On the use of high-level information in speaker and language recognition

Alberto Montero-Asenjo, Javier Gonzalez-Dominguez, Daniel Ramos-Castro, Ignacio Lopez-Moreno, Doroteo Torre Toledano y Joaquin Gonzalez-Rodriguez

ATVS (Speech and Signal Processing Group)
Escuela Politecnica Superior
Universidad Autonoma 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 using NIST 2005 Speaker Recognition Evaluation data.

For language recognition systems, the NIST 2005 Language Recognition Evaluation was selected to measure performance of a high-level language recognition systems.

1. Introduction

Speaker recognition systems are automatic systems that provide information about the identity of the speaker of a given speech segment. As a verification system it must decide whether or not the identity of the speaker is a claimed one. As an identification system the goal is to determine the identity of the speaker among a set of predefined speakers.

Regardless the operation mode, the system works by computing similarity measures (scores) between the speech segment and a previously stored model (extracted from other speech segment). Based on the computed measures, the system may perform a hard decision (yes/no, accepted/rejected) or a soft one, providing a score for a subsequent module. For the hard decisions case, a threshold must be determined, and the decision emerge from the comparison between the score and the threshold.

Despite the fact that text-independent identification of speakers by their voices has been a subject of interest for decades, the first really successful results in actual telephone conversational speech came in the 90s, where acoustic-spectral based systems [1] 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-90s [2]. 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 early 00s when the pioneering work on idiolectal differences between speakers [3] and specially the confluence of different sources of knowledge that were presented in the SuperSID project [4] gave a major impulse to multilevel and fusion approaches to automatic speaker recognition. Presently, multilevel speaker recognition systems may include generative [1] or discriminative [5] acoustic-spectral sub-systems, prosodic [6], and phonotactic [4] sub-systems among others [7].

2. Speaker recognition techniques overview

Speaker recognition techniques can be broadly classified as those relaying in short-time spectral and acoustic information and those using high-level features (phonetic, prosodic, lexical information, etc.). A third group may be considered, in which would be included those systems made by combination (fusion) of several other systems.

Short-time spectral information systems include Gaussian Mixture Models (GMM)[1] and Support Vector Machines (SVM) [5]. GMM has been the reference in the past decade, but SVM systems are now competitive.

High-level features based systems include prosodic and phonotactic based systems. In both situations, the speech is converted to streams of tokens used to compute the score. Prosodic systems use pitch and energy as tokens while phonotactic systems use phones in some language.

GMM and SVM systems perform much better than prosodic or phonotactic systems, but the combination of them may improve the final systems [8][7].

3. High-level speaker recognition techniques

The interest in the use of these higher level features was motivated by the work of Doddington [3], 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 [4] 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 [4][6], giving rise to the field known as prosodic speaker recognition. As in the initial work of Doddington [3], all of these higher-level techniques were particularly useful in combination with traditional acoustic-only speaker recognition systems.

3.1. Statistical grammar modelling

The most common modelling technique for tokens sequences is statistical modelling, where the probability of a sequence given a language model is used as the basis for scoring.

Given a sequence of tokens (words, phones, prosodic tokens, data driven units, etc.)

$$S = (w_1, w_2, \ldots, w_m)$$

the probability of occurrence can be decomposed as a product of conditional probabilities

$$P(w_1, w_2, \ldots, w_m) = \prod_{i=1}^{m} P(w_i \mid w_1, \ldots, w_{i-1}) \quad (1)$$

Usually eq.1 is approximated by limiting the context:

$$P(w_1, w_2, \ldots, w_m) \simeq \prod_{i=1}^{m} P(w_i \mid w_{i-n+1}, \ldots, w_{i-1}) \quad (2)$$

for some $n \geq 1$. Due to reasons of data sparsity n is usually selected in the range of 1 to 4.

Estimates of probabilities in n-gram models are commonly based on maximum likelihood estimates – that is, by counting events in context on some given training text:

$$P(w_i|w_{i-n+1}, \ldots, w_{i-1}) = \frac{C(w_{i-n+1}, \ldots, w_i)}{C(w_{i-n+1}, \ldots, w_{i-1})} \quad (3)$$

where $C(.)$ is the count of a given word sequence in the training text. For robust estimation, probability smoothing techniques can be applied.

3.2. Process overview

The process starts with speaker models training, from speech transcriptions or tokenization. Direct robust estimation of probabilities requires a big amount of data, usually not available for a single speaker. The procedure to train more robust models is to train the UBM model using a lot of data and then adapt that model to the data available for a particular speaker. This adaptation is made by linearly interpolating the n-grams models of the UBM and the one trained only with speaker data. That interpolation is governed by an weighting factor that has to be empirically determined.

To obtain a score related to a given sequence, usually a log-likelihood detector is used. The target model SPM_i will be the n-gram model adapted to transcriptions from speech from the speaker. The alternate model will be a n-gram model trained with transcriptions from speech from many speakers (UBM model). The final score for sequence S of m tokens is obtained as

$$s_i = \frac{1}{m} \log \frac{P(S/SPM_i)}{P(S/UBM)} \quad (4)$$

where $P(S/SPM_i)$ and $P(S/UBM)$ are calculated as shown in eq. 2.

3.3. Phonotactic systems

Phonotactic systems use phonetic transcribers to convert speech into a sequence of tokens where each token is a phone.

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 pre-existing speech recognizer or trained ad hoc. Any speech recognition technology can be used, but usually phonetic decoders are based on Hidden Markov Models and null grammars.

Reported experiments have been performed using phonetic decoders for Castillian Spanish (using the Albayzin Corpus [9]), American English (using the TIMIT corpus) and Basque (using Basque-SpeechDAT corpus).

Once the phonetic sequence has been obtained the scoring process is performed as explained above.

3.4. Prosodic systems

A prosodic speaker recognition system consists of two main building blocks: the prosodic tokenizer, which analyses the prosody, and represents it as a sequence of prosodic labels or tokens, and the n-gram statistical language modeling stage, which models the frequencies of prosodic tokens and their sequences for each particular speaker, as explained above. The tokenization process carried out consists of two stages. Firstly, for each speech utterance, both temporal trajectories of the prosodic features, (fundamental frequency or pitch- and energy) are extracted. Secondly, both contours are segmented and labelled by means of a slope quantification process. 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 [6] .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 a certain amount (tipically 30 ms) are removed from the sequence of joint-state classes.

Additionally to the sixteen tokens defined for the joint-dynamic of the prosodic features, a special token must be added to represent unvoiced segments.

3.5. Data-driven phonotactic systems

Both phonetic and phonotactic systems use as tokens items with linguistic meaning: phones, energy trajectories, pitch. But from an engineering point of view the tokens do not need to be limited to these. Some researchers have proposed the use of units extracted automatically from speech samples.

ALISP systems are based on this idea. Units are similar to phones but the information modelled by each unit is determined by means of an automatic clustering process. Results reported

in [10] show that this approach can outperform classical phonetic systems.

3.6. PhoneSVM systems

PhoneSVM systems share many elements with a phonotactic system. Both rely on phonetic transcriptions and n-gram model estimation, but differ in the way the score is computed. While in classical phonotactic systems the scores are calculated as log-likelihood ratios, in phoneSVM systems a SVM is used to determine the score.

The training process starts by estimating n-gram models. One for the speaker under consideration and many more from other different speakers. The probabilities of the different n-grams are considered as the coordinates of a vector, thus transforming the problem into a one suitable for training a SVM. Results presented in [11] show major improvements.

3.7. Phonetic decoding using phone lattices

Phone lattices is a technique initially proposed for language recognition [12] that was later successfully applied to speaker recognition [13].

The underlying idea is how to increase the amount of data available for n-gram estimation. Extracting a single phone sequence from a speech segment to estimate n-grams is an idea borrowed from the speech recognition field, where only one sequence is useful as the transcription of a speech segment (unless further processing is performed, but at last, only one sequence will be output). But in speaker and language recognition fields, the sequence itself is not important, but for estimation purposes, so if instead of considering the best sequence, the N best sequences are considered, the amount of data for parameter estimation largely increases, thus relying to better estimates. The decrease in error rates (as reported in [12] and [13]) largely compensates the increase in execution time.

4. Multilevel speaker recognition techniques fusion

There are many works related to the combination of different speaker characteristics and modelling methods for a speaker recognition system, such as [14][3][8]. State of the art systems as [15] are commonly not a single system but the fusion of several of them. 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.

Some research has been done in the adaptation of fusion schemes to each user [16]. While there is some research effort in fusion of multiple biometrics systems [17][18], it is not very common within different speaker recognition systems, but improvements can be obtained from that adaptation.

The effect of fusion on error rates will be shown in sections 5 where specific systems are described and tested on reference data.

5. The ATVS speaker recognition system at NIST 2006 Speaker Recognition Evaluation

The evaluation is a speaker detection task, where the goal is to determine if a specified speaker is present in a given segment of conversational speech. Systems must provide a decision about the segment (T/F) and a confidence score.

The different task conditions are defined by a training and a test condition. For the core task, also called 1conv4w-1conv4w, 5 minutes of speech are provided (before silence removal) and other five minutes for testing.

ATVS systems have been tested in the core task, 8conv4w-1conv4w task and 1conv4w-10sec4w. For the 8conv4w training condition 8 speech segments of about 5 minutes are provied, and in the 10sec4w testing condition only 10 seconds of speech are available. Four different systems were developed and fused in different ways for the different tasks.

5.1. Development process

For the development process several partions of data were defined. For the training of the UBM model for all systems, data from NIST 2005, NIST 2004, SWITCHBOARD I and SWITCHBOARD II Extended-data task was used.

5.2. GMM system

A root UBM (needed for feature mapping[19]) was trained using 5 hours of channel- and gender-balanced speech after silence removal. Data from MIXER (NIST SRE 2004 and 2005), Switchboard I and Switchboard II was used. The UBM was trained using 1024 Gaussian mixtures and ML estimation via EM algorithm. Fourteen channel models (7 per gender) were adapted from the UBM in order to perform Feature Mapping. An average value of 2 hours of speech was used for each channel model training.

Target models were 1024 mixtures GMM, MAP adapted with one iteration (only means) from the 1024 root UBM. Only 5 Gaussian per frame were used in likelihood computations.

Score normalization was performed using Tnorm [20] and KL-Tnorm [21].

5.3. Acoustic SVM system

The acoustic SVM system uses a explicit normalized three degree polynomial expansion [22] followed by a decomposed Generalized Linear Discriminant Sequence Kernel (GLDS) as described in [5]. SVMTorch [5] was used to train the target models.

5.4. Prosodic system

The submitted prosodic systems was similar to the one described in section 3.4.

HTK 3.2.1 n-gram modeling tools [23] are used to train gender-dependent UBMs and the target-speaker models. Male and female training data from NIST 2005, NIST 2004, SWITCHBOARD I and SWITCHBOARD II Extended-data task has been used to train the male and female UBMs, respectively. N-gram modelling used trigrams. The target-speaker models are created by linear interpolation of the corresponding UBM (gender-dependent) and the speaker training data. The interpolation coefficients are set to 0.8 for the speaker data and 0.2 for the UBM. By including the general knowledge provided by the UBM into the target-speaker models, the amount of data needed for a good estimation of the trigram models is reduced. TNorm technique was applied for score normalization. Cohorts consist of 60 models from NIST SRE 2004 database.

5.5. Phonotactic system

The phonotactic system was built over three independent speech recognisers (English, Spanish and Basque) based on Hidden Markov Models (HMMs). The HMM topology is three- state left-to-right with no skips. The output pdfs of each state are modelled as GMMs. The number of Gaussians per state were adjusted on the NIST SRE05 data task corpus to minimize speaker recognition EER.

Two schemes have been used for feature extraction:

- The Advanced Distributed Speech Recognition Standard Front-End defined in the standard ETSI ES 202 050 [24].
- Sphinx [25] feature extraction system. This system is based on 13 MFCC coefficients along with delta and double delta coefficients and C0.

The set of English phone HMMs was trained on the TIMIT corpus. Since this corpus is microphone speech sampled at 16 kHz, audio was filtered to simulate the telephone channel and then downsampled it to 8 kHz. One Gaussian/state was used to model output pdfs. The set of Spanish phone HMM was trained on the Albayzin corpus. The same subsampling process as described above was applied for this case. Five gaussians/state were used to model output pdfs. Both systems use the ETSI ES 202 050 parameterizer. Basque SpeechDAT was used in order to train the Basque phone HMM set, modelling output pdfs with 20 gaussians per state. The parameterisation was performed using the Sphinx parameteriser. All sets were trained using HTK v3.2.1.

Acoustic-phonetic decoding (phone recognition) was performed with every recogniser on Switchboard I, Switchboard II, NIST SRE04, NIST SRE05, NIST SRE06 train and test files using HTK v3.2.1, the trained models and a null grammar. The only information used from the acoustic-phonetic decoding was the phone streams. The output phone streams were filtered to avoid repetitions of inter-word silences.

The Universal Background Phone Model (UBM) is a trigram language model trained with data from Switchboard I, Switchboard II, NIST SRE04 and NIST SRE05. Smoothing of unlikely trigrams was performed with absolute discounting. No cut-off factor was applied. A different UBM was used for each phonetic decoder. Speaker Phone Models (SPM_i) are created by linear interpolation of the 8 sides training material for each target speaker from NIST SRE06 training data. The interpolation factor (weight of the UBM) for this adaptation was adjusted on NIST SRE05 extended data task and was found to be optimal for an UBM weight of 0.7.

For score normalization Tnorm was applied using as cohort a gender dependent set of 60 models extracted from NIST SRE04.

5.6. The system for the 8conv4w-1conv4w task

High-level systems require lots of data for reliable parameter estimation, so they can only be applied to certain tasks. The 8conv4w-1conv4w task provides about 40 minutes of untranscribed speech for model training (prior to silence removal) and 5 minutes for testing.

The submitted system for this task was a combination of all available systems: GMM, SVM, prosodic system, phonotactic systems. All systems were fused using a linear SVM.

For the development stage, NIST05 data was used to train and test the fusion rule. This makes shown results a little bit optimistic, but as the used classifier is a very simple one (a linear SVM) the overfitting is not expected to be strong. With this setup, the obtained results are shown in figure 1. This figure shows that acoustic systems (GMM's and SVM's) perform much better than high-level systems, but the fusion of all improves results significantly.

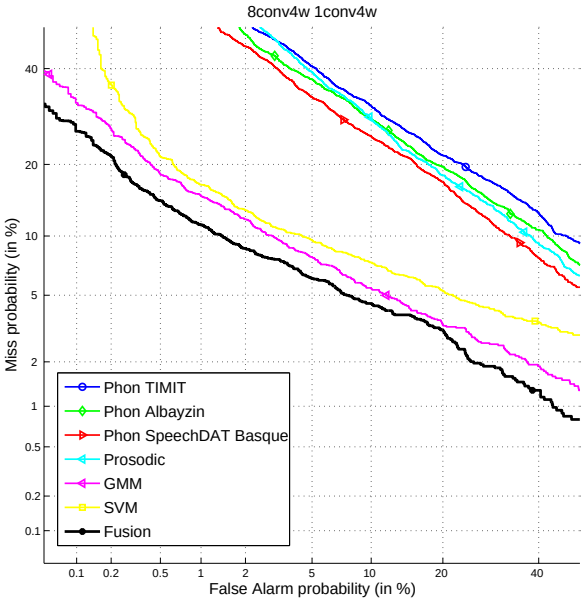

Figure 1: 8conv4w-1conv4w task involved systems results

6. Application of high-level speaker recognition techniques to language recognition

The language recognition field shares many techniques with the speaker recognition field. In [26] there is a detailed explanation of several techniques for language recognition.

First research efforts showed that high-level systems performed better than acoustic systems [26], but there have been important improvements in acoustic systems, as can be shown in [27] and [28]. Both techniques are applied in similar ways in both fields.

One of the most common technique for language recognition is an extension of phonotactic systems called Parallel Phone Recognition and Language Modelling (PPRLM). Basically it consists on the fusion of several phonotactic systems as described in section 3.3 related to phonetic decoders in several languages, not necessarily related to the target ones. Using the transcriptions, statistical grammars are applied and the scoring process is performed in the same way as for speaker recognition. Sum fusion is the most common applied fusion technique. In order to train each of the underlying phonetic recognisers, multilingual speech corpus are required, but they do not need to contain labelled speech in the target language. The only requirement is to have labelled in a certain number of language (and in the appropriate amount to train a phonetic recogniser). The most common corpus for this purpose is OGI Multilingual Telephone Speech [29]. This corpus has speech in 11 languages, and for 6 of them it contains labelled speech. Other corpora used to train language recognition systems is SpeechDAT. The main advantage of SpeechDAT over OGI Mul-

tilingual Telephone Speech is that the former is a much larger database, which seems to be important, as shown in [30].

Language recognition can also be performed using speech recognisers in the target languages. This technique is called Parallel Phone Recognition (PPR) and it basically consists of N parallel speech recognisers (one for each target language), each one providing a speech recognition score, used for decisions. To perform language recognition using PPR labelled speech in all target languages is required, in order to train the speech recognisers. As this may be difficult for certain languages, PPR is not very common.

6.1. The ATVS PPRLM speaker recognition system at NIST 2005 Language Recognition Evaluation

For the 2003 evaluation, twelve target languages were defined (Arabic, English, Farsi, French, German, Hindi, Japanese, Korean, Mandarin, Spanish, Tamil and Vietnamese). Test material was extracted mainly from the CallFriend corpus from LDC, with some test segment from other corpora.

In the 2005 evaluation, the target languages set was reduced to seven (English, Hindi, Japanese, Korean, Mandarin, Spanish and Tamil) along with some dialectal variations (American/Indian English, Mainland Mandarin and Taiwanese Mandarin). A new database was collected, including a big amount of cellular data, that was not present in previous evaluations data.

The testing segments have nominal durations of 30, 10 and 3 seconds.

The full description of submitted system along with the details of the development process can be found in [31].

The best submitted system was a fusion of two PPRLM systems (as described above) based on phonetic recognisers trained with the OGI Multilingual Telephone Speech corpus. Both PPRLM systems were the fusion of 6 language recognisers based on phonetic recognisers of Mandarin, German, Japanese, Spanish, Hindi and English, using in all cases trigrams. The two PPRLM systems were similar but differ in the the number of Gaussian used to model HMM state output pdf's in the phonetic recognisers. To normalise scores, Tnorm was used. (This system appears in plots as ATVS1).

Other system was submitted, but consists only of a single PPRLM system, built using 10 Gaussian per state. (This system appears in plots as ATVS2).

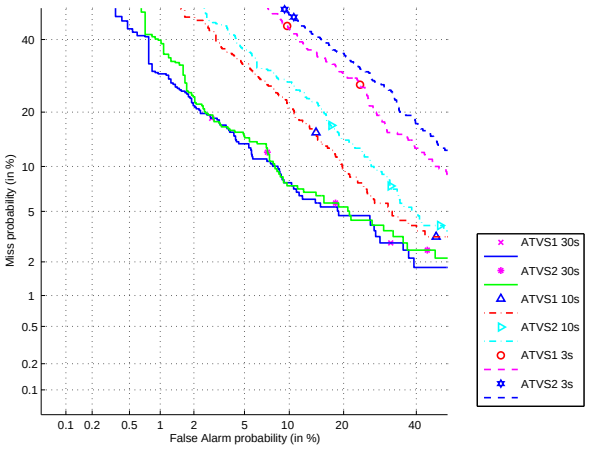

Figure 2: ATVS PPRLM results tested on a NIST 2003 LRE subset involving only NIST 2005 LRE target languages.

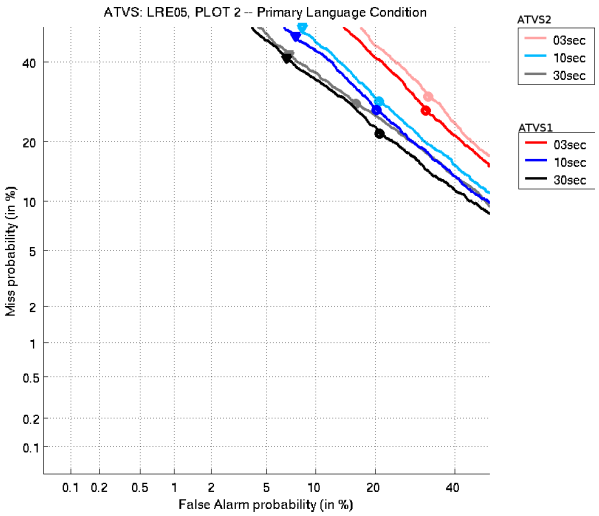

Figure 3: ATVS PPRLM submitted systems results at NIST 2005 LRE.

Figure 2 shows the results of the development process, with testing data coming from 2003 evaluation data, but restricted to the 2005 evaluation target languages. Figure 3 shows the results obtained in the 2005 evaluation. As aforementioned, in 1996 and 2003 editions, evaluation data was mainly extracted from the CallFriend database. But for the 2005 evaluation a new corpus was collected by OHSU. This corpus includes a large proportion of cellular data, which is not present neither in the CallFriend database nor in the NIST 2003 Evalu- ation data. The aforementioned systems were submitted to the 2005 evaluation, performing as shown 3. Those new channel conditions could explain the similar degradation in re- sults, by a factor of two or even more, that affected all LRE05 participants, from dev to eval data.

7. Conclusions

Several techniques are available in the field of speaker recognition. They are usually divided into acoustics (GMM, SVM) and high-level (phonotactic and prosodic) systems. Results obtained by each type of systems are quite different, and acoustics systems clearly outperform high-level systems. The main purpose of high-level systems is to be combined, in order to improve results. The kind of knowledge, parameters and modelling techniques these systems use are quite different from one to another (discriminative vs. generative approaches, phones, energy), thus providing a chance for fusions, which is performed at score level, providing important improvements in recognition rates.

8. References

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

[2] "Speaker recognition evaluations," http://www.nist.gov/speech/tests/spk/.

[3] G. Doddington, "Speaker recognition based on idiolectal differences between speakers," in *Proceedings of EUROSPEECH*, 2001, vol. 4, pp. 2517–2520.

[4] D. Reynolds et al., "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)*, Apr. 2003.

[5] W.M. Campbell, "Generalized linear discriminant sequence kernels for speaker recognition," in *Proceedings of the International Conference on Acoustics Speech and Signal Processing (ICASSP)*, 2002, pp. 161–164.

[6] 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)*, 2003, vol. IV, pp. 788–791.

[7] J. Gonzalez-Rodriguez et al, "On the use of high-level information for speaker recognition: the atvs-uam system at nist sre 05," *IEEE Aerospace and Electronic Systems Magazine*, (in press).

[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 Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2003, pp. 229–232.

[9] A. Moreno, D.Poch, A.Bonafonte, E.Lleida, J.Llisterri, and C.Nadeu J.Marino, "Albayzin speech database: Design of the phonetic corpus," *Proc. Eurospeech'93*, vol. vol. 1, pp. 175–178, 1993.

[10] Asmaa El Hannani, Doroteo T. Toledano, Dijana Petrovska-Delacrétaz, Alberto Montero-Asenjo, and Jean Hennebert, "Using data-driven and phonetic units for speaker verication," in *Proceedings of Odyssey06: The speaker and language recognition workshop*, 2006.

[11] W. M. Campbell, J. P. Campbell, D. A. Reynolds, D. A. Jones, and T. R. Leek, "High-level speaker verification with support vector machines," in *Proc. of ICASSP*, 2004, pp. 73–76.

[12] Jean-Luc Gauvain, Abdel Messaoudi, and Holger Schwenk, "Language Recognition Using Phone Lattices," Jeju Island, October 2004, pp. 1283–1286.

[13] Andrew O. Hatch, Barbara Peskin, and Andreas Stolcke, "Improved phonetic speaker recognition using lattice decoding," in *Proc. of ICASSP*, 2005, pp. 165–168.

[14] William M. Campbell, Douglas A. Reynolds, and Joseph P. Campbell, "Fusing discriminative and generative methods for speaker recognition: experiments on switchboard and NFI/TNO field data," in *Odyssey: The Speaker and Language Recognition Workshop*, 2004, pp. 41–44.

[15] D. A. Reynolds et al., "The 2004 MIT Lincoln Labs speaker recognition system," in *Proceedings of ICASSP*, 2005, pp. 177–180.

[16] Julian Fierrez-Aguilar et al., "Speaker verification using adapted user-dependent multilevel fusion," in *Proc. 6th IAPR Intl. Workshop on Multiple Classifier Systems, MCS, Springer LNCS-3541*, 2005, pp. 356–365.

[17] J. Kittler et al., "On combining classifiers," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 20, pp. 226–239, 1998.

[18] J. Fierrez-Aguilar et al., "Bayesian adaptation for user-dependent multimodal biometric authentication," *Pattern Recognition*, vol. 38, no. 8, August 2005.

[19] D. A. Reynolds, "Channel robust speaker verification via feature mapping," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2003, vol. 2, pp. 53–56.

[20] Roland Auckenthaler, Michael Carey, and Harvey Lloyd-Thomas, "Score normalization for text-independent speaker verification systems," *Digital Signal Processing*, vol. 10, pp. 42–54, 2000.

[21] D. Sturim and D. A. Reynolds, "Speaker adaptive cohort selection for tnorm in text-independent speaker verification," in *Proc. of ICASSP*, 2005.

[22] V. Wan and W. Campbell, "Support vector machines for speaker verification and identification," 2000.

[23] Steve Young et al., *The HTK Book*, Dec. 2002.

[24] ETSI ES 202 050 (v1.1.3), "Speech processing, transmission and quality aspects (STQ); Distributed speech recognition; Advanced front-end features extraction algorithm; Compression algorithms.," http://www.etsi.org/.

[25] "Sphinx systems homepage," http://cmusphinx.sourceforge.net/html/cmusphinx.php.

[26] Marc A. Zissman, "Comparison of Four Approaches to Automatic Language Identification of Telephone Speech," *IEEE Transactions on Speech and Audio Processing*, pp. 31–44, Jan. 1996.

[27] Pedro A. Torres-Carrasquillo et al., "Approaches to language identification using gaussian mixture models and shifted delta cepstral features," in *Proc. of ICSLP*, 2002.

[28] W. M. Campbell et al., "Language recognition with support vector machines," in *Proceedings of Odyssey04: The speaker and language recognition workshop*, 2004, pp. 285–288.

[29] OGI, "OGI Multilanguage Telephone Speech v1.2," http://cslu.cse.ogi.edu/corpora/mlts/.

[30] Matejka Pavel, Schwarz Petr, Cernocky Jan, and Chytil Pavel, "Phonotactic language identification using high quality phoneme recognition," in *Interspeech'2005 - Eurospeech - 9th European Conference on Speech Communication and Technology*, Lisbon, 2005, pp. 2237–2240.

[31] Alberto Montero-Asenjo, Doroteo T. Toledano, Javier Gonzalez-Dominguez, Joaquin Gonzalez-Rodriguez, and Javier Ortega-Garcia, "Exploring PPRLM performance for nist 2005 language recognition evaluation," in *Proceedings of Odyssey06: The speaker and language recognition workshop*, 2006.