Speaker Verification Using Speaker- And Test-Dependent Fast Score Normalization

Daniel Ramos-Castro [a,*], Julian Fierrez-Aguilar [a],

Joaquin Gonzalez-Rodriguez [a], Javier Ortega-Garcia [a]

[a] *ATVS (Speech and Signal Processing Group), Escuela Politecnica Superior, Avda. Francisco Tomas y Valiente 11, Campus de Cantoblanco, Universidad Autonoma de Madrid E-28049 Madrid, Spain*

Abstract

A novel score normalization scheme for speaker verification is presented. The proposed technique is based on the widely used test-normalization method (Tnorm), which compensates test-dependent variability using a fixed cohort of impostors. The new procedure selects a speaker-dependent subset of impostor models from the fixed cohort using a distance-based criterion. Selection of the sub-cohort is made using a distance measure based on a fast approximation of the Kullback-Leibler (KL) divergence for Gaussian Mixture Models (GMM). The proposed technique has been called KL-Tnorm, and outperforms Tnorm in computational efficiency. Experimental results using NIST 2005 Speaker Recognition Evaluation protocol also show a stable performance improvement of our method on standard speaker recognition systems.

Key words: speaker verification, score normalization, Tnorm, Kullback-Leibler

1 Introduction

Automatic speaker recognition, also known as *voice biometrics*, is defined as the use of a machine to recognize persons from their voice (Campbell, 1997). It is a difficult task with many variable factors among different trials, e. g. speaker identity, transmission channel, utterance length, gender, session, speaking style, etc. We may classify these variabilities into: i) speaker-dependent, when the variability comes from the speaker data; and ii) test-dependent, when the variability comes from the test segment. It has been shown in the literature that these variations have a direct negative impact in the system performance (Doddington et al., 1998; Reynolds et al., 2000). Thus, compensation techniques are needed to cope with speech variability.

Successful compensation techniques have been proposed at different levels, e. g., at the feature (Pelecanos and Sridharan, 2001; Reynolds, 2003), model (Teunen et al., 2000) or score levels (Auckenthaler et al., 2000; Reynolds et al., 2000). Variability compensation at the score level is also referred to as score normalization (Bimbot et al., 2005). These techniques are defined as a transformation to the output scores of a speaker verification system in order to reduce misalignments in the score ranges due to variations in the conditions

* Corresponding author. Tel.: +34-91-4973212; fax: +34-91-4972235
Email addresses: daniel.ramos@uam.es (Daniel Ramos-Castro), julian.fierrez@uam.es (Julian Fierrez-Aguilar), joaquin.gonzalez@uam.es (Joaquin Gonzalez-Rodriguez), javier.ortega@uam.es (Javier Ortega-Garcia).

of a trial. Figure 1 shows a speaker verification scheme including score normalization.

Many score normalization techniques have been presented in the literature, either for speaker- and test-dependent variability compensation (Bimbot et al., 2005). One of the most popular is the so-called *test-normalization* method, also called *Tnorm* (Auckenthaler et al., 2000). This test-dependent technique estimates an impostor score distribution for each test utterance by performing a set of non-target trials with a population of impostor speaker models, or *cohort*. The similarity between the test utterance and the speaker model is then normalized using this distribution. Tnorm has become widely used in the speaker recognition community in the last years due to its significant improvement in system performance at low false acceptance rates (Navratil and Ramaswamy, 2003). It has been recently shown (Reynolds et al., 2000; Sturim and Reynolds, 2005) that an improvement in the system can be achieved when considering both speaker- and test-dependent variabilities. In this paper we propose one technique that compensates both variabilities by performing a fast selection of cohorts of speaker models for test-normalization. The contribution of this work relies on the use of a fast approximation of the Kullback-Leibler (KL) divergence as a distance between GMM models for Tnorm cohort selection. This model selection technique not only improves the system accuracy achieved using Tnorm, but also enhances the computational efficiency of the system. We have called this novel technique KL-Tnorm.

The paper is organized as follows. Introduction is completed with some definitions. Related work and motivation are described in Section 2. Section 3 describes the traditional test-normalization technique. In Section 4 the fast approximation to Kullback-Leibler divergence as a distance between GMM

models is presented, and its main properties are highlighted. KL-Tnorm is then described in Section 5. Experiments using NIST 2005 Speaker Recognition Evaluation (SRE) protocol are reported in Section 6. Finally, conclusions are drawn in Section 7.

1.1 Definitions

In most of speaker recognition systems an input speech utterance is compared to an enrolled *target* speaker model, resulting in a similarity measure between them, also called a *similarity score*. The target model is obtained from a set of training speech utterances from a known speaker. The process of computing a score from a speaker model and a test speech utterance is usually called a *trial*. The trials may be classified as *target* and *non-target* trials depending on whether the training and test speech are respectively generated by the same individual or not.

The way in which the similarity score is processed by the system defines an operation mode. In the verification mode the system has to decide whether the identity of the speaker is the same as a claimed one or not. This output decision is generated by performing a trial with a test speech utterance and a speaker model representing the claimed identity. The score is then compared to a threshold to obtain the final decision – accepted or rejected. The users attempting to access the system are referred to as *genuine* users when their identity is the same as the claimed one, otherwise they are called *impostors*. In a speaker verification system there are two types of verification error: false rejections (or "missed detections", when a genuine user is rejected) and false acceptances (or "false alarms", when an impostor is accepted). The assessment

of speaker verification systems is typically performed by means of decision theory tools such as ROC or DET curves (Martin et al., 1997), which plot both types of error in a two-dimensional graph.

2 Related work and motivation

The use of score normalization for simultaneously compensating test- and speaker-dependent variability is not new. A simple way to accomplish this objective is using different normalization techniques simultaneously (see, e. g., Vogt et al. (2005)). The main drawbacks of this approach are the high computational burden and the need of additional background speech sets. Moreover, the background data should be carefully selected to consider speaker or test conditions, and therefore these approaches can significantly increase system complexity.

Another approach to simultaneously achieve speaker- and test-dependent normalization is based on the incorporation of speaker-dependent knowledge to test-dependent techniques. One early example of such approach can be found in the context of likelihood normalization based on cohorts of speakers (Rosenberg et al., 1992). Cohort- and test-normalization techniques are closely related, as both of them are based on the estimation of the distribution of the likelihoods or scores that are obtained from trials using the test segment and a set of impostor models. As shown in Reynolds (1997) and Finan et al. (1997), cohort normalization performance was improved when the cohorts used for normalization were different for every speaker.

Recently, a speaker-dependent cohort approach has been proposed in the con-

text of test-normalization, namely adaptive Tnorm or ATnorm (Sturim and Reynolds, 2005). In this case, the $K-$nearest cohort models to the speaker model were used to normalize each score at each trial. A pool of utterances is compared both to the speaker and all cohort models, generating scores. A distance measure using these scores is used to select the $K-$nearest cohort models of each speaker.

The proposed KL-Tnorm technique follows ATnorm, but we use a distance measure based on an approximation of the Kullback-Leibler divergence (Cover and Thomas, 1991) for GMM (Reynolds et al., 2000). Our approach presents the following attractive properties:

- It is extremely fast for mean-adapted GMM from the same Universal Background Model (UBM).
- It does not require additional background data.

These properties motivate us to the use of the KL approximation described below as a distance measure to speaker-dependent cohort selection for test-normalization. This distance measure was presented in Do (2003) and has been successfully used in speaker diarization and detection (Ben et al., 2004, 2005). Other efficient distance measures approaches for GMM may be found in Aronowitz et al. (2004); Aronowitz and Burshtein (2005); Goldberger and Aronowitz (2005) in the context of efficient speaker identification and retrieval, and may be used as alternatives for the proposed methodology.

3 Test-normalization (Tnorm)

Test normalization (Tnorm) was presented in (Auckenthaler et al., 2000) and works as follows.

Let's assume that we have a sequence of feature vectors $O = \{\mathbf{o}_1, \mathbf{o}_2, ..., \mathbf{o}_N\}$ extracted from a test utterance, and a speaker model λ_t, and that we compute a score $s(O, \lambda_t)$ by comparing the observation O with the model λ_t. Tnorm uses a cohort of impostor models $\Lambda_I = \{\lambda_{I,1}, \cdots, \lambda_{I,N}\}$ to obtain the impostor scores $S_I = \{s(O, \lambda_{I,1}), \cdots, s(O, \lambda_{I,N})\}$. The normalized scores are computed as follows:

$$s_{Tnorm}(O, \lambda_t) = \frac{s(O, \lambda_t) - \mu_{Tnorm}}{\sigma_{Tnorm}} \tag{1}$$

where μ_{Tnorm} and σ_{Tnorm} are respectively the mean and standard deviation of the impostor scores S_I, assuming a Gaussian distribution. Figure 2 illustrates this technique.

4 Fast approximation to Kullback-Leibler divergence for GMM

The Kullback-Leibler (KL) divergence between two probability density functions (Cover and Thomas, 1991), also known as relative entropy, is defined by the following expression:

$$D\left(f \,\middle|\, \hat{f}\right) \equiv \int f \log \frac{f}{\hat{f}} \tag{2}$$

where f and $\hat{f}$ are arbitrary probability density functions (pdfs). KL divergence can be interpreted as an asymmetric measure of the difference between two probability distributions. The solution of Equation 2 is classically obtained by computer-intensive algorithms such as Monte-Carlo estimation. However, these techniques usually demand high computational costs, especially when high-dimension distributions are handled, representing a problem in real applications.

It has been recently shown in the literature (Do, 2003) that Equation 2 can be upper-bounded by a single expression when two Hidden Markov Models (HMM) are involved, and therefore we can particularize this expression to the GMM case. Formally, let $f = \sum_{i=1}^{M} w_i f_i$ and $\hat{f} = \sum_{i=1}^{M} \hat{w}_i \hat{f}_i$ be two pdfs associated with their corresponding $d-$dimensional GMM models λ and $\hat{\lambda}$, where w_i and $\hat{w}_i$ are real non-negative numbers (weights) and:

$$\begin{aligned} &\sum_{i=1}^{M} w_i = 1 \quad ; \quad \sum_{i=1}^{M} \hat{w}_i = 1 \\ &f_i = N\left(\mu_i, \Sigma_i\right) \quad ; \quad \hat{f}_i = N\left(\hat{\mu}_i, \hat{\Sigma}_i\right) \end{aligned} \tag{3}$$

We assume a correspondence between Gaussian components of pdfs f and $\hat{f}$, because in our case both distributions come from the same UBM via mean-only MAP adaptation [1]. Without loss of generality we assume that f_i and $\hat{f}_j$ are corresponding when $i = j$. Given these assumptions, we can develop the definition of the KL divergence between f and $\hat{f}$ in the following way:

[1] In Reynolds et al. (2000) it is shown that this adaptation scheme outperforms other MAP approaches where weights or covariances are also involved.

$$D\left(f\left|\hat{f}\right.\right) = D\left(\sum_{i=1}^{M} w_i f_i \left| \sum_{i=1}^{M} \hat{w}_i \hat{f}_i \right.\right) = \int \left(\sum_{i=1}^{M} w_i f_i\right) \log \frac{\sum_{i=1}^{M} w_i f_i}{\sum_{i=1}^{M} \hat{w}_i \hat{f}_i} \tag{4}$$

We now consider the *log-sum inequality* (Cover and Thomas, 1991):

$$\left(\sum_{i=1}^{n} a_i\right) \log \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i} \leq \sum_{i=1}^{n} a_i \log \frac{a_i}{b_i} \tag{5}$$

where $a_1, a_2, \ldots, a_n$ and $b_1, b_2, \ldots, b_n$ are two sets of non-negative numbers. By using Equation 5 the right hand side term on Equation 4 can be further developed as follows:

$$\begin{aligned} \int \left(\sum_{i=1}^{M} w_i f_i\right) \log \frac{\sum_{i=1}^{M} w_i f_i}{\sum_{i=1}^{M} \hat{w}_i \hat{f}_i} &\leq \int \sum_{i=1}^{M} \left[w_i f_i \log \frac{w_i f_i}{\hat{w}_i \hat{f}_i} \right] \\ &= \sum_{i=1}^{M} w_i \log \frac{w_i}{\hat{w}_i} + \sum_{i=1}^{M} w_i \int f_i \log \frac{f_i}{\hat{f}_i} \end{aligned} \tag{6}$$

Therefore, the KL divergence between the two GMM models is upper-bounded by two terms:

- The first term is the KL divergence between the weights of both pdfs.
- The second term is the weighted sum of the individual KL divergences of the corresponding Gaussian mixtures. These individual divergences can be computed using the following formula (Do, 2003):

$$\begin{aligned} \int f_i \log \frac{f_i}{\hat{f}_i} = \frac{1}{2} \Bigg[&\log \frac{\det\left(\Sigma_i\right)}{\det\left(\hat{\Sigma}_i\right)} - \dim\left(\Sigma_i\right) + tr\left(\hat{\Sigma}_i^{-1} \Sigma_i\right) \\ &+ \left(\mu_i - \hat{\mu}_i\right)^t \hat{\Sigma}_i^{-1} \left(\mu_i - \hat{\mu}_i\right) \Bigg] \end{aligned} \tag{7}$$

In our case, both GMM are adapted from the same UBM using mean-only MAP adaptation (Reynolds et al., 2000). Therefore, the weight vectors and covariance matrices are the same in both models, and the first term in the right hand of Equation 6 is canceled. The KL divergence approximation in this situation is reduced to the following expression:

$$D\left(f\left|\hat{f}\right.\right) \leq D_a\left(f\left|\hat{f}\right.\right) = \sum_{i=1}^{M} w_i \int f_i \log \frac{f_i}{\hat{f}_i}$$
$$= \sum_{i=1}^{M} \frac{w_i}{2}\left[\left(\mu_i - \hat{\mu}_i\right)^t \Sigma_i^{-1}\left(\mu_i - \hat{\mu}_i\right)\right] \quad (8)$$

The approximation $D_a\left(f\left|\hat{f}\right.\right)$ in Equation 8 has several attractive properties:

Low resource cost. The computational cost required to compute $D_a\left(f\left|\hat{f}\right.\right)$ is much lower compared with other techniques for KL divergence estimation such as Monte-Carlo methods.

Symmetry. $D_a\left(f\left|\hat{f}\right.\right)$ between two mean-adapted GMM from the same UBM is symmetric, i. e.,

$$D_a\left(f\left|\hat{f}\right.\right) = D_a\left(\hat{f}\left|f\right.\right) \quad (9)$$

Tightness. In Do (2003), experiments show that $D_a\left(f\left|\hat{f}\right.\right)$ is reasonably tight to the Monte-Carlo estimated KL divergence.

Correlation. Ben et al. (2004) show that $D_a\left(f\left|\hat{f}\right.\right)$ is highly correlated with the symmetric expression of the KL divergence computed via Monte-Carlo estimation, namely $D_2\left(f\left|\hat{f}\right.\right) = D\left(f\left|\hat{f}\right.\right) + D\left(\hat{f}\left|f\right.\right)$.

Interpretation as a weighted sum of Mahalanobis distances. The Mahalanobis distance between two multidimensional Gaussian pdfs having the same covariance matrix, namely $g = N\left(\mu_g, \Sigma_g\right)$ and $\hat{g} = N\left(\hat{\mu}_g, \Sigma_g\right)$ is defined as follows:

$$D_m\left(g\left|\hat{g}\right.\right) = \left[\left(\mu_g - \hat{\mu}_g\right)^t \Sigma_g^{-1} \left(\mu_g - \hat{\mu}_g\right)\right] \tag{10}$$

So, it can be noted that $D_a\left(f\left|\hat{f}\right.\right)$ is a weighted sum of Mahalanobis distances between each of the corresponding Gaussian components in each GMM.

The aforementioned properties of $D_a\left(f\left|\hat{f}\right.\right)$ in the model domain make it useful in areas where it may be necessary to compute distances between models, e. g. speaker diarization (Ben et al., 2004), speaker detection (Ben et al., 2005) and the proposed score normalization.

5 KL-Tnorm: speaker-dependent test-normalization

The approximation of KL divergence in Equation 8 is proposed to select speaker-dependent cohorts for Tnorm in speaker verification systems. We have called this novel technique KL-Tnorm. For each score $s\left(O, \lambda_t\right)$, the application of KL-Tnorm can be described as follows:

Computation of distances. For each target speaker model λ_t, we compute a set of KL divergence approximations to each model of a given cohort $\Lambda_I = \{\lambda_{I,1}, \cdots, \lambda_{I,N}\}$, namely $D_{t,I} = \{D_a\left(f_t\left|f_{I,1}\right.\right), \cdots, D_a\left(f_t\left|f_{I,N}\right.\right)\}$ using Equation 8.

Selection of K-nearest models. We select the K-nearest impostor models to λ_t (with $K < N$) following the KL divergence approximation criterion, and so we will obtain a set of impostor models $\Lambda_{KL-I} = \{\lambda_{KL-I,1}, \cdots, \lambda_{KL-I,K}\}$, being $\Lambda_{KL-I} \subset \Lambda_I$.

Computation of KL-Tnorm scores. We compute the impostor scores $S_{KL-I} =$

$\{s(O, \lambda_{KL-I,1}), \cdots, s(O, \lambda_{KL-I,K})\}$.

Normalization. KL-Tnorm is finally performed as:

$$s_{KL-Tnorm}(O, \lambda_t) = \frac{s(O, \lambda_t) - \mu_{KL-Tnorm}}{\sigma_{KL-Tnorm}} \tag{11}$$

where $\mu_{KL-Tnorm}$ and $\sigma_{KL-Tnorm}$ are respectively the mean and standard deviation of S_{KL-I} assuming a Gaussian distribution.

KL-Tnorm technique is illustrated in Figure 3. It is important to remark that KL-Tnorm can be applied to any non-GMM speaker recognition system as well. One straightforward way to apply KL-Tnorm with other verification approaches is to compute $D_{t,I}$ using GMM modeling and then use this distance set to select the KL-Tnorm cohorts. The system at hand can then be used to obtain $s(O, \lambda_t)$ and S_{KL-I} distribution parameters in Equation 11.

6 Experiments

In this section we present results that show that the application of KL-Tnorm to the output scores of two common speaker verification systems outperforms the classical Tnorm technique. This section is organized as follows. The baseline speaker verification systems used are briefly sketched in Section 6.1. The experimental protocol and databases are described in Section 6.2. Results are reported in Section 6.3 including both development, evaluation and post-evaluation experiments and a discussion of computational efficiency issues.

6.1 Baseline systems

6.1.1 GMM system description

The system is based on a likelihood ratio detector with target and alternative probability distributions modelled by Gaussian Mixture Models (Reynolds et al., 2000). The similarity scores are computed by means of the following log-likelihood ratio formula:

$$s\left(O, \lambda_t\right) = \log f\left(O \,|\lambda_t\right) - \log f\left(O \,|\lambda_{UBM}\right) \tag{12}$$

where $f\left(O \,|\lambda_t\right)$ and $f\left(O \,|\lambda_{UBM}\right)$ are the probability density functions for the target model and a UBM (Reynolds et al., 2000), which are modeled as mixtures of Gaussians, and O is the sequence of feature vectors from the test utterance.

Target speaker models are derived using means-only MAP adaptation, which can be described as follows: first, a GMM is trained using speech from multiple speakers, namely the Universal Background Model (UBM), which represents the common feature distribution shared among speakers. Target speaker models are then trained by adapting a GMM from the UBM using the target speaker training data. This is done by means-only MAP adaptation, i. e., the UBM means are moved to fit the target speaker training data (see Reynolds et al. (2000)). In order to perform this adaptation, the Expectation Maximization (EM) algorithm has been used (Duda et al., 2001; Reynolds et al., 2000).

Feature extraction in order to obtain the O sequence for each utterance is

performed as follows: 19 Mel Frequency Cepstral Coefficients (MFCC) are extracted from the speech signal using overlapped Hamming windows. These windows have a length of 20 ms. and they are overlapped 10 ms. No band-limiting has been performed in the extracted features. As channel mismatch seriously affects the performance of speaker recognition systems, several channel compensation techniques have been applied to the obtained feature vectors. First, Cepstral Mean Normalization (CMN) has been used in order to remove linear channel distortion. Then, RASTA filtering (Hermansky and Morgan, 1994) has been performed. Finally, a short-time Gaussianization, namely feature warping (Pelecanos and Sridharan, 2001), has been used in order to achieve robustness against channel and noise effects.

6.1.2 SVM system description

The SVM speaker recognition system is also based on the spectral characteristics of the speech, as the GMM system described above. However, in this case the similarity computation is based on discriminative Support Vector Machines (SVM) (Campbell et al., 2005). A kernel expansion is performed on the whole observation sequence O, and a separating hyperplane is computed between the speaker features and the background model.

ATVS acoustic SVM system uses a polynomial expansion of degree two (Wan and Campbell, 2000) followed by a Generalized Linear Discriminant Sequence Kernel (GLDS) as described in Campbell (2002). We use a channel compensation matrix (Solomonoff et al., 2004) in order to avoid channel mismatch effects.

The system uses the same feature extractor as the GMM system. Two gender-

dependent and channel-independent data sets from the development set are used for background modelling. KL-Tnorm score normalization is performed using approximations of the KL divergences computed from GMM models as described in Section 5.

6.2 Databases and experimental framework

Experiments have been performed using the evaluation protocol proposed by NIST in its 2005 Speaker Recognition Evaluation (SRE) (NIST SRE, 2005), which is similar to the one used in 2004 (van Leeuwen et al., 2005). The database used in this evaluation is a subcorpus of the MIXER database (Campbell et al., 2004). The acquisition conditions include different communication channels (landline, GSM, CDMA, etc.), different handsets and microphones (carbon button, electret, earphones, cordless, etc.) and different languages (American English, Arabic, Spanish, Mandarin, etc.). The evaluation protocol defines the following training conditions: 10 seconds, 1, 3 and 8 conversation sides; and the following test conditions: 10 seconds, 1 conversation side, 3 full conversations in a mixed channel and multichannel microphone data. Each conversation side has an average duration of 5 minutes, with 2.5 minutes of speech on average after silence removal. Although there are speakers of both genders in the corpus, no cross-gender trials are defined. Details can be found in the NIST 2005 SRE Evaluation Plan (NIST SRE, 2005). We carry out our experiments using the ATVS GMM and SVM systems submitted to NIST 2005 SRE, which are described in Section 6.1.

Before the evaluation, a development set, consisting of the NIST 2004 SRE database, was selected, which is also a subset of MIXER. Trials performed

using this development set follow the NIST 2004 SRE protocol (van Leeuwen et al., 2005). Additional background data needed for development trials (UBM, normalization cohorts, etc.; namely background data) were selected from the same development set. For NIST 2005 SRE tests we used NIST 2004 SRE database as background data. Therefore the Tnorm cohorts consist of the NIST 2004 SRE target models for each training condition. The total number of models N in each cohort is shown in Table 1. Target and cohort models conditions regarding gender and amount of training data match in all cases. The experiments are performed using 1 conversation side for testing and both 1 and 8 conversation sides for training (1c-1c and 8c-1c conditions respectively).

6.3 Results

6.3.1 Development experiments

Figure 4 summarizes the experiments performed in the development set. We vary the number of models K used for KL-Tnorm and plot the EER in (%). By observing Figure 4 we can compare the use of Tnorm ($K = N$) and several operating points of KL-Tnorm ($K = 25, ..., 150$). We observe that KL-Tnorm improves the system performance in terms of Equal Error Rate (EER) especially for $K = 50$ and $K = 75$. We also observe that the optimum EER value is obtained for $K = 50$. Figure 5 shows the optimum Detection Cost Function (DCF) as defined by NIST evaluations (van Leeuwen et al., 2005) for the same experiments as in Figure 4. DCF measures the quality of the decisions that a speaker verification system takes according to fixed costs assigned to wrong decisions, and is used as a scalar value to rank the systems in NIST evaluations. We observe a general trend of DCF improvement exists

for KL-Tnorm in all cases.

6.3.2 Evaluation experiments

In order to perform KL-Tnorm in NIST 2005 SRE, we set $K = 75$ for both systems (GMM and SVM) and conditions (1c-1c and 8c-1c). Figure 6 shows the performance of the GMM and SVM systems in both 1c-1c and 8c-1c conditions when no normalization, Tnorm and KL-Tnorm are used. For all these cases, Tnorm has been applied using the total number of models in the cohorts (Table 1). We note an improvement in system performance for KL-Tnorm with respect to Tnorm in the 8c-1c condition, whereas this is not appreciated for the 1c-1c condition.

The efficiency gain of KL-Tnorm can be illustrated by estimating the processing time of Tnorm and KL-Tnorm:

$$\begin{aligned} t_{Tnorm} &= N \cdot t_{score} \\ t_{KL-Tnorm} &= N \cdot t_{KLa} + K \cdot t_{score} \end{aligned} \tag{13}$$

where t_{score} is the time needed to compute a score using the fast scoring technique presented in Reynolds et al. (2000) and t_{KLa} is the time required to perform D_a using Equation 8 [2]. Defining $R = t_{score}/t_{KLa}$ we obtain:

$$t_{KL-Tnorm} = t_{Tnorm}\left(\frac{1}{R} + \frac{K}{N}\right) \tag{14}$$

[2] We have considered the rest of processing times (sorting the distances, memory access, etc.) negligible compared to t_{KLa} and t_{score}

Note that if $(1/R + K/N) < 1$ then we obtain a computational gain. We have estimated from our experiments that these values are, on average, $R \simeq 10$ and $K/N \simeq 0.3$. Therefore the computational cost of the score normalization technique is decreased to $t_{KL-Tnorm} \simeq 0.4 \cdot t_{Tnorm}$.

6.3.3 Post-evaluation experiments

In order to compare KL-Tnorm to Tnorm at similar operating points in terms of computational burden, we compare here both of them with the same number of models in the cohorts. As before, we use $K = 75$ models in KL-Tnorm for all experiments. In an analogous way, we apply Tnorm using $K = 75$ models randomly selected from each whole cohort. In order to perform statistically significant experiments, we evaluate the system using 10 different random selections for Tnorm cohorts, and averaging the results. Table 2 for the GMM system show significant EER improvement in all conditions when KL-Tnorm is used. On the other hand, a slight improvement in DCF values is appreciated in all cases. Table 3 shows the same results for the SVM system. In this case the performance gain is even higher.

7 Conclusions

In this paper we have presented a novel technique for score normalization, namely KL-Tnorm, which performs speaker-dependent Tnorm by selecting the nearest models to each speaker model from a given cohort. The contribution of this work relies on the selection technique proposed, which is based on a fast approximation of the Kullback-Leibler divergence for Gaussian Mixture Models. Experiments following NIST 2005 Speaker Recognition Evaluation

protocol have shown that KL-Tnorm outperforms Tnorm in verification performance. Moreover, we have observed stable improvements over a wide range of number of models in the cohort and for various experimental scenarios. It has demonstrated that KL-Tnorm also outperforms Tnorm in computational efficiency, with an average reduction of processing time by a factor 0.4 in the standard benchmark defined by NIST.

KL-Tnorm is useful in speaker recognition applications to select adapted subcohorts from large fixed set of models presenting different conditions. As has been demonstrated with two standard systems, the proposed technique can be applied to any speaker recognition paradigm. Furthermore, our approach can be directly applied to other biometric systems as well. This method may be useful in forensic applications due to strong variability present in the biometric samples (Gonzalez-Rodriguez et al., 2005). Also, performance improvements in other biometric systems can be expected if they are based on traits showing strong user- and test-dependencies, such as the written signature (Fierrez-Aguilar et al., 2005). Future work consider the exploration of different distance measure schemes (Goldberger and Aronowitz, 2005) and the use of the proposed cohort selection methodology in different normalization schemes such as TZnorm (Aronowitz and Burshtein, 2005) and other speaker adaptation frameworks.

Acknowledgements

This work was in part supported by the Spanish Ministry for Science and Technology under projects TIC2003-09068-C02-01 and TIC2003-08382-C05-01. The authors D. R.-C. and J. F.-A. also thank Consejeria de Educacion

de la Comunidad de Madrid and Fondo Social Europeo for supporting their doctoral research.

References

Aronowitz, H., Burshtein, D., 2005. Efficient speaker identification and retrieval. In: Proc. of Interspeech. pp. 2433–2436.

Aronowitz, H., Burshtein, D., Amir, A., 2004. Speaker indexing in audio archives using test utterance gaussian mixture modeling. In: Proc. of ICSLP. pp. 609–612.

Auckenthaler, R., Carey, M., Lloyd-Tomas, H., 2000. Score normalization for text-independent speaker verification systems. Digital Signal Processing 10, 42–54.

Ben, M., Betser, M., Bimbot, F., Gravier, G., 2004. Speaker diarization using bottom-up clustering based on a parameter-derived distance between adapted GMM. In: Proc. of ICSLP. pp. 2329–2332.

Ben, M., Gravier, G., Bimbot, F., 2005. A model space framework for efficient speaker detection. In: Proc. of Interspeech. pp. 3061–3064.

Bimbot, F., Bonastre, J. F., Fredouille, C., Gravier, G., Magrin-Chagnolleau, I., Meignier, S., Merlin, T., Ortega-Garcia, J., Petrovska-Delacretaz, D., Reynolds, D. A., 2005. A Tutorial on Text-Independent Speaker Verification. Journal on Applied Signal Processing 4, 430–451.

Campbell, J. P., 1997. Speaker verification: A tutorial. Proceedings of the IEEE 85, 1437–1462.

Campbell, J. P., Nakasone, H., Cieri, C., Miller, D., Walker, K., Martin, A. F., Przybocki, M. A., 2004. The MMSR bilingual and crosschannel corpora for speaker recognition research and evaluation. In: Proc. of Odyssey. pp. 29–32.

Campbell, W., 2002. Generalized linear discriminant sequence kernels for speaker recognition. In: Proc. of ICASSP. pp. 161–164.

Campbell, W., Campbell, J., Reynolds, D., Singer, E., Torres-Carrasquillo, P., 2005. Support Vector Machines for speaker and language recognition. Computer Speech and Language (to appear).

Cover, T. M., Thomas, J. A., 1991. Elements of Information Theory. Wiley Interscience.

Do, M. N., 2003. Fast approximation of Kullback-Leibler distance for dependence trees and Hidden Markov Models. IEEE Signal Processing Letters 10, 115–118.

Doddington, G., Liggett, W., Martin, A., Przybocki, M., Reynolds, D. A., 1998. Sheeps, goats, lambs and wolves: A statistical analysis of speaker performance in the NIST 1998 speaker recognition evaluation. In: Proc. of ICSLP.

Duda, R. O., Hart, P. E., Stork, D. G., 2001. Pattern Classification. Wiley.

Fierrez-Aguilar, J., Ortega-Garcia, J., Gonzalez-Rodriguez, J., 2005. Target dependent score normalization techniques and their application to signature verification. IEEE Trans. on Systems, Man and Cybernetics, part C 35 (3), 418–425.

Finan, R., Sapeluk, A., Damper, R., 1997. Impostor cohort selection for score normalisation in speaker verification. Pattern Recognition Letters 18, 881–888.

Goldberger, J., Aronowitz, H., 2005. A distance measure between gmms based on the unscented transform and its application to speaker recognition. In: Proc. of Interspeech. pp. 1985–1988.

Gonzalez-Rodriguez, J., Fierrez-Aguilar, J., Ramos-Castro, D., Ortega-Garcia, J., 2005. Bayesian analysis of fingerprint, face and signature evidences with

automatic biometric systems. Forensic Science International 155 (2-3) 126–140.

Hermansky, H., Morgan, N., 1994. RASTA processing of speech. IEEE Transactions on Speech and Audio Processing 2 (4), 578–589.

Martin, A., Doddington, G., Kamm, T., Ordowski, M., Przybocki, M., 1997. The DET curve in assessment of decision task performance. In: Proc. of Eurospeech. pp. 1895–1898.

Navratil, J., Ramaswamy, G., 2003. The awe and mystery of T-Norm. In: Proc. of Eurospeech. pp. 2009–2012.

NIST, 2005. Speaker recognition evaluation plan: http://www.nist.gov/speech/tests/spk/2005/index.htm.

Pelecanos, J., Sridharan, S., 2001. Feature warping for robust speaker verification. In: Proc. of Odyssey. pp. 213–218.

Reynolds, D. A., 1997. Comparison of background normalization methods for text-independent speaker verification. In: Proc. of Eurospeech. pp. 963–966

Reynolds, D. A., 2003. Channel robust speaker verification via feature mapping. In: Proc. of ICASSP. pp. 53–56.

Reynolds, D. A., Quatieri, T. F., Dunn, R. B., 2000. Speaker verification using adapted Gaussian Mixture Models. Digital Signal Processing 10, 19–41.

Rosenberg, A. E., DeLong, J., Lee, C. H., Juang, B. H., , Soong, F. K., 1992. The use of cohort normalized scores for speaker verification. In: Proc. of ICSLP. p. 599-602.

Solomonoff, A., Quillen, C., Campbell, W., 2004. Channel compensation for SVM speaker recognition. In: Proc. of Odyssey. pp. 57–62.

Sturim, D., Reynolds, D. A., 2005. Speaker adaptive cohort selection for Tnorm in text-independent speaker verification. In: Proc. of ICASSP. pp. 741–744.

Teunen, R., Shahshahani, B., Heck, L., 2000. A model-based transformational approach to robust speaker recognition. In: Proc. of ICSLP. pp. 495–498.

van Leeuwen, D., Martin, A., Przybocki, M., Bouten, J., 2005. The NIST 2004 and TNO/NFI speaker recognition evaluations. Computer Speech and Language (to appear).

Vogt, R., Baker, B., Sridharan, S., 2005. Modelling session variability in text-independent speaker verification. In: Proc. of Interspeech. pp. 3117–3120.

Wan, W., Campbell, W., 2000. Support Vector Machines for speaker verification and identification. In: Proc. of IEEE International Workshop on Neural Networks for Signal Processing. pp. 775–784.

DANIEL RAMOS-CASTRO received his M.S. in Electrical Engineering in 2001 from Universidad Politecnica de Madrid, Spain. Since 2004 he is with Universidad Autonoma de Madrid, where he is currently working towards the Ph.D. degree on speaker recognition with forensic applications. His research interests are focused on speech and signal processing, pattern recognition, biometrics with forensic applications and Bayesian inference. He has participated in the development of the ATVS-UPM speaker recognition system for the NIST 2004 and 2005 evaluations. He has been part of the organizing committee for "Odyssey-04, The ISCA Speaker Recognition Workshop".

JULIAN FIERREZ-AGUILAR received the M.S. degree in Electrical Engineering in 2001, from Universidad Politecnica de Madrid. Since 2004 he is with Universidad Autonoma de Madrid, where he is currently working towards the Ph.D. degree on multimodal biometrics. His research interests include signal and image processing, pattern recognition and biometrics. He was the recipient of the Best Poster Award at AVBPA 2003 and led the development of the UPM signature verification system ranked 2nd in SVC 2004. He has been part of the organizing committee for "Odyssey-04, The ISCA Speaker Recognition Workshop".

JOAQUIN GONZALEZ-RODRIGUEZ received the Ph.D. degree in Electrical Engineering in 1999 from Universidad Politecnica de Madrid. He is currently an Associate Professor at Universidad Autonoma de Madrid. His research interests are focused on speech and signal processing, biometrics with forensic applications and language recognition. He is an invited member of ENFSI-FSAAWG and has been vice-chairman for "Odyssey-04, The ISCA Speaker Recognition Workshop".

JAVIER ORTEGA-GARCIA received the Ph.D. degree in Electrical Engineering in 1996 from Universidad Politecnica de Madrid. He is currently an Associate Professor at Universidad Autonoma de Madrid. His research interests are focused on forensic acoustics, biometrics signal processing and security applications. He has participated in several scientific and technical committees, and has chaired "Odyssey-04, The ISCA Speaker Recognition Workshop".

Figures

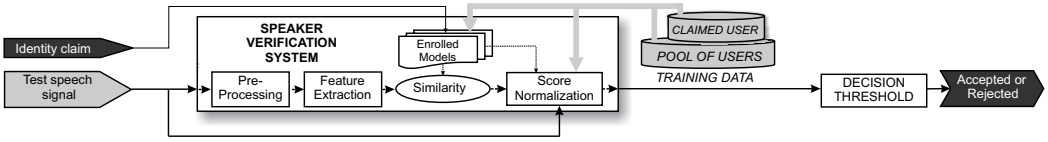


Fig. 1. General scheme for a Speaker Verification System with Score Normalization.

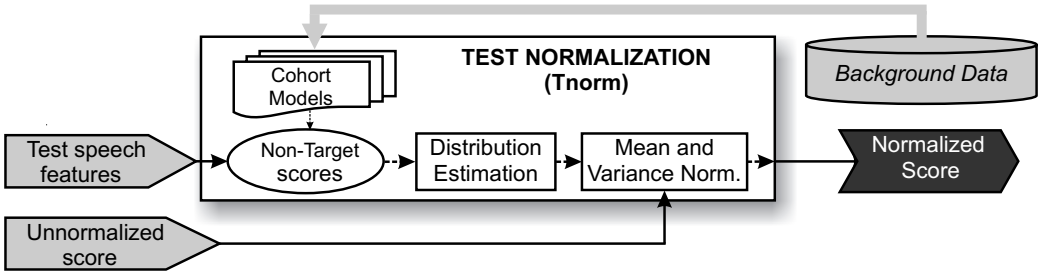


Fig. 2. Test-normalization Technique (Tnorm).

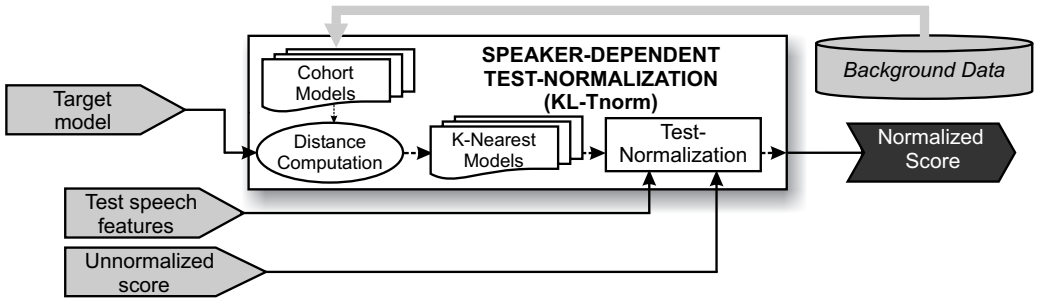


Fig. 3. KL-Tnorm.

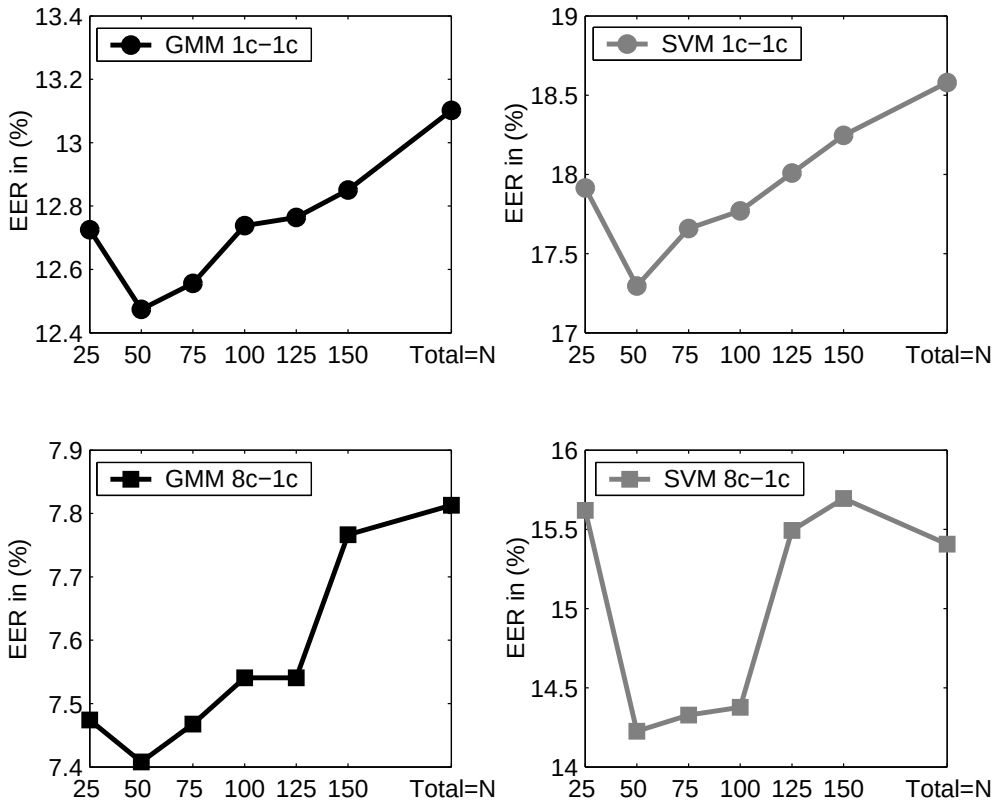


Fig. 4. EER performance in the development set. The number of models in the cohort, K, is represented in the horizontal axis for the GMM (left) and SVM (right) systems and for 1c-1c (up) and 8c-1c (down) conditions.

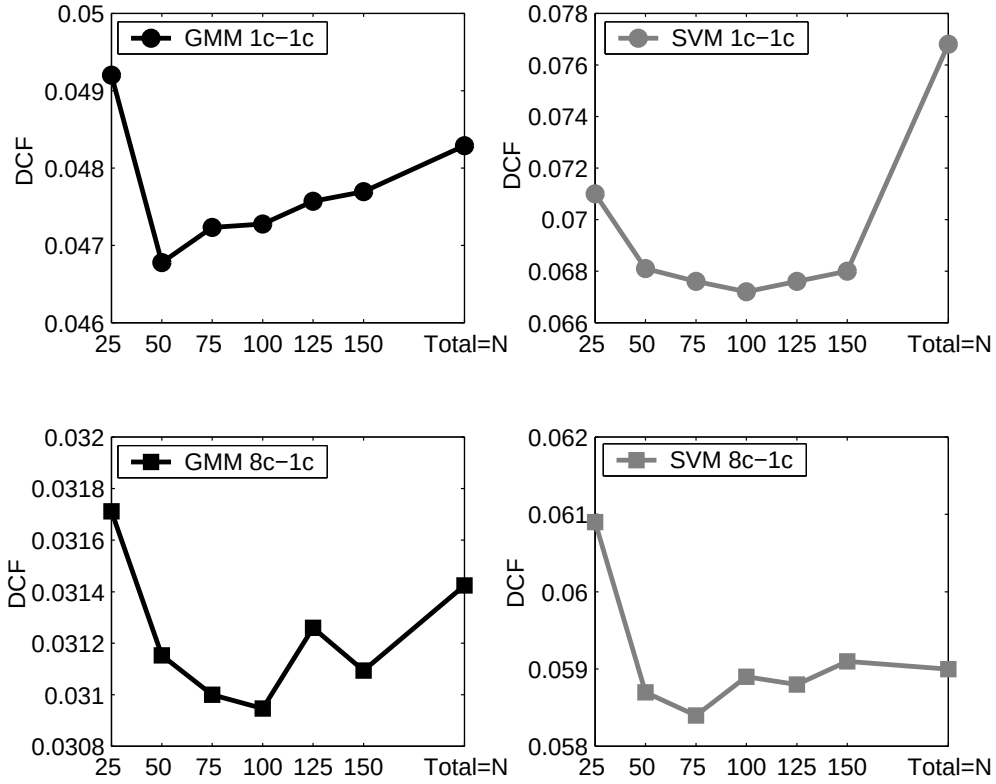
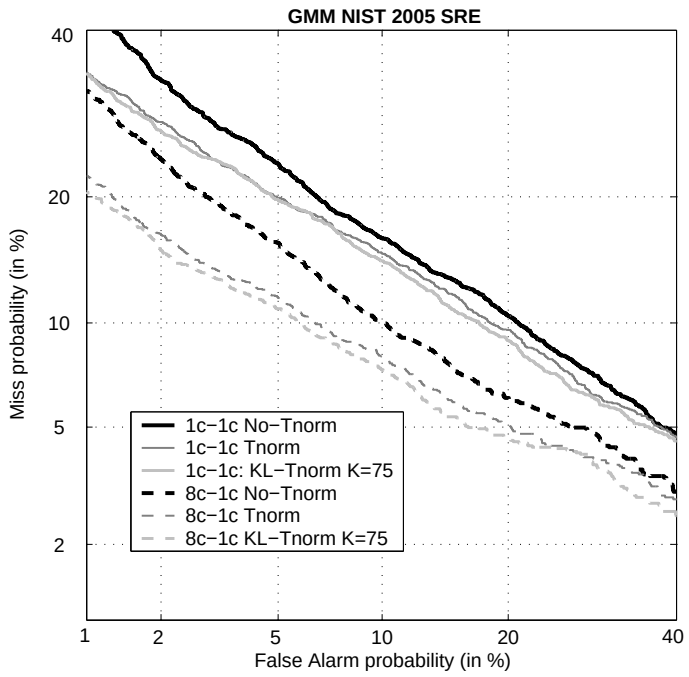
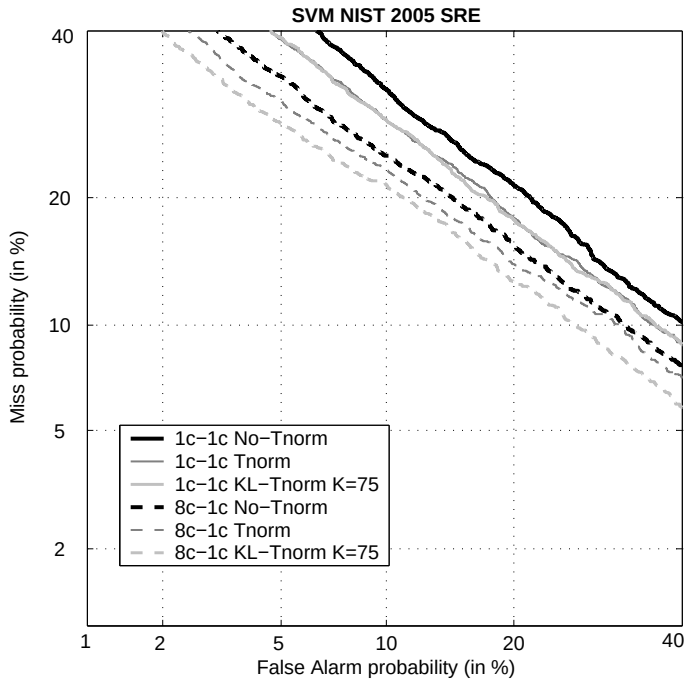


Fig. 5. DCF performance (as defined by NIST) in the development set. The number of models in the cohort, K, is represented in the horizontal axis for the GMM (left) and SVM (right) systems and for 1c-1c (up) and 8c-1c (down) conditions.



(a)



(b)

Fig. 6. KL-Tnorm in blind NIST 2005 SRE. (a) GMM system and (b) SVM system.

Tables

Table 1

Total number of models N in each Tnorm cohort

	1c-1c		8c-1c	
	male	female	male	female
N in cohort	246	370	170	205

Table 2

Tnorm and KL-Tnorm in GMM system using $K = 75$ models for both techniques

(Tnorm values are averaged from 10 random selection trials)

GMM	1c-1c		8c-1c	
$K = 75$	male	female	male	female
EER Tnorm (Av.)	11.14	14.62	7.78	9.57
EER KL-Tnorm	10.76	13.88	7.25	9.12
EER Av. Improvement	3.4%	5.0%	6.8%	4.7%
DCF Tnorm (Av.)	0.041	0.048	0.030	0.033
DCF KL-Tnorm	0.039	0.047	0.029	0.031

Table 3

Tnorm and KL-Tnorm in SVM system using $K = 75$ models for both techniques (Tnorm values are averaged from 10 random selection trials)

SVM	1c-1c		8c-1c	
$K = 75$	male	female	male	female
EER Tnorm (Av.)	19.22	19.27	16.58	16.79
EER KL-Tnorm	17.19	17.87	14.15	14.59
EER Av. Improvement	11.6%	7.3%	14.7%	13.1%
DCF Tnorm (Av.)	0.073	0.075	0.060	0.060
DCF KL-Tnorm	0.063	0.073	0.057	0.059