systems.

7. Results

7.1. Independent linguistically-constrained systems

7.1.1. Overall performance per constraint

In this section we show the performance of each linguistically-constrained system independently. Table 2 shows the result for each metric on the evaluation dataset for the 10 best-performing phone-constraints (results for every phone are given in Table A.1), while Figure 5 shows the EER as a function of the frequency of occurrence for each of the 41 analysed phone-constraints. In both cases male and female trials are independently analysed; it can be seen that the constraints show similar behaviour for both genders in relative terms

NIST 2006 SRE, English-only trials									
Male					Female				
Phone	EER (%)	minDCF	C_{llr}	C_{llr}^{loss}	Phone	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 phone-constraints (extended results for every phone are given in Table A.1).

(Figure 5) except for the shift in absolute performance in favour of male speakers, which has been also reported in NIST SRE frameworks for cepstral-based systems [33].

It can be seen from Table 2 that, while each of the constraints have limited discriminative performance by themselves, they have good calibration properties. As an example, probability density functions of the logLRs provided for the best-performing phone-constraint

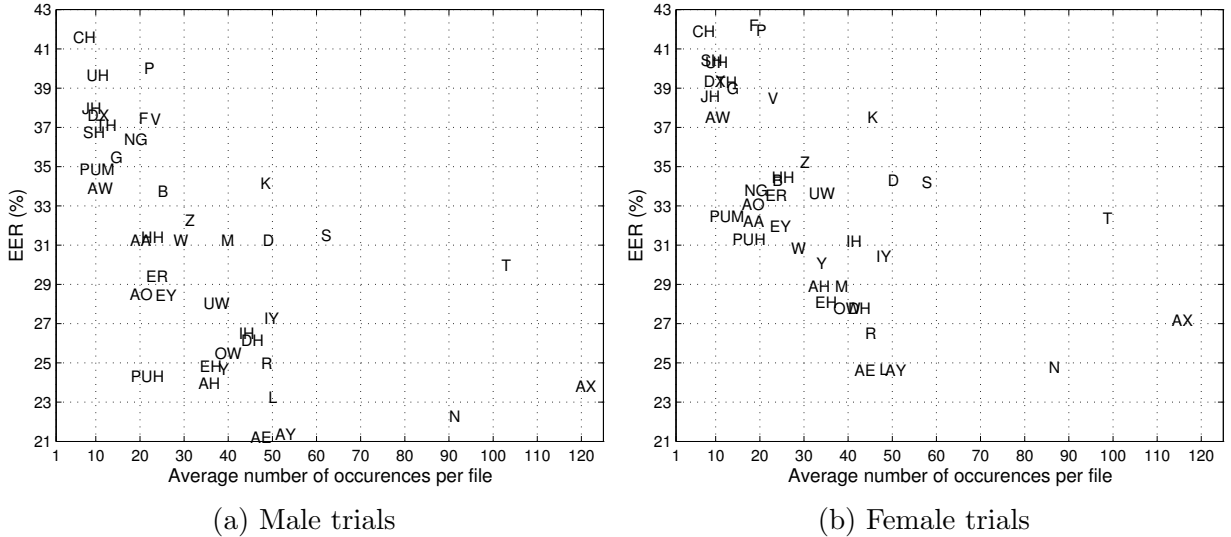


Figure 5: *EER vs frequency of occurrence for phone-constraints on the English-only trials of the core condition of the NIST 2006 SRE. Detailed frequency of occurrence in Table B.1.*

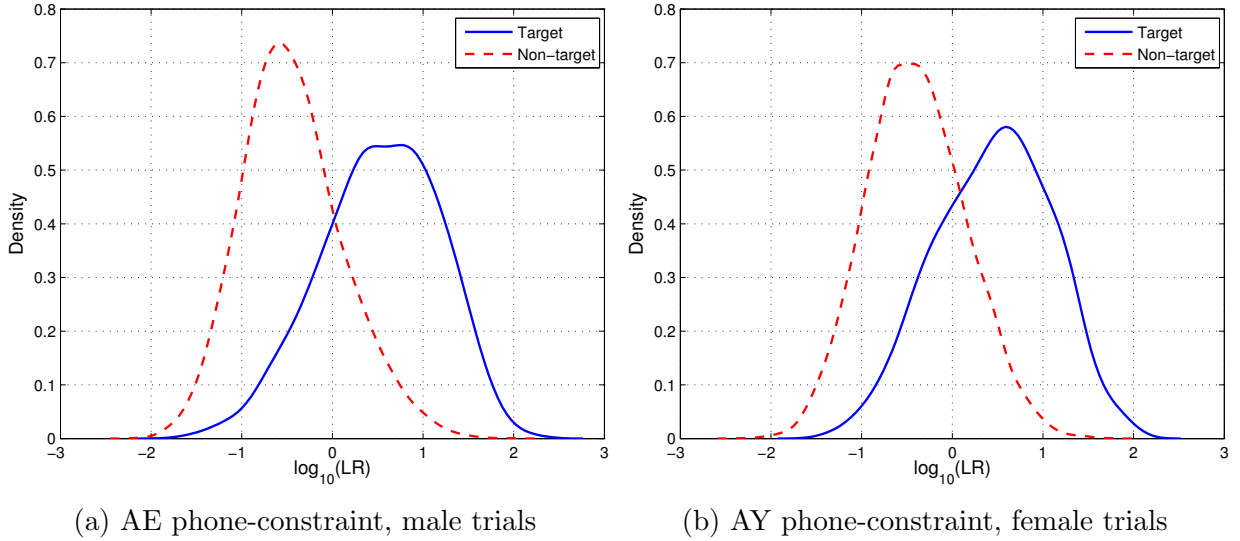


Figure 6: *Target and non-target \log_{10} -LRs probability density functions for the best phone-constrained system on the English-only trials of the core condition of the NIST 2006 SRE.*

are shown in Figure 6. This allows to obtain informative calibrated LR for voice comparisons from isolated linguistic units, as it has been suggested by some forensic-phonetics practitioners [24] [4].

Regarding the relationship between discriminative abilities and frequency of occurrence of each phone-constraint (Figure 5), there is a clear relationship between them in general terms, obtaining lower EERs those constraints with higher frequency of occurrence. However, for a subset of phone-constraints with similar frequency of occurrence, the range of EERs obtained may be wide, suggesting that different linguistic units present different discriminative abilities. In fact, some of the best performing units ('AE', 'AY', 'L', 'R') are not among those with the highest frequency of occurrence. However, it should be noted at this point that neither the formant tracking nor the ASR are error-free processes, and some particular phone units may present more errors than others, affecting to the relative difference in performance among them.

In the case of diphone-constraints, it can be seen from Table 3 that the best performing diphone-constraints are those combining some of the best performing phones (results for every diphone are given in Table A.2). This is a consequence of the combination of instantaneous frequency values with the derivative coefficients, which do not characterize the formant dynamics along the whole unit but in a local vicinity. However, there is not a clear relationship between performance and frequency of occurrence (Figure 7) unlike for phone-constraints, being in fact the best performing constraint, Y-AE, one of the least frequent in the database. This suggests that there is significant speaker-distinguishing information in formant dynamics in the transition between Y and AE phones: although these isolated phone-constraints are two of those with better performance, other two phone combinations among the 10-best performing phone-constraints obtain lower performance despite having a higher frequency of occurrence (e.g. AE-N or AX-N). However, it can be seen that, in aver-

NIST 2006 SRE, English-only trials									
Male					Female				
Diphone	EER (%)	minDCF	C_{llr}	C_{llr}^{loss}	Diphone	EER (%)	minDCF	C_{llr}	C_{llr}^{loss}
Y-AE	25.26	0.0867	0.7533	0.0245	Y-AE	28.32	0.0894	0.7960	0.0130
Y-UW	27.69	0.0928	0.8005	0.0135	AX-N	29.34	0.0942	0.8190	0.0154
AX-N	27.79	0.0947	0.7829	0.0132	L-AY	29.71	0.0943	0.8154	0.0199
AE-T	28.42	0.0987	0.8103	0.0195	N-OW	30.71	0.0957	0.8489	0.0159
L-AY	28.42	0.0957	0.8215	0.0277	AE-N	31.71	0.0962	0.8528	0.0097
DH-AE	28.82	0.0949	0.8269	0.0171	AE-T	31.90	0.0997	0.8614	0.0128
AE-N	28.88	0.0945	0.8132	0.0129	L-IY	32.20	0.1000	0.8728	0.0114
L-IY	30.15	0.0966	0.8331	0.0130	Y-UW	32.65	0.0948	0.8667	0.0114
N-D	31.80	0.0974	0.8469	0.0152	N-D	33.00	0.0978	0.8736	0.0082
N-OW	32.15	0.0959	0.8665	0.0128	S-OW	33.22	0.0996	0.8811	0.0116

Table 3: Results on the evaluation dataset for the 10 best-performing diphone-constraints (extended results for every diphone are given in Table A.2). Sample words for listed diphones are: yeah (Y-AE), you (Y-UW), second (AX-N), at (AE-T), like (L-AY), that (DH-AE), an (AE-N), firstly (L-IY), and (N-D), know (N-OW), so (S-OW).

age, diphone-constraints are less discriminative than phone-constraints due to their smaller average frequency of occurrence, although they also present good calibration properties (Table 3).

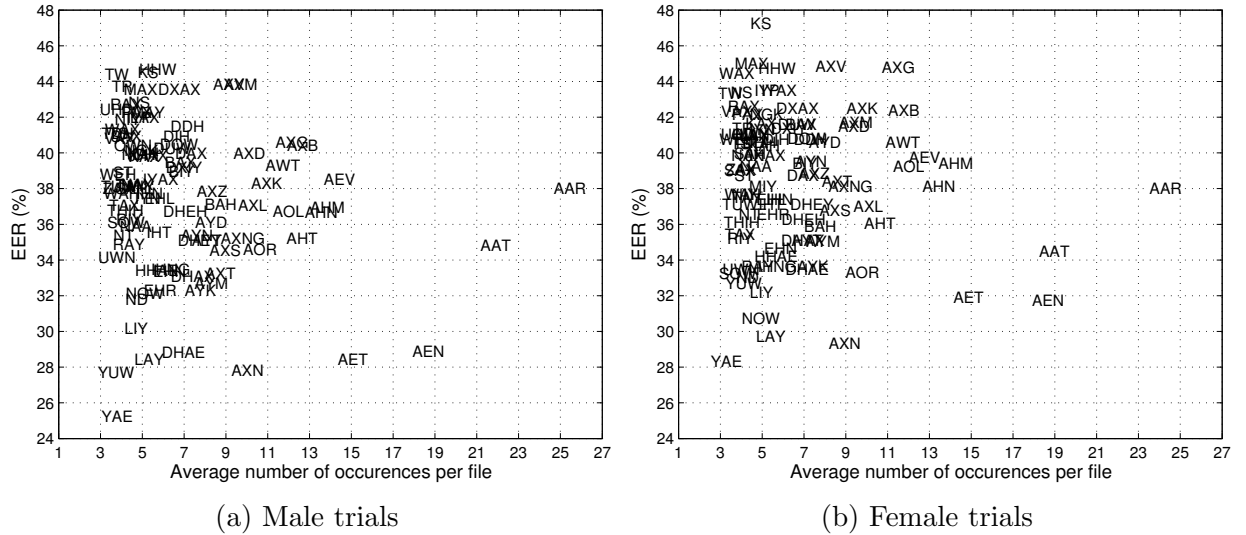
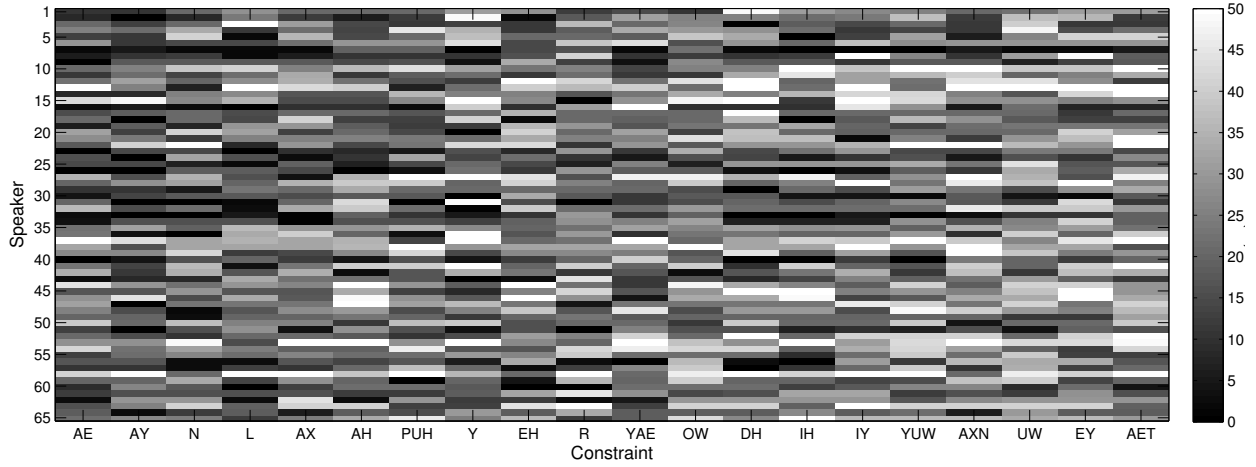
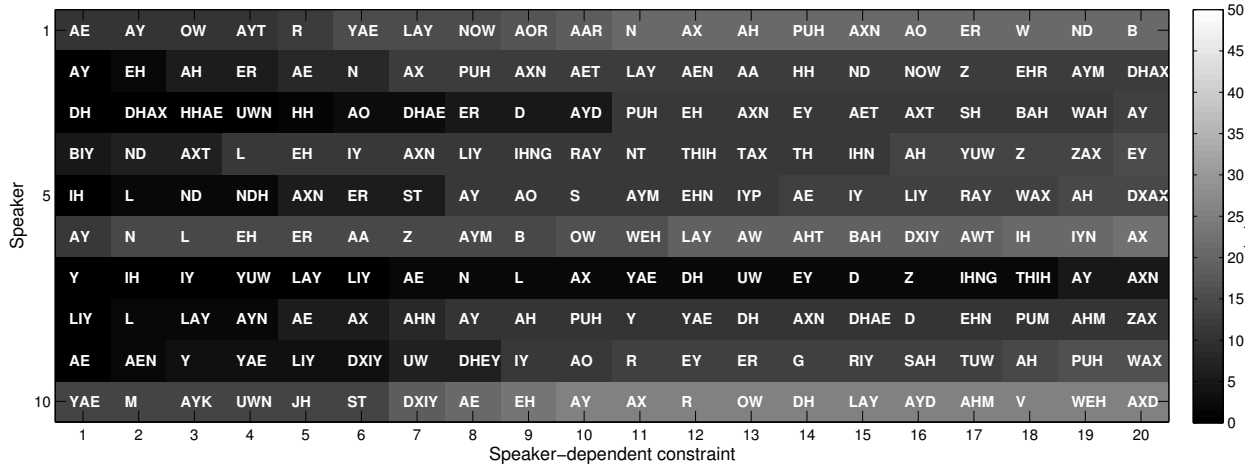


Figure 7: EER vs frequency of occurrence for diphone-constraints on the English-only trials of the core condition of the NIST 2006 SRE. Detailed frequency of occurrence in Table B.2.



(a) Constraints sorted by overall performance



(b) Constraints sorted by speaker-dependent performance (only 10 speakers are shown)

Figure 8: *EER (%) per speaker and constraint (only 20 first constraints are shown) on the English-only male trials of the core condition of the NIST 2006 SRE. In (a), the same unit (columns) performs very differently for different speakers. In (b), for every speaker (rows), the set and order of best constraints vary widely.*

7.1.2. Speaker-dependent performance of different constraints

It is also interesting to analyse how different constraints behave for different speakers, instead of the average behaviour per unit showed in the previous section. While both automatic formant tracking and ASR systems may present different behaviour for different speakers and units, this reflects, in fact, some speaker specificities that are combined with the discriminative abilities of formant frequencies. Figure 8a shows the EER per speaker for the 20-best performing constraints, sorted by overall performance on the evaluation dataset. As the EER has to be computed per each speaker, enough target trials per speaker are needed in order to obtain reliable metrics; with this aim, in this section only those speakers with at least 5 target trials have been used, yielding this 65 male speaker-set (only results for male speakers are shown in Figures 8-11, as similar conclusions can be drawn for female

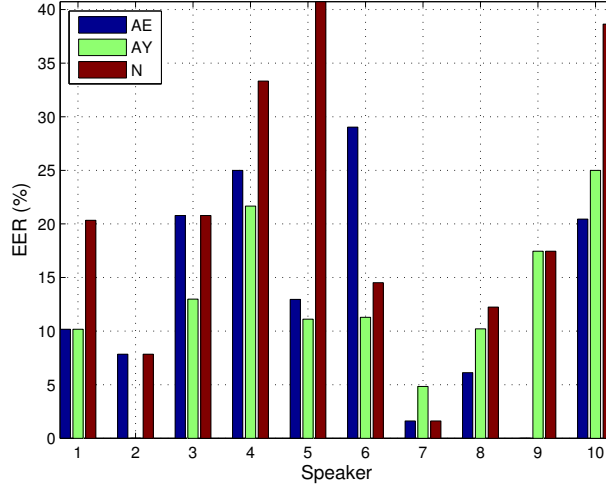


Figure 9: *EER (%) per speaker (only 10 speakers are shown) for the 3-best overall-performing constraints. Missing bars indicates that the EER is equal to zero.*

speakers).

This analysis shows (Figure 8a) that different constraints present different behaviour for different speakers. In fact, the best overall-performing constraint (the phone unit 'AE') may not be the best-performing one for a particular speaker, but even one of the worst-performing. For example, this constraint (first column in Figure 8a) presents a high EER (light grey) for speaker 13 while the performance is much better (dark grey) for speaker 14 and many others. Similarly, the constraint 'AE-T' (last column in Figure 8a), as having a much lower overall-performance (28.42% EER) than the constraint 'AE' (21.21% EER), presents a high EER (light grey) for several speakers, while it still presents a very good performance (dark

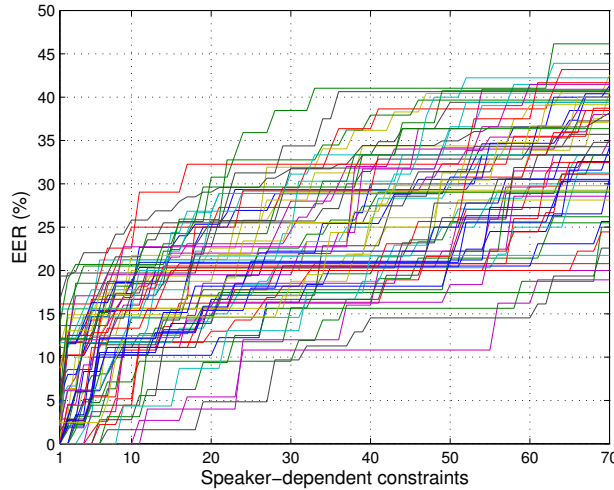


Figure 10: *EER (%) per constraint sorted by speaker-dependent performance for the 65 speakers (each line represents a different speaker). As shown, all 65 speakers have a subset of at least 10 speaker-dependent units with significant discriminative performance (EER per unit below 25%).*

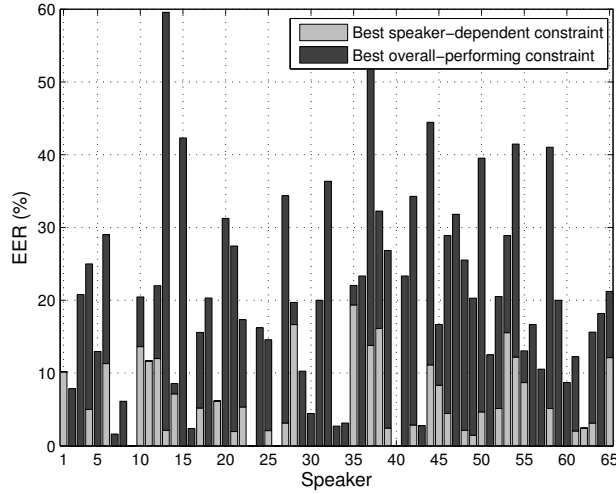


Figure 11: *EER (%) per speaker using the best overall constrained-system and the best speaker-dependent constrained-system. Missing bars indicates that the EER is equal to zero.*

grey) for some others. Similar information is shown in higher detail in Figure 9 in a slightly different way. Here, the EER of the 3-best overall-performing constraints ('AE', 'AY' and 'N') is represented for 10 different speakers, showing their highly variable performance from one speaker to another.

Moreover, when constraints are sorted by performance independently for each speaker, the set of N best-performing constraints can be very different from one speaker to another, as it can be seen in Figure 8b (where only 10 speakers and their 20 best-performing constraints are shown). This analysis also shows that a very good performance (low EER) could be achieved for most of the speakers if a speaker-dependent set of constraints is used, as it is shown in Figure 10 for all 65 speakers and their 70 best-performing constraints (conversely to Figure 8b, the particular constraints are not shown). It can be seen that, for every speaker, there is at least one constraint (and usually between 5 and 10 constraints) with better performance than the best overall-performing constraint (21.21% EER). Moreover, all 65 speakers have a subset of at least 10 speaker-dependent units with significant discriminative performance (EER per unit below 25%).

As an independent system is built for each isolated constraint in this approach, it would be possible to take advantage of this fact by using a different linguistically-constrained system for each speaker in order to adapt to his/her particular specificities if they were known in advance. For example, in the NIST 2012 SRE the target speakers were known in advance and several utterances per target speaker were provided; similar conditions may exist in real-life applications like access control or wiretapping. In such a case, the performance of the different constrained-systems could be analysed for each target speaker on a development dataset.

Figure 11 shows how the EER per speaker could be highly improved if the best constraint is selected in a speaker-dependent way instead of taking the best overall-performing constraint. While for these 65 speakers the average EER using the best overall-performing

constraint ('AE' phone) is 19.49%, the average EER using the best-performing constraint of each speaker would be 4.10%, a remarkable result as, for this speaker-set, the average EER of the reference cepstral system is 3.31%. Although this last result is optimistic as it is obtained knowing the best-performing speaker-dependent units over the evaluation dataset, it shows that improved results could be obtained adopting speaker-dependent strategies.

7.2. Performance of speaker-independent combinations of constraints

7.2.1. Comparison of fusion techniques

Figure 12a shows the EER of the fused system as a function of the number of fused constraints on the sre05-val development dataset for male trials for the two fusion techniques analysed in this work (namely, the average rule and logistic regression). While the EER of the fused system through the average rule obtains a minimum value for a certain number of fused constraints and then begin to increase, the EER of the fused system through logistic regression keeps going down as the number of fused constraints increases. The logistic regression fusion, being a trained fusion rule, benefits from the increasing amount of data provided by the additional constraints to be fused.

However, these are optimistic results as they are obtained in the development dataset, and the combination of constraints on the evaluation dataset can degrade if the performance of fused constraints varies from that obtained in development. This effect can be seen in Table 4. For the logistic regression technique, while the EER of the best fused system on the development dataset decreases as long as we take into account more constraints (from 41 phones to 41 phones + 98 diphones), the difference with the evaluation results increases, making them less robust to dataset mismatch for a large number of fused constraints. Conversely, the average fusion rule benefits from a higher number of constraints even in the case of dataset mismatch.

On the other hand, it can be seen also from Figure 12b that the calibration loss increases

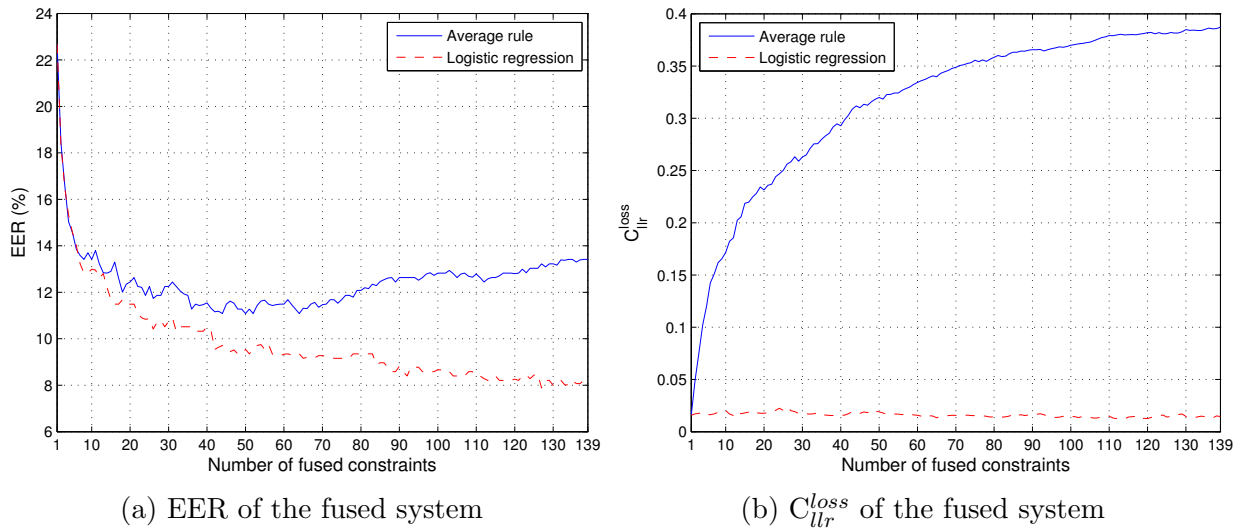


Figure 12: Comparison of fusion techniques on male trials of the sre05-val development dataset.

		Male/female EER (%) for the N-best rule		
		Phones (P)	Diphones (D)	P+D
Average	sre05-val	12.35 / 13.67	12.63 / 15.06	11.08 / 13.55
	SRE06	10.33 / 14.82	12.72 / 15.10	9.93 / 13.50
Log. Reg.	sre05-val	11.68 / 12.14	10.51 / 12.23	7.88 / 9.84
	SRE06	9.57 / 12.89	12.72 / 15.36	11.26 / 12.62

Table 4: *EER (%) for male/female trials in development and evaluation datasets when combining different types of linguistic units through the N-best rule.*

		Male/female EER (%) for the SFS rule		
		Phones (P)	Diphones (D)	P+D
Average	sre05-val	12.25 / 13.18	11.87 / 14.56	10.70 / 12.58
	SRE06	10.17 / 14.45	12.99 / 16.15	9.66 / 13.89
Log. Reg.	sre05-val	11.87 / 12.58	11.39 / 13.91	10.70 / 12.08
	SRE06	11.15 / 14.11	12.34 / 15.72	10.33 / 14.17

Table 5: *EER (%) for male/female trials in development and evaluation datasets when combining different types of linguistic units through the SFS rule.*

for the average fusion as the number of fused constraint increases, while it remains almost constant for the logistic regression. This makes the logistic regression the preferred fusion option as eliciting calibrated LR is among our main objectives.

7.2.2. Comparison of constraint-selection strategies

Table 5 shows the results for the SFS constraint-selection strategy as Table 4 does for the N-best one. It can be seen that both strategies give similar results on the evaluation dataset for the average fusion rule, being the EER of the fused systems reduced when constraints from different linguistic-unit types (phones and diphones) are combined. However, in the case of the logistic regression fusion technique, there is no such gain for the N-best strategy on male trials and slight differences on female trials due to the over-fitting and database mismatch between development and evaluation datasets observed in the previous section. The SFS strategy does not suffer from this over-fitting as it does not select a number of constraints as high as the N-best strategy, as constraints that do not increase the performance of the fused system are discarded. In this way, it still benefits from incorporating diphone units, which can provide additional dynamic information present in the transition between phone units.

Finally, Table 6 shows the performance on different evaluation metrics for the best combinations of constraint-selection strategies and fusion techniques. In this table, we can see that logistic regression technique has the advantage of providing well calibrated likelihood

		Male/female results on the NIST 2006 SRE, English-only trials			
		EER (%)	minDCF	C_{llr}	C_{llr}^{loss}
SFS	Phones	10.17 / 14.45	0.0495 / 0.0585	0.6759 / 0.6928	0.3069 / 0.2229
Average	P+D	9.66 / 13.89	0.0463 / 0.0585	0.8163 / 0.8439	0.4741 / 0.3975
N-best	Phones	9.57 / 12.89	0.0456 / 0.0543	0.3742 / 0.4361	0.0277 / 0.0117
Log. reg.	P+D	11.26 / 12.62	0.0503 / 0.0590	0.4046 / 0.4531	0.0317 / 0.0202

Table 6: Comparison of the best combinations between constraint-selection strategies and fusion techniques on the evaluation dataset.

ratios also on the evaluation dataset, as we saw in Figure 12 for the development dataset. Being this a highly desirable property, the following analysis in Section 7.3 focus on the best combination of constraints through logistic regression, which is the one using N-best selection from phone-constraints. In order to highlight the calibration properties of the elicited LRs from the best formant-based fused system, in Figure 13 we show the \log_{10} LR target and non-target probability density functions.

7.3. Fusion of formant- and cepstral-based systems

Table 7 show the results on the evaluation dataset for the best formant-based fused system (that using logistic regression fusion of the N-best selected phone units), for the cepstral-based reference system, and for the average fusion of both. For female trials, although the EER of the fused system is almost the same, there are significant improvements

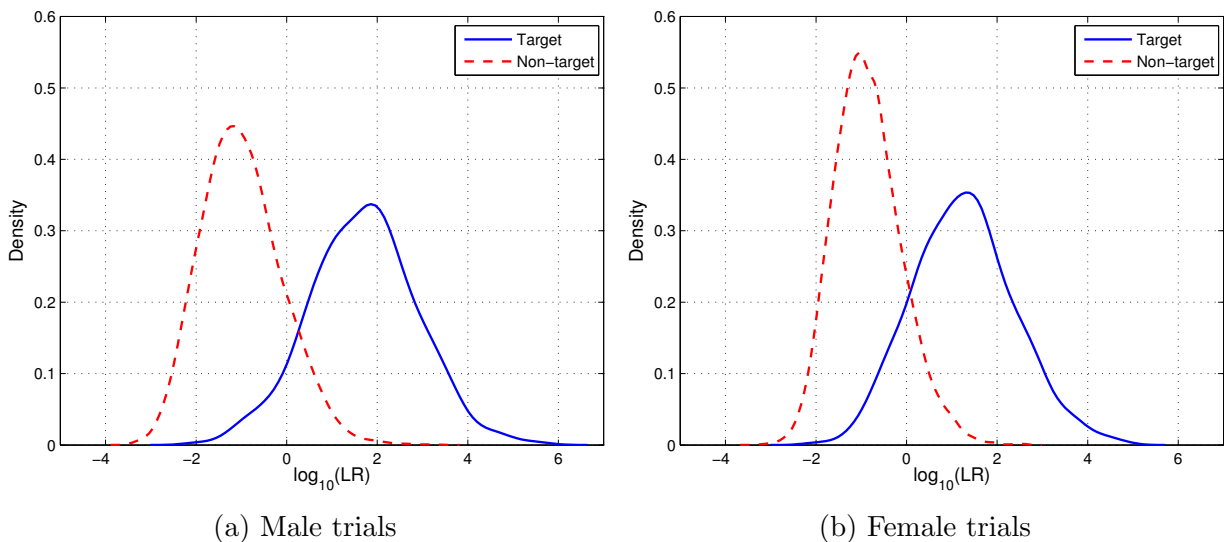


Figure 13: Target and non-target \log_{10} -LRs probability density functions for the best formant-based fused system on the English-only trials of the core condition of the NIST 2006 SRE.

	Male/female results on the NIST 2006 SRE, English-only trials			
	EER (%)	minDCF	C_{llr}	C_{llr}^{loss}
Formant-based	9.57 / 12.89	0.0456 / 0.0543	0.3742 / 0.4361	0.0277 / 0.0117
Cepstral-based	6.21 / 6.87	0.0303 / 0.0352	0.2293 / 0.2927	0.0232 / 0.0321
Average fusion	5.41 / 6.86	0.0248 / 0.0311	0.2179 / 0.2789	0.0368 / 0.0393

Table 7: Results on the evaluation dataset for the best formant-based fused system, the cepstral-based reference system and the average fusion of both.

in both minDCF and C_{llr} metrics. In the case of male trials, there is also a relative improvement of $\sim 18\%$ in terms of the EER. Considering both genders, the fused system obtains relative improvements of 7% and 15% in terms of gender-averaged EER and minDCF, respectively. Although both approaches are based on spectral features, it is shown that they present a high complementarity like other high-level approaches on the same evaluation framework [49] [12] [15].

7.4. Comparison with other higher-level systems

Finally, an objective comparison of different high-level approaches in the same evaluation framework (core condition of the NIST 2006 SRE) is given in Table 8, extending that presented in [10] with some later works, sorted by performance.

The best performing systems are those based on cepstral information (1,2), using either cepstral-derived features (coefficients from MLLR transforms between cepstral-based GMMs) or MFCC (and prosodic) contours, where ASR is used either only for feature extraction (2) or also for region conditioning (1). Then, there is a group of systems based on several prosodic (usually including energy, pitch and duration) and/or formant features (3-7), most of them having very similar performance (ranging from 10.41% to 11.9% EER for the four best performing ones). Next two systems are based only on duration information: (8) models the number of frames of the three states in phone HMMs, while (9) directly models the duration of phones within specific words. Finally, system (10) is a lexico-prosodic approach with similar performance to (9).

Among them, our approach is the only one based only on formant frequencies and where feature extraction does not rely on ASR labels, which are only used for region conditioning. Also, it is worth noting that our formant-based system does not include NIST 2005 SRE in the background dataset in order to avoid using overoptimistic scores in the calibration training; in this way, it is possible to obtain well calibrated LRs per constraint, but better discriminative performance may be achieved using a richer and larger background database for UBM and total variability training. However, being its features obtained from short-term windows every 10 ms, system parameters can be properly trained on limited background data.

Male+female results on the NIST 2006 SRE, English-only trials		
System (feature type and model)	EER (%)	Reference
(1) Cepstral-derived MLLR SVM	4.00	[51]
(2) Prosodic and MFCC contours JFA	7.66	[15]
(3) Syllable-based prosody sequence SVM	10.41	[52], [53]
(4) Prosodic contours JFA	11.00	[13]
(5) Formants+Δs i-vector	11.23	-
(6) Formant and prosodic contours JFA	11.9	[12]
(7) Prosodic contours JFA	14.6	[49]
(8) State-in-phone-duration GMM	16.02	[54]
(9) Phone-in-word-duration GMM	22.22	[54]
(10) Duration-conditioned word N-gram SVM	23.46	[55]

Table 8: *Results on the core condition of the NIST 2006 SRE (English-only trials) for several high-level systems compared to our formant-based approach (5).*

8. Conclusions and future work

In this work, we have explored the discriminative abilities of formant frequencies and their dynamics within linguistic units through fully-automatic linguistically-constrained i-vector systems.

Automatic formant tracking have been used for feature extraction, and dynamic information is included through derivative coefficients. In this way, it is possible to combine both static and dynamic information of formant frequencies while maintaining the frame-by-frame feature observation rate, instead of reducing each constraint to a single observation feature vector as it is done in some approaches that code the whole trajectory within a speech region. This procedure allows us to robustly train the parameters of the system even with limited background data (NIST SRE 2004) compared with similar higher-level approaches based on coded trajectories, as it has been shown in Section 7.4.

Then, ASR is used in order to constrain the set of features to be used by each subsystem, corresponding each of them to a different linguistic unit among two main groups: phones and diphones. For each of such constraints, one independent i-vector system is developed. Although linguistically-constrained systems have limited performance by themselves, we have shown that well calibrated log-likelihood ratios can be provided for each linguistic unit. Regarding the relative differences in performance among units, it should be noted that they can be due not only to the different discriminative abilities of their formant frequencies but also to a different behaviour of the automatic systems involved in the feature extraction (formant tracking) and region conditioning (ASR labels) processes, which may lead to a non-uniform distribution of errors among different units. It would be of broad interest to perform an equivalent analysis in a manually labelled database in order to avoid the effect of the errors introduced by these automatic systems, but large datasets of spontaneous conversational speech as those used in this work ($\sim 10,000$ 5-minute conversations) seem unlikely to be manually annotated (both formant frequencies and phonetic transcriptions).

On the other hand, a different behaviour of the formant tracking and ASR systems across speakers for a particular unit is considered to reflect the specificities of the different speakers.

This fine-grained detail provided by linguistically-constrained systems can be exploited through speaker-dependent strategies when selecting the constraints to be used. For example, in Section 7.1.2 it has been shown that using only the best-performing speaker-dependent constraint instead of the best overall-performing one for every speaker, the average EER in the analysed speaker set improves from 19.49% to 4.10%. Furthermore, most of the speakers in the analysed set presents a subset of several constraints (usually between 5 and 10) that perform better than the overall-performing constraint, so using any of those (different) constraints for every speaker will lead to an overall performance improvement. Although this is an optimistic result as it is obtained knowing the best-performing speaker-dependent units over the evaluation dataset, it shows that improved results could be obtained adopting speaker-dependent strategies. As a future work, some of this strategies would be tested on an experimental framework that allows to estimate in advance the best speaker-dependent set of linguistic units to be used for the different target speakers.

Moreover, we have presented several speaker-independent constraint-combination strategies in order to integrate the speaker distinguishing information spread over the different linguistic units, achieving for some of them a remarkable combined performance taking into account the limited size of the background dataset and the nature of features being used. For these fused systems, discriminative and well calibrated log-likelihood ratios are also provided.

Finally, significant improvements have been achieved by combining these formant-based systems with a cepstral-based reference system, showing the complementarity of cepstral and formant-based approaches.

9. Acknowledgements

This work has been supported by the Spanish Ministry of Economy and Competitiveness (project CMC-V2: Caracterización, Modelado y Compensación 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.

References

- [1] T. Kinnunen and H. Li, “An overview of text-independent speaker recognition: From features to supervectors,” *Speech Communication*, vol. 52, no. 1, pp. 12–40, 2010.
- [2] J. Darch, B. Milner, X. Shao, S. Vaseghi, and Q. 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.
- [3] F. Nolan and C. Grigoras, “A case for formant analysis in forensic speaker identification,” *International Journal of Speech Language and the Law*, vol. 12, no. 2, 2005.
- [4] P. Rose, *Forensic Speaker Identification*, ser. Forensic Science. Taylor and Francis, 2002.

- [5] T. Becker, M. Jessen, and C. Grigoros, “Forensic speaker verification using formant features and gaussian mixture models,” in *Proceedings of the 9th Annual Conference of the International Speech Communication Association (INTERSPEECH 2008)*, Brisbane, Australia, September 22-26, 2008, 2008, pp. 1505–1508.
- [6] J. Gonzalez-Rodriguez, P. Rose, D. Ramos, D. Toledano, and J. Ortega-Garcia, “Emulating DNA: Rigorous Quantification of Evidential Weight in Transparent and Testable Forensic Speaker Recognition,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 7, pp. 2104–2115, Sept 2007.
- [7] F. Nolan, *The phonetic bases of speaker recognition*. Cambridge (UK): Cambridge University Press, 1983.
- [8] K. 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.
- [9] J. Gonzalez-Rodriguez, J. Gil, R. Pérez, and J. 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.
- [10] E. Shriberg, “Higher-level features in speaker recognition,” in *Speaker Classification I. Fundamentals, Features, and Methods*, ser. Lecture Notes in Computer Science, vol. 4343. Springer Berlin Heidelberg, 2007, pp. 241–259.
- [11] T. Bocklet and E. 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.
- [12] N. Dehak, P. Kenny, and P. 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.
- [13] M. Kockmann, L. Burget, and J. Cernocký, “Investigations into prosodic syllable contour features for speaker recognition,” in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2010)*, 14-19 March 2010, Sheraton Dallas Hotel, Dallas, Texas, USA, 2010, pp. 4418–4421.
- [14] D. A. Reynolds, W. D. Andrews, J. P. Campbell, J. Navrátil, B. Peskin, A. Adami, Q. Jin, D. Klusacek, J. S. Abramson, R. Mihaescu, J. J. Godfrey, D. A. Jones, and B. Xiang, “The supersid project: exploiting high-level information for high-accuracy speaker recognition,” in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2003)*, Hong Kong, April 6-10, 2003, 2003, pp. 784–787.
- [15] M. Kockmann and L. Burget, “Contour modeling of prosodic and acoustic features for speaker recognition,” in *Proceedings of the IEEE Spoken Language Technology Workshop (SLT 2008)*, Dec 2008, pp. 45–48.
- [16] J. 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.
- [17] J. Franco-Pedroso, F. Espinoza-Cuadros, and J. 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.
- [18] S. Furui, “Cepstral analysis technique for automatic speaker verification,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 29, no. 2, pp. 254–272, Apr 1981.
- [19] F. Soong and A. Rosenberg, “On the use of instantaneous and transitional spectral information in speaker recognition,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 36, no. 6, pp. 871–879, Jun 1988.
- [20] J. Benesty, M. M. Sondhi, and Y. Huang, Eds., *Springer Handbook of Speech Processing*. Berlin:

- Springer, 2008.
- [21] P. Rose and E. 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 2010)*, December 2010, pp. 42–45.
 - [22] C. Zhang, G. S. Morrison, and P. Rose, "Forensic speaker recognition in chinese: a multivariate likelihood ratio discrimination on /i/ and /y/," in *Proceedings of the 9th Annual Conference of the International Speech Communication Association (INTERSPEECH 2008)*, Brisbane, Australia, September 22–26, 2008, 2008, pp. 1937–1940.
 - [23] F. Nolan, "The 'telephone effect' on formants: a response," *International Journal of Speech Language and the Law*, vol. 9, no. 1, 2002.
 - [24] G. S. Morrison, "Likelihood-ratio-based forensic speaker comparison using parametric representations of vowel formant trajectories," *Journal of the Acoustical Society of America*, no. 125, pp. 2387–2397, 2009.
 - [25] K. Sjölander and J. 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.
 - [26] P. Boersma and D. Weenink, "Praat: doing phonetics by computer." [Online]. Available: <http://www.praat.org/>
 - [27] Y. Laprie, "Winsnoori 1.34 – free software for speech analysis," 2014. [Online]. Available: <http://www.loria.fr/~laprie/WinSnoori/>
 - [28] "Snack Sound Toolkit — Wikipedia, The Free Encyclopedia," 2014. [Online]. Available: http://en.wikipedia.org/wiki/Snack_Sound_Toolkit
 - [29] "Tcl — Wikipedia, The Free Encyclopedia," 2015. [Online]. Available: <http://en.wikipedia.org/wiki/Tcl>
 - [30] "Snack v2.2.8 manual." [Online]. Available: <http://www.speech.kth.se/snack/man/snack2.2/tcl-man.html>
 - [31] "Arpabet — Wikipedia, The Free Encyclopedia," 2014. [Online]. Available: <http://en.wikipedia.org/wiki/Arpabet>
 - [32] A. Stolcke, E. Shriberg, L. Ferrer, S. Kajarekar, K. Sonmez, and G. Tur, "Speech recognition as feature extraction for speaker recognition," in *Proceedings of the IEEE Workshop on Signal Processing Applications for Public Security and Forensics (SAFE 2007)*, April 2007, pp. 1–5.
 - [33] 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.
 - [34] S. S. Kajarekar, N. Scheffer, M. Graciarena, E. Shriberg, A. Stolcke, L. Ferrer, and T. 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.
 - [35] L. Rabiner and B. Juang, "Speech recognition: Statistical methods," in *Encyclopedia of Language and Linguistics*, K. Brown, Ed. Elsevier, 2006.
 - [36] S. Goldwater, D. Jurafsky, and C. D. Manning, "Which words are hard to recognize? Prosodic, lexical, and disfluency factors that increase speech recognition error rates," *Speech Communication*, vol. 52, no. 3, pp. 181–200, 2010.
 - [37] J. E. Shoup, "Phonological aspects of speech recognition," in *Trends in Speech Recognition*, W. A. Lea, Ed. Englewood Cliffs: Prentice Hall, 1980, pp. 125–138.
 - [38] D. Garcia-Romero and C. 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.
 - [39] R. Auckenthaler, M. J. Carey, and H. Lloyd-Thomas, "Score normalization for text-independent speaker verification systems," *Digital Signal Processing*, vol. 10, no. 1–3, pp. 42–54, 2000.
 - [40] N. Brümmer and J. du Preez, "Application-independent evaluation of speaker detection," in *Computer*

- Speech and Language*, vol. 20, no. 2 - 3, 2006, pp. 230 – 275.
- [41] N. Brümmer, “Toolkit for evaluation, fusion and calibration of statistical pattern recognizers.” [Online]. Available: <https://sites.google.com/site/nikobrummer/focal>
 - [42] A. K. Jain, R. P. W. Duin, and J. Mao, “Statistical pattern recognition: A review,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 1, pp. 4–37, 2000.
 - [43] S. Pigeon, P. Druyts, and P. 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.
 - [44] N. Brümmer, L. Burget, J. Cernocký, O. Glembek, F. Grézl, M. Karafiát, D. A. van Leeuwen, P. Matejka, P. Schwarz, and A. Strasheim, “Fusion of heterogeneous speaker recognition systems in the STBU submission for the NIST speaker recognition evaluation 2006,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 7, pp. 2072–2084, 2007.
 - [45] J. 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.
 - [46] “The NIST Year 2004 Speaker Recognition Evaluation Plan.” [Online]. Available: http://www.itl.nist.gov/iad/mig/tests/spk/2004/SRE-04_evalplan-v1a.pdf
 - [47] “The NIST Year 2005 Speaker Recognition Evaluation Plan.” [Online]. Available: http://www.itl.nist.gov/iad/mig/tests/sre/2005/sre-05_evalplan-v6.pdf
 - [48] “The NIST Year 2006 Speaker Recognition Evaluation Plan.” [Online]. Available: http://www.itl.nist.gov/iad/mig/tests/spk/2006/sre-06_evalplan-v9.pdf
 - [49] 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.
 - [50] D. van Leeuwen and N. Brümmer, “An Introduction to Application-Independent Evaluation of Speaker Recognition Systems,” in *Speaker Classification I*, ser. Lecture Notes in Computer Science, C. Müller, Ed. Springer Berlin Heidelberg, 2007, vol. 4343, pp. 330–353.
 - [51] A. Stolcke, L. Ferrer, S. S. Kajarekar, E. Shriberg, and A. Venkataraman, “MLLR transforms as features in speaker recognition,” in *Proceedings of the 9th European Conference on Speech Communication and Technology (INTERSPEECH 2005 - EUROSPEECH)*, Lisbon, Portugal, September 4-8, 2005, 2005, pp. 2425–2428.
 - [52] L. Ferrer, E. Shriberg, S. S. Kajarekar, and M. K. Sönmez, “Parameterization of prosodic feature distributions for SVM modeling in speaker recognition,” in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2007)*, Honolulu, Hawaii, USA, April 15-20, 2007, 2007, pp. 233–236.
 - [53] E. Shriberg and L. Ferrer, “A text-constrained prosodic system 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. 1226–1229.
 - [54] L. Ferrer, H. Bratt, V. R. R. Gadde, S. S. Kajarekar, E. Shriberg, M. K. Sönmez, A. Stolcke, and A. Venkataraman, “Modeling duration patterns for speaker recognition,” in *Proceedings of the 8th European Conference on Speech Communication and Technology (INTERSPEECH 2003 - EUROSPEECH)*, Geneva, Switzerland, September 1-4, 2003, 2003.
 - [55] G. Tür, E. Shriberg, A. Stolcke, and S. S. Kajarekar, “Duration and pronunciation conditioned lexical modeling 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. 2049–2052.

Appendix A. Extended results

Table A.1: *Results on the evaluation dataset for every phone-constraint.*

NIST 2006 SRE, English-only trials									
Male					Female				
Phone	EER (%)	minDCF	C_{llr}	C_{llr}^{loss}	Phone	EER (%)	minDCF	C_{llr}	C_{llr}^{loss}
AA	31.21	0.0971	0.8397	0.0112	AA	32.20	0.0982	0.8707	0.0113
AE	21.21	0.0850	0.6668	0.0143	AE	24.59	0.0876	0.7308	0.0131
AH	23.96	0.0964	0.7286	0.0158	AH	28.89	0.0974	0.8185	0.0079
AO	28.47	0.0996	0.8264	0.0151	AO	33.08	0.0989	0.8767	0.0115
AW	33.88	0.0986	0.8907	0.0108	AW	37.48	0.1000	0.9352	0.0075
AX	23.80	0.0844	0.7001	0.0150	AX	27.15	0.0932	0.7764	0.0100
AY	21.38	0.0825	0.6580	0.0158	AY	24.59	0.0841	0.7101	0.0111
B	33.74	0.0956	0.8774	0.0141	B	34.28	0.0988	0.8851	0.0090
CH	41.57	0.0997	0.9665	0.0102	CH	41.87	0.0999	0.9757	0.0061
D	31.21	0.0951	0.8333	0.0130	D	34.28	0.0982	0.8868	0.0097
DH	26.16	0.0932	0.7514	0.0125	DH	27.79	0.0940	0.7876	0.0114
DX	37.59	0.0997	0.9289	0.0134	DX	39.35	0.1000	0.9613	0.0093
EH	24.83	0.0972	0.7544	0.0140	EH	28.06	0.0990	0.8196	0.0157
ER	29.40	0.0996	0.8409	0.0157	ER	33.53	0.0996	0.8816	0.0144
EY	28.42	0.0995	0.8164	0.0132	EY	31.96	0.0978	0.8608	0.0080
F	37.47	0.0986	0.9288	0.0095	F	42.22	0.0995	0.9701	0.0109
G	35.45	0.1000	0.9126	0.0137	G	38.97	0.0999	0.9416	0.0078
HH	31.37	0.0980	0.8624	0.0149	HH	34.46	0.0980	0.8933	0.0112
IH	26.48	0.0931	0.7736	0.0174	IH	31.18	0.0998	0.8554	0.0160
IY	27.25	0.0996	0.7854	0.0151	IY	30.44	0.0975	0.8384	0.0128
JH	37.96	0.0994	0.9338	0.0151	JH	38.58	0.1000	0.9457	0.0098
K	34.13	0.0979	0.8779	0.0101	K	37.51	0.0994	0.9263	0.0109
L	23.24	0.0839	0.7083	0.0133	L	24.68	0.0869	0.7355	0.0127
M	31.21	0.0961	0.8445	0.0114	M	28.89	0.0944	0.8169	0.0121
N	22.26	0.0812	0.6896	0.0168	N	24.77	0.0839	0.7256	0.0112
NG	36.40	0.0975	0.9139	0.0165	NG	33.77	0.0991	0.8809	0.0110
OW	25.49	0.0927	0.7445	0.0152	OW	27.79	0.0936	0.7830	0.0098
P	39.99	0.0998	0.9439	0.0082	P	41.96	0.0999	0.9732	0.0069
PUH	24.32	0.0933	0.7296	0.0137	PUH	31.28	0.0976	0.8420	0.0097
PUM	34.86	0.0983	0.8985	0.0217	PUM	32.46	0.0984	0.8764	0.0142
R	24.96	0.0937	0.7380	0.0149	R	26.49	0.0932	0.7681	0.0132
S	31.48	0.0947	0.8390	0.0095	S	34.18	0.0963	0.8840	0.0123
SH	36.73	0.1000	0.9255	0.0121	SH	40.40	0.0998	0.9571	0.0122
T	29.98	0.0925	0.8247	0.0161	T	32.37	0.0959	0.8624	0.0095
TH	37.10	0.0978	0.9387	0.0159	TH	39.31	0.1000	0.9506	0.0054
UH	39.64	0.0999	0.9471	0.0075	UH	40.31	0.1000	0.9593	0.0106
UW	28.02	0.0950	0.8016	0.0121	UW	33.63	0.0993	0.8980	0.0121
V	37.39	0.0998	0.9286	0.0106	V	38.49	0.0999	0.9456	0.0082
W	31.21	0.0948	0.8307	0.0111	W	30.81	0.0973	0.8509	0.0177
Y	24.68	0.0915	0.7325	0.0180	Y	30.08	0.0923	0.8303	0.0124
Z	32.27	0.0952	0.8658	0.0150	Z	35.20	0.0997	0.9083	0.0091

Table A.2: *Results on the evaluation dataset for every diphone-constraint.*

NIST 2006 SRE, English-only trials									
Male					Female				
Diphone	EER (%)	minDCF	C_{llr}	C_{llr}^{loss}	Diphone	EER (%)	minDCF	C_{llr}	C_{llr}^{loss}
AA-R	37.99	0.1000	0.9414	0.0085	AA-R	38.03	0.0988	0.9355	0.0123
AA-T	34.80	0.0982	0.9027	0.0175	AA-T	34.46	0.1000	0.9020	0.0122
AE-N	28.88	0.0945	0.8132	0.0129	AE-N	31.71	0.0963	0.8529	0.0098
AE-T	28.42	0.0987	0.8103	0.0195	AE-T	31.90	0.0997	0.8614	0.0128
AE-V	38.53	0.1000	0.9433	0.0131	AE-V	39.77	0.0998	0.9554	0.0078
AH-M	36.93	0.0996	0.9249	0.0120	AH-M	39.38	0.1000	0.9505	0.0090
AH-N	36.67	0.0996	0.9320	0.0102	AH-N	38.10	0.1000	0.9419	0.0124
AH-T	35.20	0.0986	0.9040	0.0137	AH-T	36.02	0.0996	0.9101	0.0102
AO-L	36.73	0.1000	0.9312	0.0101	AO-L	39.22	0.1000	0.9530	0.0092
AO-R	34.60	0.0992	0.8978	0.0149	AO-R	33.28	0.0990	0.8909	0.0103
AW-T	39.32	0.0980	0.9406	0.0190	AW-T	40.59	0.0999	0.9600	0.0095
AX-B	40.39	0.0998	0.9593	0.0118	AX-B	42.38	0.1000	0.9727	0.0102
AX-D	39.94	0.1000	0.9668	0.0105	AX-D	41.50	0.0999	0.9695	0.0070
AX-G	40.57	0.1000	0.9682	0.0088	AX-G	44.77	0.0997	0.9825	0.0078
AX-K	38.26	0.0991	0.9391	0.0102	AX-K	42.51	0.1000	0.9678	0.0077
AX-L	37.07	0.0985	0.9327	0.0123	AX-L	37.00	0.1000	0.9297	0.0083
AX-M	43.84	0.0999	0.9930	0.0204	AX-M	41.71	0.0999	0.9724	0.0068
AX-N	27.79	0.0947	0.7829	0.0132	AX-N	29.34	0.0942	0.8190	0.0154
AX-NG	35.18	0.1000	0.9294	0.0154	AX-NG	38.12	0.1000	0.9321	0.0080
AX-S	34.53	0.0983	0.8957	0.0149	AX-S	36.77	0.0999	0.9240	0.0098
AX-T	33.26	0.1000	0.8946	0.0126	AX-T	38.37	0.0999	0.9482	0.0111
AX-V	43.84	0.1000	0.9835	0.0084	AX-V	44.84	0.0999	0.9873	0.0049
AX-Z	37.86	0.1000	0.9430	0.0113	AX-Z	38.85	0.1000	0.9523	0.0094
AY-D	36.13	0.0987	0.9185	0.0157	AY-D	40.59	0.0996	0.9486	0.0074
AY-K	32.27	0.0976	0.8699	0.0153	AY-K	33.65	0.0988	0.8825	0.0119
AY-M	32.67	0.0970	0.8616	0.0122	AY-M	35.01	0.0993	0.9157	0.0161
AY-N	35.37	0.0988	0.9096	0.0139	AY-N	39.50	0.1000	0.9493	0.0078
AY-T	35.07	0.0989	0.9006	0.0154	AY-T	35.09	0.0998	0.9146	0.0129
B-AH	37.12	0.0993	0.9309	0.0148	B-AH	35.90	0.0998	0.9210	0.0141
B-AX	39.46	0.0998	0.9588	0.0146	B-AX	41.59	0.1000	0.9682	0.0085
B-IY	38.95	0.0997	0.9476	0.0141	B-IY	39.40	0.1000	0.9444	0.0081
D-AX	39.99	0.1000	0.9585	0.0071	D-AX	38.75	0.1000	0.9458	0.0076
D-DH	41.49	0.0998	0.9670	0.0103	D-DH	40.75	0.1000	0.9574	0.0098
DH-AE	28.82	0.0950	0.8270	0.0172	DH-AE	33.47	0.0954	0.8769	0.0127
DH-AX	33.07	0.0996	0.8868	0.0123	DH-AX	35.11	0.1000	0.9142	0.0069
DH-EH	36.73	0.0998	0.9281	0.0134	DH-EH	36.29	0.0999	0.9174	0.0124
DH-EY	35.07	0.0996	0.9194	0.0121	DH-EY	37.12	0.0990	0.9306	0.0111
D-IH	40.95	0.0999	0.9665	0.0119	D-IH	40.74	0.1000	0.9614	0.0076
D-OW	40.47	0.0998	0.9644	0.0099	D-OW	40.81	0.0990	0.9621	0.0161
D-UW	40.27	0.1000	0.9573	0.0133	D-UW	41.62	0.1000	0.9718	0.0097
DX-AX	43.52	0.0998	0.9782	0.0096	DX-AX	42.51	0.1000	0.9782	0.0078
DX-IY	39.19	0.0995	0.9462	0.0112	DX-IY	41.36	0.1000	0.9742	0.0086
EH-L	37.47	0.0998	0.9409	0.0139	EH-L	37.40	0.0999	0.9453	0.0109
EH-N	33.34	0.0999	0.8867	0.0150	EH-N	34.65	0.1000	0.8986	0.0120
Continued on next page									

Table A.2 – continued from previous page

Male					Female				
Diphone	EER (%)	minDCF	C_{llr}	C_{llr}^{loss}	Diphone	EER (%)	minDCF	C_{llr}	C_{llr}^{loss}
EH-R	32.27	0.0998	0.8820	0.0111	EH-R	36.53	0.0997	0.9183	0.0125
HH-AE	33.39	0.0993	0.8873	0.0162	HH-AE	34.19	0.1000	0.8944	0.0124
HH-W	44.68	0.1000	0.9875	0.0061	HH-W	44.70	0.1000	0.9878	0.0089
IH-N	37.72	0.0996	0.9299	0.0105	IH-N	37.39	0.1000	0.9261	0.0100
IH-NG	33.47	0.0998	0.8902	0.0156	IH-NG	33.65	0.0995	0.8833	0.0107
IH-T	35.55	0.0996	0.9216	0.0162	IH-T	37.12	0.0998	0.9347	0.0095
IY-AX	38.49	0.0992	0.9368	0.0117	IY-AX	43.52	0.1000	0.9810	0.0075
IY-N	37.47	0.0969	0.9225	0.0126	IY-N	41.31	0.1000	0.9727	0.0060
IY-P	41.98	0.1000	0.9794	0.0137	IY-P	43.52	0.0999	0.9768	0.0093
JH-AX	39.85	0.0994	0.9556	0.0169	JH-AX	39.86	0.1000	0.9574	0.0144
K-AH	39.99	0.0997	0.9510	0.0108	K-AH	40.40	0.1000	0.9566	0.0115
K-AX	39.81	0.0997	0.9573	0.0083	K-AX	41.65	0.1000	0.9738	0.0056
K-S	44.51	0.1000	0.9915	0.0069	K-S	47.25	0.1000	0.9948	0.0054
L-AX	41.98	0.1000	0.9690	0.0122	L-AX	40.96	0.1000	0.9630	0.0061
L-AY	28.42	0.0957	0.8216	0.0277	L-AY	29.71	0.0944	0.8154	0.0199
L-IY	30.15	0.0966	0.8331	0.0130	L-IY	32.20	0.1000	0.8728	0.0114
M-AX	43.58	0.0998	0.9829	0.0103	M-AX	44.98	0.1000	0.9850	0.0076
M-AY	42.25	0.1000	0.9683	0.0110	M-AY	41.05	0.0988	0.9574	0.0110
M-IY	38.12	0.0984	0.9547	0.0248	M-IY	38.12	0.0997	0.9365	0.0122
N-AA	35.86	0.0972	0.9145	0.0157	N-AA	39.22	0.1000	0.9497	0.0088
N-AX	38.21	0.1000	0.9409	0.0136	N-AX	39.86	0.1000	0.9556	0.0066
N-D	31.80	0.0974	0.8469	0.0152	N-D	33.00	0.0978	0.8736	0.0082
N-DH	39.91	0.0998	0.9534	0.0113	N-DH	40.56	0.1000	0.9615	0.0079
NG-K	40.12	0.1000	0.9598	0.0103	NG-K	42.18	0.1000	0.9658	0.0086
N-IY	41.86	0.0993	0.9752	0.0189	N-IY	37.65	0.1000	0.9395	0.0087
N-OW	32.15	0.0959	0.8665	0.0128	N-OW	30.71	0.0958	0.8490	0.0159
N-S	42.82	0.0999	0.9798	0.0111	N-S	43.40	0.1000	0.9809	0.0057
N-T	35.35	0.0990	0.9099	0.0147	N-T	36.57	0.0997	0.9178	0.0096
OW-N	40.39	0.0993	0.9540	0.0116	OW-N	41.05	0.1000	0.9608	0.0092
P-AX	42.25	0.0997	0.9726	0.0107	P-AX	42.15	0.0998	0.9722	0.0093
R-AX	42.69	0.0999	0.9804	0.0163	R-AX	42.60	0.1000	0.9760	0.0053
R-AY	34.88	0.0996	0.9107	0.0239	R-AY	33.65	0.0999	0.8984	0.0174
R-IY	36.25	0.1000	0.9160	0.0138	R-IY	35.20	0.0999	0.9113	0.0124
S-AH	37.99	0.0994	0.9306	0.0124	S-AH	39.95	0.0999	0.9468	0.0099
S-AX	40.92	0.0997	0.9546	0.0114	S-AX	39.00	0.1000	0.9452	0.0083
S-OW	36.13	0.0992	0.9107	0.0102	S-OW	33.22	0.0996	0.8811	0.0116
S-T	38.92	0.0981	0.9333	0.0107	S-T	38.75	0.0999	0.9527	0.0105
T-AX	37.08	0.0985	0.9273	0.0123	T-AX	35.43	0.0999	0.9095	0.0096
T-AY	38.20	0.0983	0.9393	0.0116	T-AY	37.50	0.0996	0.9288	0.0138
T-DH	41.06	0.0995	0.9583	0.0127	T-DH	40.56	0.1000	0.9597	0.0056
TH-IH	36.80	0.0986	0.9122	0.0253	TH-IH	36.11	0.0999	0.9187	0.0106
T-R	43.70	0.0999	0.9832	0.0055	T-R	41.39	0.1000	0.9685	0.0054
T-S	41.06	0.0995	0.9618	0.0081	T-S	40.54	0.0998	0.9591	0.0086
T-UW	38.12	0.0997	0.9393	0.0127	T-UW	37.12	0.0995	0.9279	0.0123
T-W	44.38	0.1000	0.9875	0.0099	T-W	43.33	0.1000	0.9869	0.0073
UH-D	42.45	0.1000	0.9712	0.0090	UH-D	41.02	0.1000	0.9678	0.0067

Continued on next page

Table A.2 – continued from previous page

Male					Female				
Diphone	EER (%)	minDCF	C_{llr}	C_{llr}^{loss}	Diphone	EER (%)	minDCF	C_{llr}	C_{llr}^{loss}
UW-N	34.15	0.0935	0.8771	0.0161	UW-N	33.46	0.0967	0.8781	0.0140
V-AX	40.79	0.1000	0.9763	0.0100	V-AX	42.33	0.1000	0.9718	0.0074
W-AH	37.70	0.1000	0.9415	0.0095	W-AH	37.66	0.1000	0.9341	0.0092
W-AX	41.31	0.1000	0.9632	0.0115	W-AX	44.46	0.1000	0.9831	0.0090
W-EH	38.78	0.0992	0.9397	0.0101	W-EH	40.72	0.0999	0.9614	0.0083
Y-AE	25.26	0.0867	0.7533	0.0245	Y-AE	28.32	0.0894	0.7960	0.0130
Y-UW	27.69	0.0928	0.8005	0.0135	Y-UW	32.65	0.0948	0.8667	0.0114
Z-AX	37.99	0.0986	0.9306	0.0111	Z-AX	39.07	0.1000	0.9478	0.0070

Appendix B. Detailed frequency of occurrence

Table B.1: *Frequency of occurrence in NIST SRE 2004, 2005 and 2006 conversations for every phone-constraint.*

Average number of occurrences per conversation								
Phone	Male	Female	Phone	Male	Female	Phone	Male	Female
AA	17.8	16.2	EY	23.5	22.3	PUH	17.9	13.8
AE	45.1	41.4	F	19.8	17.6	PUM	6.3	8.6
AH	33.2	31.0	G	13.3	12.5	R	47.4	43.9
AO	17.7	15.9	HH	20.3	22.7	S	61.1	56.7
AW	8.1	7.6	IH	42.4	39.6	SH	7.1	6.6
AX	118.6	113.4	IY	48.1	46.4	T	101.9	97.7
AY	50.6	48.4	JH	6.9	6.7	TH	10.0	10.1
B	24.0	22.8	K	47.3	44.4	UH	7.9	7.6
CH	4.9	4.7	L	49.1	47.1	UW	34.4	31.1
D	47.8	49.0	M	38.4	37.1	V	22.4	21.8
DH	42.9	40.0	N	90.0	85.5	W	27.6	27.1
DX	8.1	7.3	NG	16.4	16.5	Y	37.8	32.8
EH	33.6	32.5	OW	36.9	36.7	Z	30.3	29.1
ER	21.5	21.3	P	21.0	19.2			

Table B.2: *Frequency of occurrence in NIST SRE 2004, 2005 and 2006 conversations for every diphone-constraint.*

Average number of occurrences per conversation								
Diphone	Male	Female	Diphone	Male	Female	Diphone	Male	Female
AA-R	24.7	23.5	DH-AE	5.9	6.1	N-DH	4.0	4.1
AA-T	21.2	18.3	DH-AX	6.4	5.9	NG-K	4.2	4.4
AE-N	17.9	17.9	DH-EH	6.0	5.9	N-IY	3.7	3.5
AE-T	14.3	14.1	DH-EY	6.7	6.3	N-OW	4.2	4.0
AE-V	13.7	12.0	D-IH	6.0	5.0	N-S	4.4	3.5
AH-M	13.0	13.4	D-OW	5.8	6.2	N-T	3.6	3.9
AH-N	12.8	12.7	D-UW	5.5	5.8	OW-N	3.7	3.5
AH-T	11.9	9.9	DX-AX	5.8	5.7	P-AX	4.0	3.6
AO-L	11.2	11.3	DX-IY	6.1	5.4	R-AX	3.5	3.4
Continued on next page								

Table B.2 – continued from previous page

Diphone	Male	Female	Diphone	Male	Female	Diphone	Male	Female
AO-R	9.8	9.0	EH-L	5.1	4.7	R-AY	3.6	4.0
AW-T	10.9	10.9	EH-N	5.5	5.1	R-IY	3.8	3.3
AX-B	11.9	11.0	EH-R	5.1	4.7	S-AH	3.7	3.7
AX-D	9.4	8.6	HH-AE	4.7	4.6	S-AX	3.5	3.2
AX-G	11.4	10.7	HH-W	4.9	4.8	S-OW	3.3	2.9
AX-K	10.2	9.0	IH-N	4.7	5.2	S-T	3.6	3.7
AX-L	9.6	9.4	IH-NG	5.4	4.8	T-AX	3.4	3.2
AX-M	8.9	8.7	IH-T	5.2	4.8	T-AY	3.7	3.5
AX-N	9.3	8.2	IY-AX	5.0	5.0	T-DH	3.1	3.6
AX-NG	8.8	8.2	IY-N	4.6	4.4	TH-IH	3.3	3.2
AX-S	8.2	7.7	IY-P	4.3	4.6	T-R	3.5	3.6
AX-T	8.0	7.9	JH-AX	4.3	4.2	T-S	3.4	3.5
AX-V	8.4	7.6	K-AH	4.3	4.3	T-UW	3.0	3.1
AX-Z	7.6	6.8	K-AX	4.4	4.2	T-W	3.2	2.9
AY-D	7.5	7.2	K-S	4.8	4.4	UH-D	3.0	3.0
AY-K	7.0	6.7	L-AX	4.4	3.9	UW-N	2.9	3.1
AY-M	7.5	7.1	L-AY	4.6	4.7	V-AX	3.2	3.0
AY-N	6.8	6.6	L-IY	4.2	4.4	W-AH	3.1	3.2
AY-T	7.4	6.4	M-AX	4.1	3.7	W-AX	3.2	2.9
B-AH	8.0	7.0	M-AY	4.5	3.7	W-EH	3.0	3.0
B-AX	6.1	6.1	M-IY	4.3	4.4	Y-AE	3.1	2.5
B-IY	6.3	6.5	N-AA	3.9	3.9	Y-UW	2.9	3.2
D-AX	6.6	6.2	N-AX	4.0	3.5	Z-AX	3.1	3.3
D-DH	6.4	6.5	N-D	4.2	3.8			