Speaker Recognition

The ATVS-UAM System at NIST SRE 05

**Joaquin Gonzalez-Rodriguez, Daniel Ramos-Castro, Doroteo Torre Toledano,
Alberto Montero-Asenjo, Javier Gonzalez-Dominguez, Ignacio Lopez-Moreno,
Julian Fierrez-Aguilar, Daniel Garcia-Romero & Javier Ortega-Garcia**
Universidad Autónoma de Madrid

ABSTRACT

Automatic Speaker Recognition systems have been largely dominated by acoustic-spectral-based systems, relying in proper modelling of the short-term vocal tract of speakers. However, there is scientific and intuitive evidence that speaker-specific information is embedded in the speech signal in multiple short- and long-term characteristics. In this work, a multilevel speaker recognition system combining acoustic, phonotactic, and prosodic subsystems is presented and assessed by blind submission to NIST 2005 Speaker Recognition Evaluation.

INTRODUCTION

Text-independent identification of speakers by their voices has been a subject of interest for decades for its potential use in areas such as intelligence and security. The first really successful results in actual telephone conversational speech came in the 1990s, where acoustic-spectral based systems [14] were able to obtain remarkable performance in really challenging out-of-laboratory tasks. The series of NIST Speaker Recognition Evaluations (SRE) has fostered research and development in this area since the mid-1990s [10]. This important forum has led to yearly significant improvements in the speaker recognition technology, which has been shared among participants to these evaluations. However, there was by that time significant room for improvement which was not taken into account in the use of higher non-acoustic levels of information. This information has demonstrated to be extremely characteristic in the inter-speaker communication process and well-known in linguistics, but it was not exploited at that time by automatic speaker recognition technology. It was in the early 2000s when the pioneering work on idiolectal differences between speakers [7] and especially the confluence of different sources of knowledge that were presented in the SuperSID project [16] gave a major impulse to multilevel and fusion approaches to automatic speaker recognition. Presently, multilevel speaker recognition systems may include generative [14] or discriminative [4, 12] acoustic-spectral sub-systems, prosodic [1], and phonotactic [3, 9] sub-systems among others.

In this contribution, a sample multilevel speaker recognition system is presented. Our research group, ATVS, has successfully participated in NIST 2001, 2002, and 2004 SREs with different progressively evolutioned versions of a UBM-MAP-GMM acoustic-spectral system, focused in the 1conv-1conv task (one side of a five-minute conversation for training – typically about two minutes of net speech – and one side of a different same size conversation for testing). However, for SRE 2005 we have also participated in the 8conv-1conv task (eight one-side conversations for training and one for testing), which allows a more effective use of high level sources of information, due to a higher amount of training data. This paper describes the different implemented systems, their individual assessment, their participation in blind conditions in NIST SRE 2005 8conv-1conv task, and an analysis of result, where the complementariness of the different levels of information is highlighted and the improvement obtained by the recently developed non-acoustic systems is objectively quantified.

ACOUSTIC SPEAKER RECOGNITION

Systems exploiting acoustic information are based on the short-term spectral identity information in the speech signal. Given a speech production model, we can argue that some

Author's Current Address:
J. Gonzalez-Rodriguez, D. Ramos-Castro, D.Torre Toledano, A. Montero-Asenjo, J. Gonzalez-Dominguez, I. Lopez-Moreno, J. Fierrez-Aguilar, D. Garcia-Romero and J. Ortega-Garcia, ATVS Biometric Research Laboratory, Escuela Politécnica Superior, Universidad Autónoma de MadridUniversidad Autónoma de Madrid, Campus de Cantoblanco, 28049 Madrid, Spain.

Refereeing of this work was handled by R.E. Trebits.

Manuscript received October 1, 2006.
Revised October 12, 2006.
Released for production October 16, 2006.

0885/8985/07/ $25.00 © 2007 IEEE

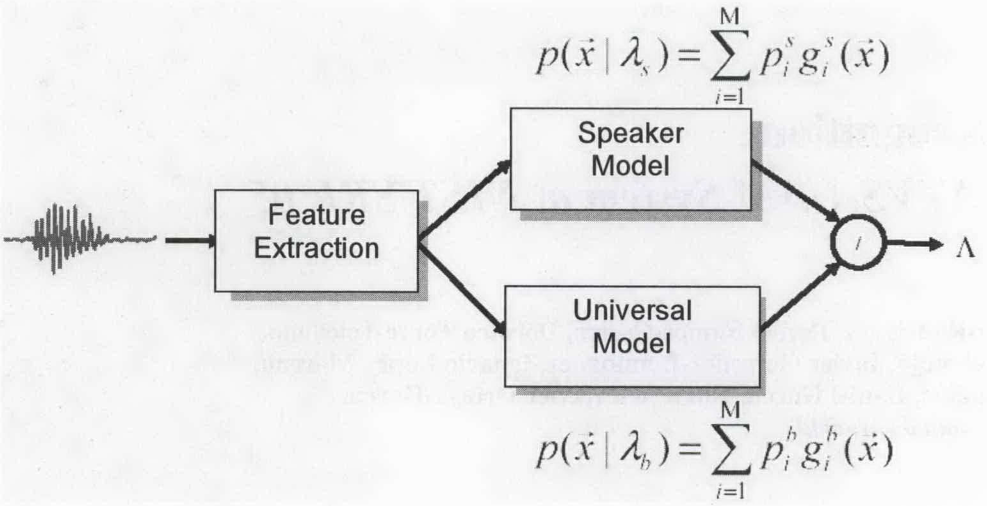
$p(\vec{x} \mid \lambda_s) = \sum_{i=1}^{M} p_i^s g_i^s(\vec{x})$
Speaker Model
Feature Extraction
Universal Model
$p(\vec{x} \mid \lambda_b) = \sum_{i=1}^{M} p_i^b g_i^b(\vec{x})$
/
Λ

Fig. 1. Likelihood ratio GMM score computation based on a speaker model and an alternate model *(Universal Background Model)*

spectral characteristics in the speech signal (formant distribution and variation, etc.) are related to speaker-dependent characteristics, such as vocal tract configuration. Therefore, this spectral information may be analyzed in order to recognize the speaker identity. Many feature extraction schemes have been proposed in the literature [6]. ATVS acoustic systems use Mel Frequency Cepstral Coefficients (MFCC) [6] obtained from a short-term windowing process. The speech signal is first windowed (using 20 ms. windows) and then each frame is processed, obtaining a MFCC vector per frame. Thus, each utterance is represented by a temporal stream of MFCC vectors.

Gaussian Mixture Models (GMM)

The state-of-the-art in text-independent speaker recognition has been widely dominated during the past decade by the Gaussian Mixture Model (GMM) approach working at the short-term spectral level [14]. This system exploits spectral characteristics of the speech in order to discriminate speakers. A GMM system will then use spectral features extracted from the speech signal in order to model speaker acoustic features in a statistical way.

The baseline ATVS GMM system is a likelihood ratio detector with target and alternative probability distributions modelled by Gaussian mixture models [14]. Briefly, let O be the set of d-dimensional feature vectors (observation vectors) representing a given utterance. Let λ_s be a speaker model, and an Universal Background Model (UBM), both represented as d-dimensional multivariate mixtures of Gaussians. The score can be computed by a likelihood ratio of both GMM models evaluated in each one of the observation vectors. Figure 1 represents the likelihood score computation process.

Speaker models in the described system are derived using Maximum A Posteriori (MAP) adaptation from the UBM using the Expectation Maximization algorithm [14]. MFCC feature extraction in order to obtain the O sequence for each utterance is performed as described above. Then, Feature Warping [11] has been used in order to compensate channel effects. The score normalization was performed by the KL-TNorm technique [13], an adaptive speaker-dependent cohort selection algorithm for T-normalization based on a fast estimation of Kullback-Leibler divergence for GMM models.

Support Vector Machines (SVM)

Support Vector Machines [12] are a discriminative learning technique based on minimum risk optimization, which aims at establishing an optimal separation boundary between classes. Because of their flexibility and their good performance in a variety of problems, they have been widely used in the last years. One of the main reasons of SVM success is the use of the so-called *kernel trick* [12], which maps each data vector into a high dimensional feature space where classes are linearly separable through a maximum margin hyperplane (MMH). Obtaining the MMH is a quadratic programming problem which can be solved with classical optimization techniques.

The objective of a SVM speaker recognition system is to obtain a likelihood score for the incoming speech taking into account the two classes involved: *target* and *non-target* speakers. From this discriminative approach, the score may be computed as a value proportional to the distance of the MMH to each vector by $score = w^{t} \bullet x$ where w is the MMH and x is the expanded featured testing vector to be classified. The kernel trick allows us to obtain this score as a function of: 1) the support vectors which represent the MMH; and 2) the testing vector to be classified. For each vector, the score is obtained without performing any explicit high dimensional mapping, and therefore the classification process is performed very efficiently [12]. The score for the whole testing utterance is finally computed as an average for all vectors extracted from it.

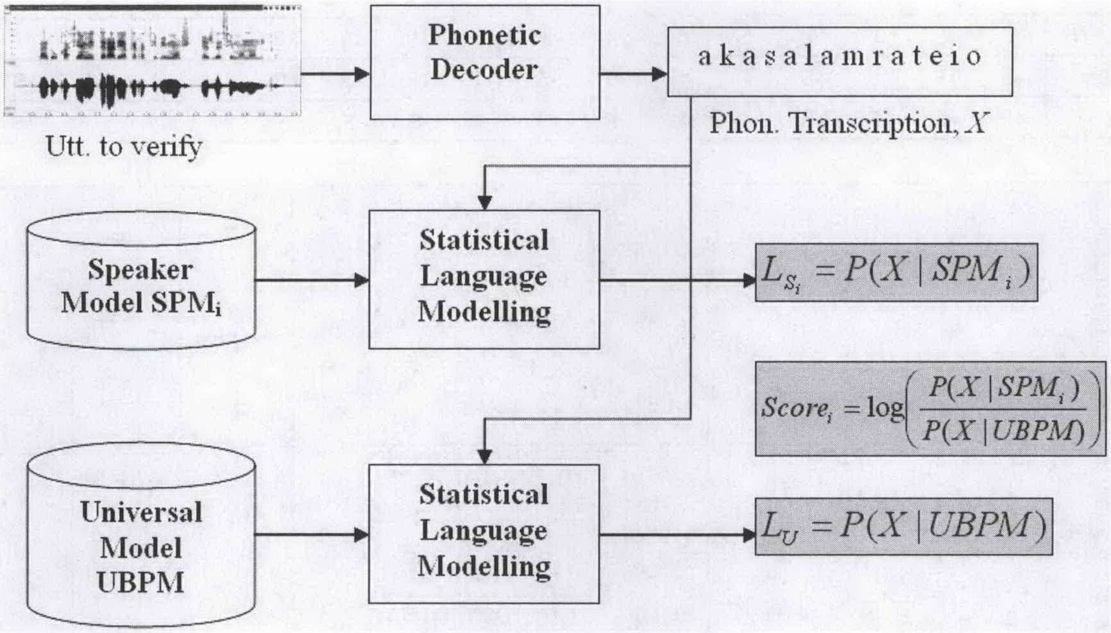
Utt. to verify
Phonetic Decoder
a k a s a l a m r a t e i o
Phon. Transcription, X
Speaker Model SPM$_i$
Statistical Language Modelling
$L_{S_i} = P(X \mid SPM_i)$
$Score_i = \log\left(\frac{P(X \mid SPM_i)}{P(X \mid UBPM)}\right)$
Universal Model UBPM
Statistical Language Modelling
$L_U = P(X \mid UBPM)$

Fig. 2. Verification of an Utterance against a speaker model in phonotactic speaker recognition

The ATVS SVM system uses the same MFCC parameters as in the GMM system described above. A spherical normalization has been performed in order to improve system accuracy. A channel compensation scheme has also been applied [17], as it has been demonstrated that channel variability seriously degrade the performance of acoustics SVM systems. The kernel trick has been applied by means of a second degree Generalized Linear Discriminant Sequence kernel proposed in [4].

HIGHER LEVEL SPEAKER RECOGNITION

Traditionally, automatic speaker recognition systems have relied only on the acoustic properties of speech, represented by statistical models like GMMs or discriminative models like SVMs (see the section entitled *Support Vector Machines*). However, recent research has shown that other features extracted from higher levels of information present in speech (e.g., pronunciation idiosyncrasies, linguistic content, prosody, etc.) can also be effectively used in automatic speaker recognition. In particular, numerous experiments have shown that, due to the complementary characteristics of acoustic and higher level features, the fusion of the information provided by these two features yields further improvements in speaker recognition.

The interest in the use of these higher level features was motivated by the work of Doddington [7], who used the lexical content of the speech, modeled through statistical language models (word n-grams), for speaker recognition using the Switchboard-II corpus. This relatively simple technique improved the results obtained by an acoustic-only speaker recognition system.

After the work of Doddington a number of research works have continued exploring the use of higher level features in the field of speaker recognition. Some of these works [2,3,9,16] made use of similar techniques (n-gram statistical language models) applied to the output of phonetic decoders (i.e. speech recognition engines configured to recognize any phonetic sequence), leading to the techniques known as *phonotactic speaker recognition*. Instead of modeling the lexical content, these techniques aim to model speaker pronunciation idiosyncrasies. This technique also yielded promising results, particularly when several phonetic decoders for different languages were used and combined. More recently, similar modeling techniques (n-gram statistical language models) have been applied to model the prosody (mainly fundamental frequency and energy) of the different speakers [1,16], giving rise to the field known as *prosodic speaker recognition*. As in the initial work of Doddington [7], all of these higher-level techniques were particularly useful in combination with traditional acoustic-only speaker recognition systems. In this section we describe in more detail our *phonotactic* and *prosodic* speaker recognition systems.

Phonotactic Speaker Recognition

A typical phonotactic speaker recognition system consists of two main building blocks: the *phonetic decoders*, which transform speech into a sequence of phonetic labels; and the *n-gram statistical language modeling stage*, which models the frequencies of phones and phone sequences for each particular speaker.

The phonetic decoders can either be taken from a preexisting speech recognizer or trained ad hoc. In our systems, phonetic decoders are based on Hidden Markov

TOKEN	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
FO	+F	+F	+S	+S	-F	-F	-S	-S	+F	+F	+S	+S	-F	-F	-S	-S	UV
E	+F	+S	+F	+S	-F	-S	-F	-S	-F	-S	-F	-F	+F	+S	+F	+S	*

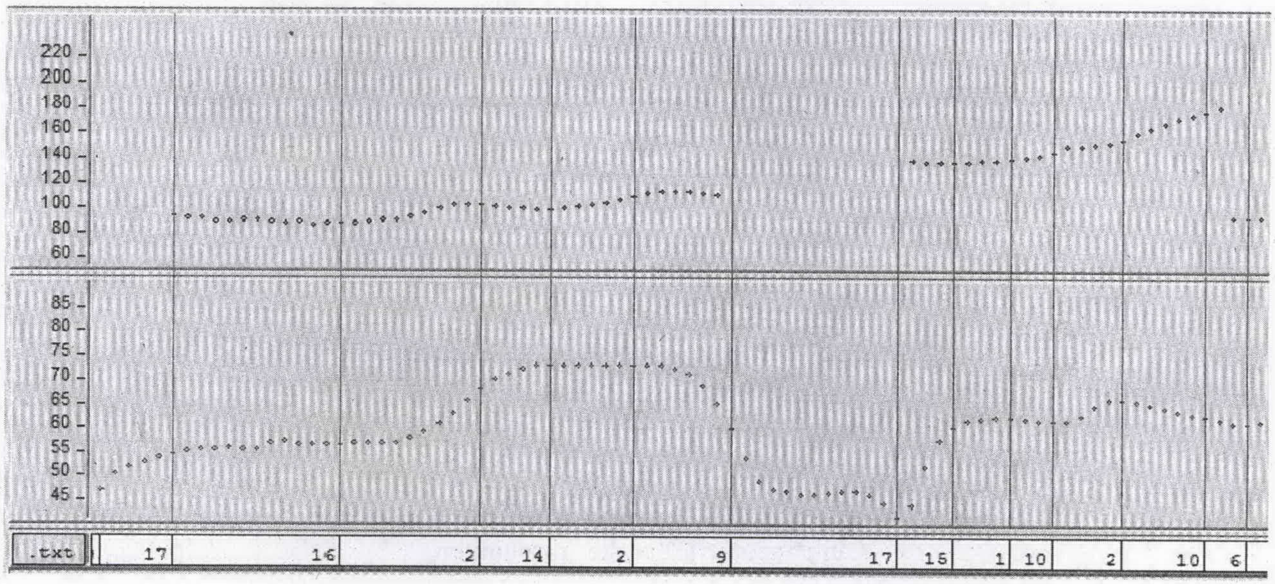

Fig. 3. Prosodic token alphabet (top table) and sample tokenization of pitch and energy contours (bottom figure)

Models (HMMs) and were implemented and trained ad hoc using the Hidden Markov Model ToolKit (HTK) (available for download at: <http://htk.eng.cam.ac.uk/>). The phonetic HMMs are three-state left-to-right models with no skips and the output probability density function of each state is modeled as a weighted mixture of Gaussians. These HMMs take as input speech features extracted using a standard front-end (the Advanced Distributed Speech Recognition Front-End defined by the European Telecommunications Standards Institute, ETSI, (available at: <www.etsi.org>). We trained context- independent phonetic HMMs for American English using the TIMIT corpus (available at: <www.ldc.upenn.edu>). 39 phones were considered for American English. At this point it is important to emphasize that, for the purpose of speaker recognition, it seems that it is not important to have accurate phonetic decoders and it is not even important to have a phonetic decoder in the language of the speakers to be recognized. This somewhat surprising fact has been analyzed by the authors [18] concluding that speaker-dependent phonetic errors made by the decoder seem to be speaker-specific, and therefore useful information for speaker recognition as long as these errors are consistent for each particular speaker.

Once a phonetic decoder is available, the phonetic decodings of many sentences from many speakers can be used to train a Universal Background Phone Model (UBPM) that models all possible speakers. Speaker Phone Models (SPM_i) are trained using several phonetic decoders of each particular speaker. Since the speech available to train a speaker model is often limited, speaker models are interpolated with the UBPM to increase robustness in parameter estimation. The optimal weight of the UBPM in this interpolation depends on several factors such as the amount of data available from the speakers and the complexity of the n-gram modeling and needs to be adjusted for each particular decoder. Once the statistical language models are trained, the procedure to verify a test utterance against a speaker model SPM_i is represented in Figure 2. The first step is to produce its phonetic decoding, X, in the same way as the decodings used to train SPM_i and UBPM. Then, the phonetic decoding of the test utterance, X, and the statistical models (SPMi, UBPM) are used to compute the likelihoods of the phonetic decoding, X, given the speaker model SPM_i and the background model UBPM. The recognition score is the log of the ratio of both likelihoods (Figure 2), where the higher the score the higher the similarity between training and test speech. This process may be repeated for different phonetic decoders (e.g., different languages or complexities) and the different recognition scores simply added or fused for better performance. For the experiments presented in this article, the language models used were trigram models.

Prosodic Speaker Recognition

Our prosodic speaker recognition system consists of two main building blocks: the *prosodic tokenizer*, which analyzes the prosody, and represents it as a sequence of prosodic labels or tokens and the *n-gram statistical language modeling*

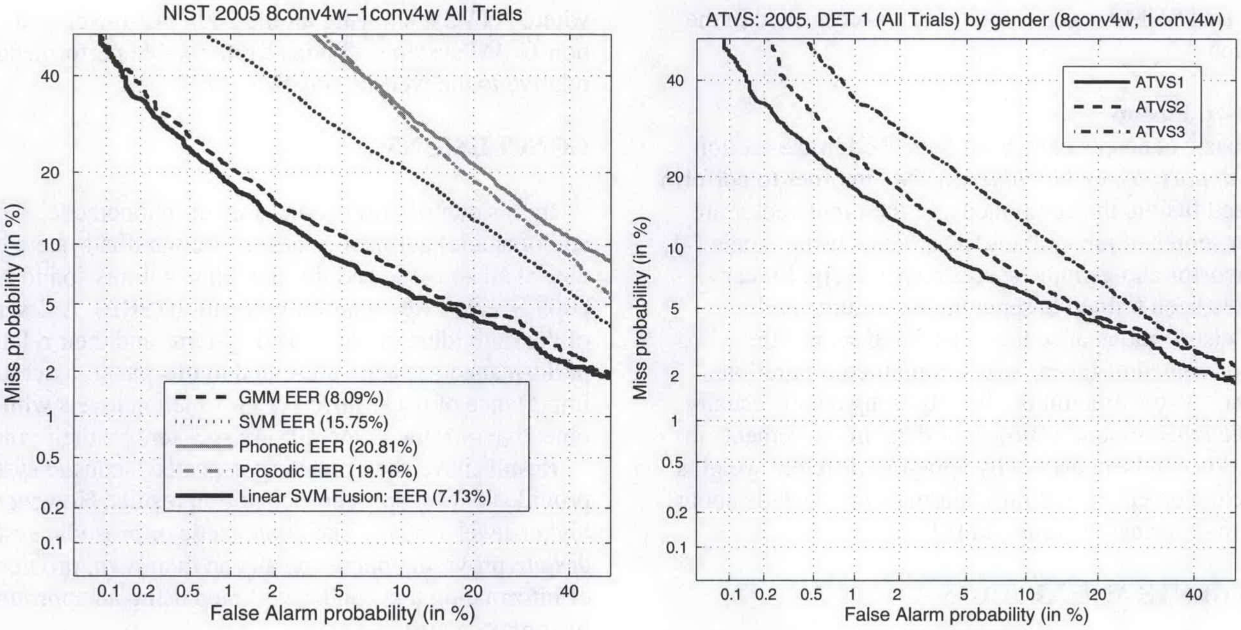

Fig. 4. NIST 2005 SRE ATVS subsystems and primary fusion results (left), and comparative performance of different submitted systems (right), where ATVS1 is our primary submissions, ATVS2 is similar to ATVS1 but with a different fusion strategy and ATVS3 is the fusion of all non-GMM systems

stage, which models the frequencies of prosodic tokens and their sequences for each particular speaker. This second block is exactly the same for phonetic and prosodic speaker recognition with only minor adjustments to improve performance (e.g., adjusting the weight of the universal model in the generation of the speaker model). For this reason this second block will not be described herein.

The tokenization process carried out in our system consists of two stages. First, for each speech utterance, both temporal trajectories of the prosodic features, (fundamental frequency – or pitch – and energy) are extracted. Second, both contours are segmented and labelled by means of a slope quantification process.

To extract contours, the Praat toolkit (available for download at: <www.praat.org>) was used. The slope quantification process was performed as follows: first, a finite set of tokens were defined using a four level quantization of the slopes (fast-rising, slow-rising, fast-falling, slow-falling) for both energy and pitch contours [1]. Thus, the combination of levels generate sixteen different tokens when combined pitch and energy contours are considered. Second, both contours were segmented using the start and end of voicing and the maximums and minimums of the contours. These points were detected as the zero-crossings of the contours derivatives using a ±2 frame span. On the other hand, silence intervals were detected with an energy-based voice activity detector. Finally, each segment was converted into a set of tokens which describe the joint-dynamic variations of slopes. Therefore, utterances with different sequences of tokens contain different prosodic information.

Since errors in the pitch and energy estimation are likely to generate small segments, all segments smaller than 30 ms were removed from the sequence of joint-state classes. Three special tokens were further included: 1) token UV, which represents unvoiced regions, and 2) tokens <s> and </s> as utterance delimiters. Figure 3 shows all possible tokens used to describe the speech utterances, and an example of a segmented utterance.

MULTI-LEVEL FUSION FOR IMPROVED SPEAKER RECOGNITION

There are many works related to the combination of different speaker characteristics and modelling methods for a speaker verification system, such as [5,7,8]. State-of-the-art systems as [15] are commonly not a single system but the fusion of several. The performance improvement of a fused system is based on the fact that different systems provide different information about the speaker, and therefore errors committed by a certain system may be cancelled out by other systems. In fact, the potential benefits from fusion increase with the uncorrelation between the involved systems. Fusion can be performed at different stages of the process, but the most common approach is to fuse individual scores provided by each system. At that stage, fusion strategies can be based on rules (as sum fusion or product fusion rules) but the problem can also be considered as a pattern classification problem, and therefore almost any classification technique like Gaussian-class classifiers, Neural Networks, and SVMs can be applied. In this article we have used SVM-based fusion, which is described in the next section.

SVM-Based Fusion

SVM basic concepts have been described in the section entitled *Support Vector Machines* (SVM) In order to perform SVM-based fusion, the components of the input vector are the output scores of the systems to be fused, using labels {-1, 1} for impostor and genuine scores respectively. Linear SVMs have been trained to separate the genuine and impostor distributions of scores. The fused scores are obtained as signed distances to the computed separating hyperplane. As the amount of client training data is usually smaller than the amount of impostor data, improvements in classification can be achieved by applying different weights to false rejection errors and false alarm errors. Details about these techniques may be found in [8].

EXPERIMENTS AND RESULTS

In order to assess the performance of the multilevel speaker recognition system, the 8side-1side task of NIST SRE 2004 has been used as a reference benchmark. Later, the submitted systems were assessed (after NIST SRE 05) with the evaluation keys (the "solutions)." A good match between both conditions (SRE 04 and 05) is expected if systems are properly designed, as the origins of the data in both evaluations was mostly the same. In fact, our experiments showed a match so good between the development (SRE 04 data) and test (SRE 05 blind data) conditions that the figures obtained are virtually the same, which highlights the good generalization of our systems. Figure 4A shows the results of all submitted ATVS individual systems in the 8conv-1conv SRE 05 task, as well as the SVM fusion of all. This task contained about 500 speaker models trained with 8 telephone conversations about 5 minutes each. These models were tested with single telephone conversations of about 5 minutes, where a total of over 23,000 trials of this kind were performed. Our newly-developed phonotactic and prosodic systems work clearly worse than the other (acoustic) systems, which was consistently found by other researchers, perhaps because the amount of prosodic and phonotactic information for this type of modeling is smaller than the acoustic information provided by the same amount of speech. It is worth noting, at this point, that our phonotactic and prosodic systems performed similarly to the best phonotactic and prosodic systems submitted to NIST SRE 2004. On the acoustic systems, our SVM system performs clearly worse than our GMM system. The main reason for this is that our GLDS-SVM system by that time for implementation reasons performed just second-order polynomial expansion, where third-order is mandatory to obtain competitive performance, as we have obtained after the evaluation.

Figure 4A shows that a significant improvement relative to the GMM performance (the unique ATVS 2004 system) is obtained with the inclusion of the 2005 just-developed systems. An important result is also shown in Figure 4B where ATVS3, showing all the 2005 just-developed non-GMM systems, obtains a remarkable performance relative to the well-established GMM One.

CONCLUSIONS

In this contribution, a multi-level (phonotactic, acoustic and prosodic) automatic speaker recognition system has been described and assessed through blind submission to NIST 2005 Speaker Recognition Evaluation (SRE). A description of the individual implemented systems and their relative performance has been presented in this paper, assessing the importance of using different information levels with the objective of reliably identifying speakers by their voices.

Results have shown that, as expected, acoustic systems provide the best speaker recognition results. However higher-level systems like phonotactic or prosodic systems, despite providing poorer results on their own, provide plenty of information that can be exploited using an appropriate fusion mechanisms.

REFERENCES

[1] A.G. Adami et al,
Modeling Prosodic Dynamics for Speaker Recognition,
in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Hong-Kong, China, 2003. Vol. IV, pp. 788-791.

[2] W. Andrews et al.,
Phonetic, idiolectal, and acoustic speaker recognition,
in Proceedings of ODYSSEY Workshop, 2001.

[3] W. Andrews et al.,
Gender-dependent phonetic refraction for speaker recognition,
in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2002, Vol. 1, pp. 149-152.

[4] W.M. Campbell,
Generalized linear discriminant sequence kernels for speaker recognition,
in Proceedings of the International Conference on Acoustics Speech and Signal Processing, 2002, pp. 161-164.

[5] W.M. Campbell, D.A. Reynolds and J.P. Campbell,
Fusing Discriminative and Generative Methods for Speaker Recognition: Experiments on Switchboard and NFI/TNO Field Data,
in Proc. of ODYSSEY 04, pp. 41-44,Toledo, Spain.

[6] J.R. Deller et al., 1999,
Discrete-Time Processing of Speech Signals,
Wiley-IEEE Press.

[7] G. Doddington,
Speaker recognition based on idiolectal differences between speakers,
in Proceedings of EUROSPEECH, Vol. 4, pp. 2517-2520, Denmark, 2001.

[8] D. Garcia-Romero, J. Fierrez-Aguilar, J. Ortega-Garcia and J. Gonzalez-Rodriguez,
Support Vector Machine fusion of idiolectal and acoustic speaker information in Spanish conversational speech,
in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP,
Vol. 2, pp. 229-232,
Hong Kong, April 2003.

[9] Q. Jin et al.,
Phonetic Speaker Identification,
in Proc. International Conference on Spoken Language Processing,
ICSLP 2002, pp. 1345-1348.

[10] NIST,
Speaker Recognition Evaluations,
http://www.nist.gov/speech/tests/spk/.

[11] J. Pelecanos and S. Sridharan,
Feature Warping for Robust Speaker Verification,
in Proceedings of A Speaker Odyssey, Paper 1038, 2001.

[12] F. Perez-Crus and O. Bousquet, 2004,
Kernel Methods and their Potential Use in Signal Processing,
IEEE Signal Processing Magazine
(Special issue on Signal Processing for Mining).

[13] D. Ramos-Castro et al., 2005,
Speaker verification using fast adaptive Tnorm based on Kullback-Leibler divergence,
Proceedings of 3rd COST 275 Workshop,
Hatfield, UK.

[14] D.A. Reynolds et al., 2000,
Speaker Verification using Adapted Gaussian Mixture Models,
Digital Signal Processing, Vol. 10, pp. 19-41.

[15] D.A. Reynolds et al.,
The 2004 MIT Lincoln Labs Speaker Recognition System,
In Proceedings of ICASSP 2005, pp. 177-180.

[16] D. Reynolds et al.,
SuperSID Project Final Report: Exploiting High-Level Information for High-Performance Speaker Recognition,
Retrieved on March 3, 2005 from http://www.clsp.jhu.edu/ws2002/groups/supersid/SuperSID_Final_Report_CLSP_WS02_2003_10_06.pdf.

[17] Solomonoff, A.C.,
Advances in channel compensation for SVM speaker recognition,
ICASSP 2005

[18] D.T. Toledano et al,
On the Relationship between Phonetic Modeling Precision and Phonetic Speaker Recognition Accuracy,
in Proceedings of the 9th European Conference on Speech Communication and Technology
(EuroSpeech-InterSpeech),
Lisbon, Portugal, 5-8 September 2005. pp. 1993-1996.