15th European Signal Processing Conference (EUSIPCO 2007), Poznan, Poland, September 3-7, 2007, copyright by EURASIP

PAIR-OF-SEQUENCES SVM SPEAKER VERIFICATION

Jérôme Louradour, Khalid Daoudi

IRIT - CNRS UMR 5505, Université Paul Sabatier
118, route de Narbonne, 31062 Toulouse cedex 04, FRANCE
phone: +33 (0)5 61 55 72 01 - fax: +33 (0)5 61 55 62 58
email: {louradou,daoudi}@irit.fr - web: www.irit.fr/recherches/SAMOVA/

ABSTRACT

We present a new concept of speaker verification based on a target independent decision system. The basic principle is to build a system that decides whether two sequences were pronounced by the same speaker. In our view, this system is aimed to complement traditional ones. While the principle is quite general, in this paper we use an SVM scheme to implement it. To do so, we conceive a kernel between pair of sequences using GMM distributions estimated on each given sequence. We present experiments on NIST Speaker Recognition Evaluation. The individual performance of the new system is similar to the GLDS-SVM, and the fusion of both outperforms the baseline GMM system.

1. INTRODUCTION

The training of a target speaker classifier is the heart of state-of-the-art text-independent speaker verification systems. That is, given some target speaker (TS) utterances to be used for system training, a TS-dependent classifier is built using some additional background data that represent impostor utterances. Given a test utterance, the classifier returns a score that is usually compared with a TS-independent threshold to automatically decide whether the utterance was pronounced by the target speaker. Fig.1 illustrates the underlying architecture when there is *one single* training utterance available for the target speaker, as with NIST Speaker Recognition Evaluation (SRE) in the "core test condition". This experimental protocol will be our setting in this paper.

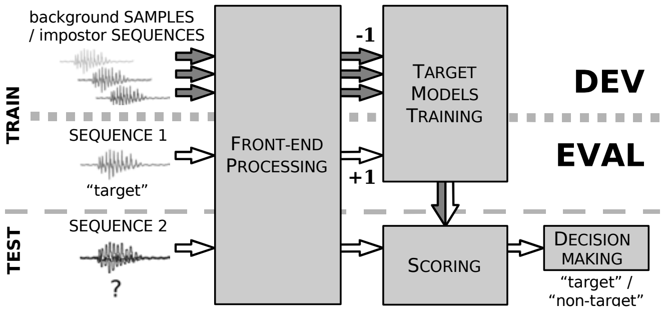

Figure 1: Block diagram of a traditional speaker verification system, with only one training utterance.

Note however that, particularly when the durations of the test and training utterances are of the same order, one can think of reversing the roles of these utterances, *i.e.* of training the target model on the test sequence and to score on the training sequence. This idea of "input swapping" was suggested during NIST SRE workshops (*e.g* see [1]).

Inspired by this observation, we investigate in this paper the idea of conceiving a speaker verification system where the inputs are pairs of sequences (playing a symmetric role) rather that single sequences. The goal of this system is to determine if "two sequences were pronounced by the same speaker"; note that this problem formulation is usually encountered in speaker segmentation. In other words, we aim at building up a classifier for which the input is the pair formed by the target and the test sequences, and the output is a decision whether these sequences were uttered by the same speaker. It has several advantages in comparison with traditional approaches:

- It enables to conceive systems with high efficiency and/or low memory requirement, because only one model has to be trained to solve the speaker verification problem.
- Expressing a training criterion on pairs of sequences is suitable to discriminate between intra- and inter-speaker variabilities. For instance, the channel mismatch problem can be directly attacked provided that the training corpus contains sequences with several recording conditions for each background speaker.

Finally, the problem reformulation leads to a novel strategy to exploit input information, that can be expected to complement classical strategies.

As will be explained, Support Vector Machines (SVM) seem to be an appropriate tool to implement the new system. The main challenge is to construct a kernel between pairs of sequences. We provide a solution using distances between Gaussian Mixture Models (GMM) estimated on the sequences of acoustic vectors. Discriminative training with the new kernel amounts to learn a distance between distributions of acoustic vectors, that is relevant for speaker recognition.

In the next section we develop the architecture of the new system. In section 3 we the discuss the implementation of such a system with an SVM scheme. In section 4, we develop the construction of the kernel between pairs of sequences required for the SVM. Finally in section 5, we show experiments on a text-independent speaker verification task using NIST SRE database.

2. PAIR-OF-SEQUENCES SPEAKER VERIFICATION SYSTEM

In the whole paper, we consider the scenario where only one sequence is available to characterize each target speaker (namely one training utterance from the traditional viewpoint). The architecture of our pair-of-sequences speaker verification system is given in Fig.2. The inputs of this system are pairs of utterances, we thus need a development set

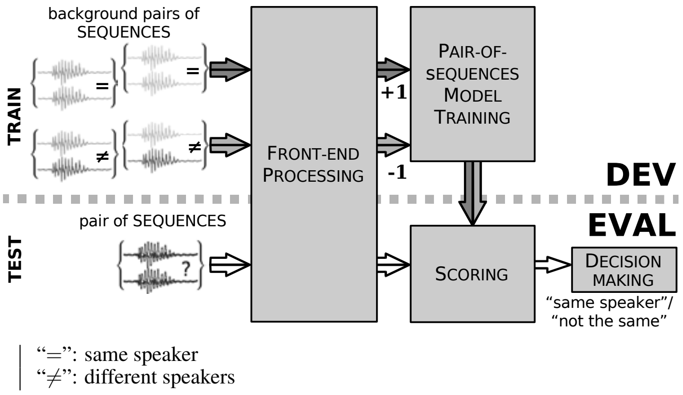

Figure 2: Block Diagram of the Pair-of-Sequences (PoS) speaker verification system.

of pairs of sequences. This requires a background set of sequences labeled according to speakers. This set is then arranged so as to provide positive (resp. negative) trials with pairs of sequences uttered by the same speaker (resp. different speakers). After the front-end processing, a *single* classification model is trained to decide whether a pair of sequences corresponds to a same speaker. Note the major difference w.r.t. traditional systems: only one classification model is trained and is independent of the target speaker. The target speaker sequence is *not* used to train the classifier. This sequence is used during testing in combination with a given test utterance. This makes this system somehow "universal" because of its independence of the target speaker. In our view, this could be an advantage if the system is used to help for decision making, for instance by fusing it with a traditional system.

We mention here that a similar an approach, although not published, was proposed by [1] under the name "Grand Logistic Regression". Like the Fisher kernel approach [2], this work is based on the derivative computation of each sequence likelihood *w.r.t* the Universal Background Model (UBM), which is estimated on a large unlabeled database and describes the prior distribution. It was reported that the resulting speaker verification system showed bad performance individually, but improved performance when fused with a classical system.

3. PAIR-OF-SEQUENCES SVM SPEAKER VERIFICATION

The new system can be implemented using any decision algorithm that enables to decide whether two sequences were pronounced by the same speaker. For instance, techniques from speaker segmentation can be borrowed for this propose, such as Bayesian Information Criterion (BIC) [3] or the Generalized Likelihood Ratio [4]. These quantities compare the hypothesis that two segments were generated or not by a same process, via the likelihoods *w.r.t* models trained on each segment separately and on both segments. Unfortunately, using these criteria for speaker verification produces poor empirical results compared to classical approaches because of the lack of discriminative training.

In this paper we use Support Vector Machines (SVM) to implement the new system for two main reasons. First, SVM allow efficient learning and can naturally handle different kinds of data including structured ones [5]. Actually, the main challenge to apply SVM in our case will be to design a suitable kernel between pairs of sequences.

Second, the resulting training procedure presents an advantage *w.r.t.* classical SVM training in speaker verification. Indeed, with the new system it is easy to collect a similar number of positive and negative trials for the training corpus. In classical SVM systems with sequence kernels, the target model is trained using the available sequences of the target speaker, which number is usually much lower than the number of impostor sequences. And in the scenario we are considering, it is one-against-all scheme. This imbalance is a limitation for the generalization capacity of SVM, as reported for example in [6]. Indeed, with most of classical SVM approaches for speaker verification, the dimension of the feature space induced by the sequence kernel is higher than the number of training utterances. In this context, separating the single target entry (+1) from all impostor entries (-1) can be easily completed, so training should be done with a hard margin criterion. This prevents from controlling the bias-variance trade-off as explicitly done by soft-margin SVM in standard situations, and also from taking into account the disproportion between positive and negative entries as recommended by [7].

With the new approach, suppose we have N background speakers in the development corpus, with $S > 2$ sequences for each speaker. Then we can simulate $NS(S-1)/2$ positive trials and up to $S^2N(N-1)/2$ negative trials.

4. A KERNEL BETWEEN PAIRS OF SEQUENCES

As said above, the main challenge is to design a suitable kernel between pairs of sequence. An option is to conceive an explicit map of pairs of sequences in a high-dimensional "Feature Space", where points corresponding to a priori similar sequences would be distant from points corresponding to dissimilar ones. Then, the kernel could be any vector kernel (dot product, polynomial, Radial Basis Functions, ...) between maps of pairs.

In this paper, we provide a solution based on the association of a probability distribution to each sequence. In practice, the distribution is a GMM estimated using a MAP adaptation of the UBM, as it is commonly done in speaker verification [8]. We then seek a suitable map of pair of GMMs that would distinguish between similar and dissimilar GMMs.

A natural and widely used measure of similarity between distributions is the Kullback-Leibler (KL) divergence [9]. This measure has been considered in several modeling strategies for speaker verification [10, 11, 12, 13]. In particular in [12], the log-sum inequality [14] is used as an upper bound of KL-divergence to derive the "GMM supervectors" kernel. In this paper we use this inequality to motivate the derivation our pair-of-sequences kernel.

4.1 Kernel construction

In the following, all GMMs share the same weights and diagonal covariance matrices[1]. Let G denote the number of Gaussian components, and g the Gaussian index shared by all GMMs. Let $\theta_{\mathrm{X}} = \{\omega_g, \mu_g^{\mathrm{X}}, \Sigma_g\}_{g=1\cdots G}$ be the set of weights,

[1] Indeed, a powerful method in speaker verification consists in training a UBM on a large database and then adapting only mean vectors to estimate GMM parameters on relatively short sequences: it enables to use complex models while avoiding over-fitting.

mean vectors, and covariance matrices of a GMM trained on a sequence X of vectors in $\Re^d$. The log-sum inequality [14] yields the following upper bound closed form expression for the KL-divergence $\mathscr{D}_{KL}$ between two GMMs:

$$0 \le \mathscr{D}_{KL}(\theta_{\mathrm{X}} \| \theta_{\mathrm{Y}}) \le \underbrace{\frac{1}{2}\sum_{g=1}^{G} \omega_g \left(\mu_g^{\mathrm{X}} - \mu_g^{\mathrm{Y}}\right)^{\top} \Sigma_g{}^{-1} \left(\mu_g^{\mathrm{X}} - \mu_g^{\mathrm{Y}}\right)}_{\mathscr{D}_{GMM}(\theta_{\mathrm{X}}, \theta_{\mathrm{Y}})} \tag{1}$$

The distance $\mathscr{D}_{GMM}$ defined in (1) can be seen as a weighted quadratic mean of Mahalanobis distances between Gaussian components. This shows that all the information required to compute the distance $\mathscr{D}_{GMM}$ (and thus to approximate $\mathscr{D}_{KL}$) is encoded in the Mahalanobis distances between corresponding Gaussian components. All these distances are normally small for similar Gaussians and high for dissimilar ones. Therefore, we can consider these distances as potential good candidates to form the map between pairs of GMMs, and by this way pairs of sequences.

We thus define the map of a pair of sequences/GMMs as the G-dimensional vector:

$$\begin{aligned} \phi(\{\mathrm{X},\mathrm{Y}\}) &= \Phi(\{\theta_{\mathrm{X}},\theta_{\mathrm{Y}}\}) \\ &= \begin{bmatrix} \sqrt{\left(\mu_1^{\mathrm{X}} - \mu_1^{\mathrm{Y}}\right)^{\top} \Sigma_1{}^{-1} \left(\mu_1^{\mathrm{X}} - \mu_1^{\mathrm{Y}}\right)} \\ \vdots \\ \sqrt{\left(\mu_G^{\mathrm{X}} - \mu_G^{\mathrm{Y}}\right)^{\top} \Sigma_G{}^{-1} \left(\mu_G^{\mathrm{X}} - \mu_G^{\mathrm{Y}}\right)} \end{bmatrix} \end{aligned} \tag{2}$$

Note that the norm of this map is simply $\mathscr{D}_{GMM}(\theta_{\mathrm{X}},\theta_{\mathrm{Y}})$.

Finally, given a vector kernel k, the kernel K between a pair of sequences is given by:

$$K\left(\{\mathrm{X},\mathrm{Y}\};\{\mathrm{X}',\mathrm{Y}'\}\right) = k\Big(\phi(\{\mathrm{X},\mathrm{Y}\});\phi(\{\mathrm{X}',\mathrm{Y}'\})\Big) \tag{3}$$

The vectorial kernel that led to the best results in our validation experiments is the widely used Gaussian RBF kernel:

$$K\left(\{\mathrm{X},\mathrm{Y}\};\{\mathrm{X}',\mathrm{Y}'\}\right) = e^{-\gamma\left\|\phi(\{\mathrm{X},\mathrm{Y}\}) - \phi(\{\mathrm{X}',\mathrm{Y}'\})\right\|^2} \tag{4}$$

where the locality parameter γ is tuned by cross-validation.

4.2 Model normalization

In [15, 13], performance of speaker verification systems based on the KL divergence are significantly improved when using a model normalization called "D-MAP". This normalization applies a linear transformation to the mean vectors of an adapted GMM so as to make all normalized GMMs equidistant from the reference UBM in terms of $\mathscr{D}_{GMM}$. Let $\theta_0 = \{\omega_g, \mu_g^0, \Sigma_g\}$ denote the parameters of the UBM. The D-MAP of the measure induced by the map Φ is given, for every Gaussian component g, by:

$$\begin{aligned} \overline{\mu}_g^{\mathrm{X}} = &\ \lambda_0^{\mathrm{X}} \mu_g^{\mathrm{X}} + \left(1 - \lambda_0^{\mathrm{X}}\right) \mu_g^0 \\ \text{with} \quad \lambda_0^{\mathrm{X}} = &\begin{cases} \dfrac{1}{2\|\Phi(\{\theta_0,\theta_{\mathrm{X}}\})\|_2} & \text{if } \theta_{\mathrm{X}} \neq \theta_0 \\ 0 & \text{otherwise} \end{cases} \end{aligned} \tag{5}$$

where $\|\cdot\|_2$ is the euclidean norm. Using this normalization, we finally consider in (2) the normalized map between pairs of sequences:

$$\overline{\phi}(\{\mathrm{X},\mathrm{Y}\}) = \Phi\left(\{\overline{\theta}_{\mathrm{X}},\overline{\theta}_{\mathrm{Y}}\}\right) \tag{6}$$

This map guarantees that all model pairs formed by the UBM and a normalized GMM has a half-unitary euclidean norm: $\|\Phi\left(\{\theta_0,\overline{\theta}_{\mathrm{X}}\}\right)\|_2 = 1/2$. Triangular inequalities using this property implies that the norm $\|\Phi\left(\{\overline{\theta}_{\mathrm{X}},\overline{\theta}_{\mathrm{Y}}\}\right)\|_2$ of the normalized map lies in the interval $[0,1]$, which is welcome for algorithm stability. Experiments show great improvement in performance when using such a model normalization.

5. EXPERIMENTS

5.1 Corpora

Experiments are carried out on the Biosecure project protocol [16]. NIST SRE 2003 and 2004 databases serve as two corpus involving distinct populations of female speakers: one is used to develop and the other one to evaluate. In traditional systems, the development data is partly used to train the UBM, or to feed discriminative training with impostor utterances. In our new system, this corpus is used to produce the training set of pairs of sequences.

All utterances (train and test sequences) contain roughly 2 minutes of conversational telephone speech. The development corpus involves about 100 speakers and includes 1616 speech utterances. To train our new classifier, 7500 pairs of sequences were formed from this set and SVM hyperparameters were tuned by cross-validation. The evaluation set consists in more than 17000 trials with about 200 target speakers.

5.2 Front-end Processing

As our main interest is in exploring modeling strategy, the front-end processing is a classical one in speaker verification. The signal frames are characterized by 32 coefficients including 16 Linear Frequency Cepstral Coefficients (LFCC) and their first derivative coefficients, obtained as follows. 24 filter bank coefficients are first computed over 20ms Hamming windowed frames at a 10ms frame rate. Filter bank coefficients are then converted to 16th order cepstral coefficients using a Discrete Cosine Transformation.

Then, speech activity detector is processed to remove low energy frames using the ALIZE toolkit [17], and parameter vectors are normalized to fit a zero-mean and unit-variance distribution on each sequence. The goal of this input normalization is to reduce channel effects.

5.3 Individual performance

Fig.3 shows the performance of our new Pair-of-Sequences (PoS) SVM system, with and without the D-MAP model normalization, using a Gaussian RBF kernel k in (3). It also shows the performance of two classical systems. The first one is a classical UBM-GMM system implemented with the ALIZE toolkit [17], with 2048 components, 10-best scoring [8] and T-norm score normalization [18]. We recall here that the UBM is used to estimate sequence distributions (by MAP adaptation) in the PoS SVM system. The second one is a classical SVM system using the GLDS kernel [19]. We thus have a generative and a discriminative system as reference for comparison.

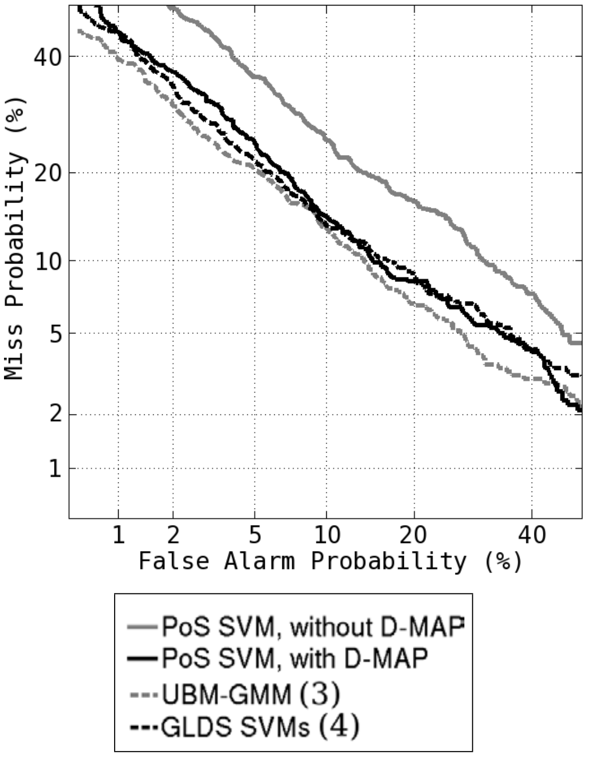

Figure 3: DET plots on NIST SRE 2004 of the new PoS SVM system, compared with two classical systems.

We can first see the high benefit provided by the D-MAP normalization. It is also interesting to note that, although it is based on a totally different strategy, the new PoS SVM system shows roughly comparable performance with UBM-GMM and GLDS-SVM. This suggests that our new concept makes great sense.

5.4 Fusion

More interestingly, Fig.4 shows the gain in performance provided by fusing the classical systems with the new one, with a linear combination of output scores. The weights of the combination, fixed for all trials, were determined by validation on the development corpus.

We can see that even if the GLDS-SVM system and our new PoS system show comparative individual performance, fusing them (even in a simple way) improves significantly. This fusion even outperforms the UBM-GMM. This confirms our initial guess about the potential complementarity of the new system with traditional ones. Nevertheless, fusing our new system with the UBM-GMM yields just a very slight improvement. This is probably due to the fact that both exploits the same "basic" information provided by the UBM.

Tab.1 sums up the results in terms of Equal Error Rates (EER) and Detection Cost Function (DCF). The DCF is a weighted sum of false alarm and false rejection rates as defined by NIST evaluation plans [20].

6. CONCLUSIONS AND FUTURE WORK

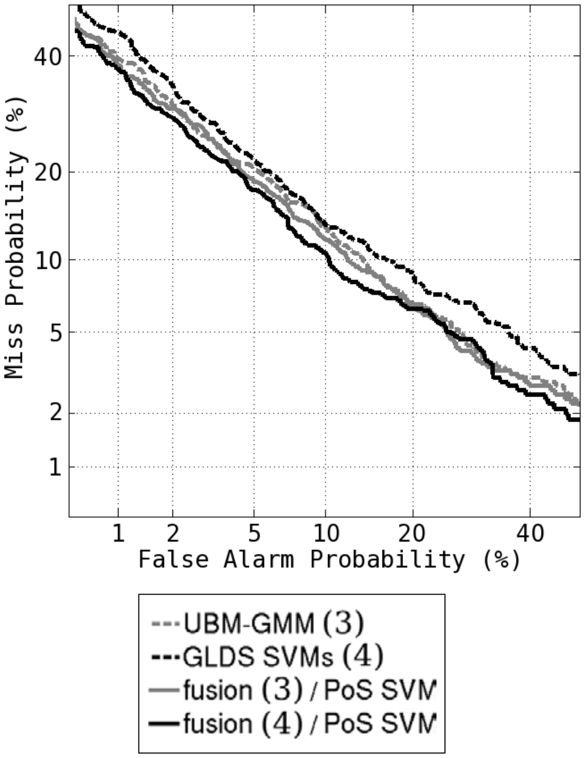

Figure 4: Fusion of the new system with classical systems

Table 1: Performance of classical systems and the new system

	EER	DCF ($\times 10^{-3}$)
(1) PoS SVM without D-MAP	17.28	67.3
(2) PoS SVM (with D-MAP)	12.29	53.4
(3) UBM-GMM	11.48	49.1
(4) GLDS SVMs	12.13	52.6
Fusion (2) / (3)	11.08	47.0
Fusion (2) / (4)	10.18	45.2

We presented a new concept of speaker verification based on a target-independent system that decides whether two utterances were pronounced by the same speaker. We developed a kernel between pairs of sequences and used it to implement the new concept in an SVM scheme. An efficient input normalization was also proposed to make the system more robust. The individual performance of the new system was comparable to two classical systems: UBM-GMM and GLDS-SVM. Moreover, its fusion with GLDS-SVM outperformed all the other systems. These results suggest that the new concept is worth considering and should be further investigated. Our future work will focus on generalizing the new concept to protocols where there are several training utterances per target speaker.

7. ACKNOWLEDGMENTS

The authors would like to thank Niko Brümmer for helpful discussions on this new way of considering the speaker verification task.

REFERENCES

[1] N. Brümmer, "Spescom datavoice and university of stellenbosch NIST2005 SRE system description," De-

scription Technique de système, workshop NIST SRE 2005, 2005.

[2] T.S. Jaakkola and D. Haussler, "Exploiting generative models in discriminative classifiers," *Advances in Neural Information Processing Systems*, vol. 11, 1998.

[3] S. Chen and P. Gopalakrishnan, "Speaker, environment and channel change detection and clustering via the bayesian information criterion," in *Proc. DARPA workshop*, 1998.

[4] A. Solomonoff, A. Mielke, M. Schmidt, and H. Gish, "Clustering speakers by their voices," in *Proc. ICASSP*, 1998.

[5] T. Gärtner, "A survey of kernels for structured data," *ACM SIGKDD Explorations Newsletter*, vol. 5, no. 1, 2003.

[6] J. Mariéthoz, *Discriminant Models for Text-independent Speaker Verification*, Ph.D. thesis, IDIAP, 2006.

[7] Y. Lin, Y. Lee, and G. Wahba, "Support vector machines for classification in non-standard situations," *Machine Learning*, vol. 46, pp. 191–202, 2002.

[8] D. Reynolds, T. Quatieri, and R. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital Signal Processing*, vol. 10, pp. 19–41, 2000.

[9] T.M. Cover and J.A. Thomas, *Elements of Information Theory*, Wiley, 1991.

[10] P. Moreno and P. Ho, "A new SVM approach to speaker identification and verification using probabilistic distance kernels," in *Proc. Eurospeech*, 2003.

[11] M. Ben, F. Bimbot, and G. Gravier, "Enhancing the robustness of bayesian methods for text-independent automatic speaker verification," in *Proc. Odyssey Speaker and Language Recognition Workshop*, 2004.

[12] W.M. Campbell, D.E. Sturim, D.A. Reynolds, and A. Solomonoff, "SVM based speaker verification using a GMM supervector kernel and NAP variability compensation," in *Proc. ICASSP*, 2006.

[13] N. Dehak and G. Chollet, "Support vector GMMs for speaker verification," in *Proc. IEEE Odyssey*, 2006.

[14] M.N. Do, "Fast approximation of kullback-leibler distance for dependence trees and hidden markov models," *Signal Processing Letters*, vol. 10, no. 4, pp. 115–118, 2003.

[15] M. Ben and F. Bimbot, "D-MAP : a distance-normalized MAP estimation of speaker models for automatic speaker verification," in *Proc. ICASSP*, 2003.

[16] "Biosecure network of excellence : Biometrics for secure authentification," http://www.biosecure.info, 2005.

[17] J.-F. Bonastre, F. Wils, and S. Meignier, "Alize, a free toolkit for speaker recognition," in *Proc. ICASSP*, 2005.

[18] R. Auckenthaler, M. 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.

[19] W.M. Campbell, "Generalized linear discriminant sequence kernels for speaker recognition," in *Proc. ICASSP*, 2002.

[20] "The NIST year 2004 speaker recognition evaluation plan," http://www.nist.gov/speech/tests/spk/2004/SRE-04_evalplan-v1a.pdf, 2004.