Multilevel and Session Variability Compensated Language Recognition: ATVS-UAM Systems at NIST LRE 2009

Javier Gonzalez-Dominguez, Ignacio Lopez-Moreno, Javier Franco-Pedroso, Daniel Ramos, *Member, IEEE*, Doroteo T. Toledano, *Member, IEEE* Joaquin Gonzalez-Rodriguez, *Member, IEEE*

***Abstract*—This work presents the systems submitted by the ATVS Biometric Recognition Group to the 2009 Language Recognition Evaluation (LRE'09), organized by NIST. New challenges included in this LRE edition can be summarized by three main differences with respect to past evaluations. Firstly, the number of languages to be recognized expanded to 23 languages from 14 in 2007, and 7 in 2005. Secondly, the data variability has been increased by including telephone speech excerpts extracted from Voice of America (VOA) radio broadcasts through Internet in addition to Conversational Telephone Speech (CTS). The third difference was the volume of data, involving in this evaluation up to 2 terabytes of speech data for development, which is an order of magnitude greater than past evaluations. LRE'09 thus required participants to develop robust systems able not only to successfully face the session variability problem but also to do it with reasonable computational resources. ATVS participation consisted of state-of-the-art acoustic and high-level systems focussing on these issues. Furthermore, the problem of finding a proper combination and calibration of the information obtained at different levels of the speech signal was widely explored in this submission. In this work, two original contributions were developed. The first contribution was applying a session variability compensation scheme based on Factor Analysis (FA) within the statistics domain into a SVM-supervector (SVM-SV) approach. The second contribution was the employment of a novel backend based on anchor models in order to fuse individual systems prior to one-vs-all calibration via logistic regression. Results both in development and evaluation corpora show the robustness and excellent performance of the submitted systems, exemplified by our system ranked 2^{nd} in the 30 second open-set condition, with remarkably scarce computational resources.**

***Index Terms*—Language Recognition, Factor Analysis, Sufficient Statistics, Linear Scoring, Anchor Models, Calibration.**

I. Introduction

RECENTLY, Spoken Language Recognition (SLR) has experienced an increase in interest mainly due to its use in a wide range of applications such as audio indexing, information retrieval or call center monitoring. While interest in the field has been latent for nearly 40 years [1], it has not been up to the last decade when systems have experienced a major research development. Among the driving factors of this rapid development and performance improvement of state-of-the-art technologies, the efforts of the US National Institute of Standards and Technologies (NIST) deserve special mention [2]. The Language Recognition Evaluations (LRE), organized by NIST since 1996, with editions in years 1996, 2003, 2005, 2007 and 2009 have established a common framework for the development and assessment of language recognition technology, succesfully focusing the efforts of the scientific community in the field. This framework includes common protocols and databases for experimental evaluation as well as well-defined evaluation methodologies [2]. Currently, the LRE evaluation has become the major and reference forum for scientific researchers and technology developers in the area who aim at adapting their systems to real-world challenges. Following such objectives, the ATVS Biometric Recognition Group of the Universidad Autonoma de Madrid (hereafter, ATVS) has been participating in LRE's since 2005, submitting systems at both lower (spectral) and higher levels (phonotactic, prosodic) for blind and public competition. From the perspective of the scientific community, the problem of automatic SLR represents a very attractive task for several reasons. On the one hand, in order to yield good performance, different levels of information across the speech signal have to be exploited. This fact implies the use of efficient methods to combine complementary information extracted from the speech signal. This is one of the major challenges in the field and it is an underlying theme in this paper. Moreover, SLR systems share most of the problems with other related research areas such as speech and speaker recognition and therefore similar solutions can be ported across to each of these fields. A good example is the inter-session variability problem, understood as the set of acoustic differences between utterances, which are not related respectively to the speaker or language to recognize. In fact, this problem, caused by several variability sources (such as channel conditions or environmental noise), is still a major source of system performance degradations in all recognition disciplines involving speech signals [3]. Because of its configuration, the LRE'09 edition clearly focused on these challenges. Session variability is present in the task by including telephone speech from Voice Of America (VOA)[1], a vast multilanguage data source new to those evaluations in addition to well-known Conversational Telephone Speech (CTS). In addition to this, a larger number of languages (23) were included, involving more language pairs difficult to distinguish (e.g Dari-Farsi, Hindi-Urdu, Bosnian-Croatian). Moreover, a huge amount of data was available to develop the systems, which required to proccess a much larger quantity of trials with respect to other evaluations. This fact highlighted the importance of systems with an acceptable balance between recognition performance and computational resources. The aim of

[1]http://www.voanews.com/english/index.cfm

this article is to describe the systems submitted by ATVS to LRE'09, which were focused on these new challenges as well as to explain some original contributions which were incorporated. The ATVS submission consisted of four different combinations of acoustic and phonotactic subsystems. The two ATVS spectral (also known as acoustic) subsystems were based in session variability compensated first-order sufficient statistics via Factor Analysis (FA) [4][5][6][7]. These statistics were calculated in our primary acoustic system which is based on the FA-GMM linear scoring framework [4], also outlined in this work as being a critical part of our acoustic systems. A novel approach, using a SVM supervector [8] acoustic system feeded from session variability compensated first-order statistics is included. The phonotactic components were based on PhoneSVM [9] composed of seven ATVS tokenizers and three tokenizers made available by Brno University of Technology (BUT). System combination is performed in a front-end-back-end configurarion. The front-end consists of recognizers trained for different languages for each of the systems used in the submission. In particular, 22 recognizers trained with VOA speech and 14 CTS recognizers trained with CTS speech were used for each system. Each recognizer for each system yielded a score, and all scores together formed a vector. After that, a back-end stage was used for classifying the resulting vector for each target language. A contribution of our submission was the use of a novel Anchor-Model approach for back-end fusion, where score vectors were classified using an SVM. Front-end scores were channel-dependent (22 VOA/14 CTS) t-normalized [10] while back-end scores are channel-independent (23 VOA+CTS) t-normalized. Calibration was achieved by the use of linear, two-class logistic regression [11], where scores were transformed into two-class, one-vs.-all log-likelihood-ratios (log-LR). In this way, a score can be interpreted as a degree of support towards any of the relevant hypotheses in the recognition process, namely θ_0 (the language in the utterance is the target language) and θ_1 (the language in the utterance is not the target language) [12]. This also allows to use Bayes thresholds for decision making, which are independent of the distribution of the output scores. The same logLR sets were submitted to the closed- and open-set conditions of the evaluation.

This paper is organized as follows. First, the ATVS individual spectral and high-level systems are described in Sections II and III respectively. Section IV presents the fusion scheme and calibration carried out in order to obtain final submitted scores, while Section V details the experimental framework for both, development and evaluation assessment. Section VI presents the ATVS submitted systems and notes on implementation details. Achieved results are presented in Section VII. Finally, future work and conclusion are outlined in Section VIII.

II. ATVS Spectral Systems

A. *FA-GMM Linear Scoring System*

The ATVS Factor Analysis Linear Scoring GMM system (hereafter, FA-GMM-LS) is based on the work developed by Niko Brummer in [4]. This system establishes a robust and efficient generative GMM framework where data sufficient statistics, relative to an Universal Background Model (UBM), play a central role. Indeed, once these are computed, both features and UBM can been discarded for next steps, with the corresponding computational savings. The *linear* term refers to novel scoring approach based on a linear approximation to log-likelihood ratios via first-order Taylor series [13]. Thus, scoring procedure simplifies to a single vector dot product. Further, session variability compensation via Factor Analysis (FA) [14] [7] is applied directly at the statistics level in both train and test stages. This subsection gives an overview of this system in four steps, where foundations for the original contributions presented in II-B are established.

*1) **Sufficient statistics**:* Given a utterance, with set of features $O = \{o_1, o_2, ..., o_n\}$ in $\Re^D$, and a reference model $\lambda_{UBM} = \{w_k, \mu_k, \Sigma_k\}$, $k = 1, ...C$, zero and first-order Baum-Welch statistics, for gaussian k of λ_{UBM}, are defined as follows:

$$zero{-}order\ statistic \longrightarrow n_k = \sum_t P_{kt} \tag{1}$$

$$first{-}order\ statistic \longrightarrow x_k = \sum_t P_{kt} o_t \tag{2}$$

where *Gaussian Occupation Probability* P_{kt} is given by:

$$P_{kt} = P(k|o_t, \lambda_{UBM}) = \frac{w_k p_k(o_t)}{\sum_{j=1}^{C} w_j p_j(o_t)} \tag{3}$$

being:

$$p_k(x) = \frac{1}{(2\pi)^{\frac{D}{2}} \Sigma^{\frac{1}{2}}} \exp(-\frac{1}{2}(x-\mu_k)' \Sigma_k^{-1} (x-\mu_k)) \tag{4}$$

For convenience first-order statistics x_k use to be measured relative to the means of the model:

$$x_{norm,k} = \sum_t P_{kt}(o_t - \mu_k) \tag{5}$$

Hereafter, we refer as $\overline{x}$ to first-order statistics supervector built as the concatenation of all $x_{norm,k}$ and N as the $CDxCD$ diagonal matrix built as C blocks defined as $N_k = n_k I$, being I the DxD identity matrix.

*2) **Classical MAP**:* As in classical GMM-UBM framework [15], a GMM for each language is derived via Maximum a Posteriori Estimation (MAP) [16] from the UBM and available training data. However, here, only means are adapted and this is performed via a single MAP iteration. This shortcut besides the linear scoring approach allow to calculate only once sufficient statistics from the data and make independent the rest of the system with respect to the UBM.
In terms of sufficient statistics, MAP process to obtain a new means of a language model λ_L can be resumed as the following equation in matrix form:

$$\overline{\mu}'_L = \overline{\mu}_{UBM} + (\tau I + N)^{-1} \overline{x} \tag{6}$$

where τ is the relevance factor and N, $\overline{x}$ resumes available training data for language L. Note that second order statistics are not neccessary because variances are not adapted.

*3) **Session variability via Factor Analysis at statistic level**:* Session variability subspace adaptation in model domain can be also seen as a mean adaptation restricted to a subspace [5][7] in the form:

$$\overline{\mu}'_L = \overline{\mu}'_{UBM} + Uz \tag{7}$$

where U is a low rank matrix whose columns define the session variability subspace, and z are the *channel factors*. Given U and assuming that z is normal distributed $N(O, I)$, it can be shown that finding a point estimate of z which maximizes (7) can be done by solving:

$$z = A^{-1}b = A^{-1}U'\Sigma^{-1}\overline{x} \tag{8}$$

where:

$$A = I + U'\Sigma^{-1}NU \tag{9}$$

$$b = U'\Sigma^{-1}\overline{x} \tag{10}$$

(N.B. adapting only means, $\Sigma = \Sigma_{UBM}$)

However, it is desirable to apply the compensation in a stage before rather than in model domain as this would allow applying the compensation to test data without the need to create a model. In order to apply channel compensation directly in the statistics domain, the work in [6] where channel compensation is applied in the feature domain will serve as inspiration.
In [6], *channel* compensation is applied in every feature of an utterance i as follows:

$$\hat{o_t^{(i)}} = o_t^{(i)} - \sum_k P_{kt}U_k z \tag{11}$$

This idea can be reused in statistics domain in order to get a channel-compensated first-order statistic $\overline{y}$, in the following way:

$$\overline{y} = \overline{x} - Uz = \overline{x} - NUA^{-1}U'\Sigma^{-1}\overline{x} \tag{12}$$

This approach has the desirable property of avoiding the need of a computational expensive *frame by frame* compensation.

*4) **Classical scoring vs linear scoring**:* Classical GMM-UBM scoring of a dataset X and a target model λ_L is presented as a likelihood ratio as:

$$score_{X,\lambda_L} = \frac{P(X|\lambda_L)}{P(X|\lambda_{UBM})} \tag{13}$$

taking logarithms for practical issues this simplifies to:

$$score_{X,\lambda_L} = log(P(X|\lambda_L)) - log(P(X|\lambda_{UBM})) \tag{14}$$

Linear scoring proposes a linear approximation of $log(P(X|\lambda_L))$ based on its first-order Taylor's series expansion evaluated in $\overline{\mu}_{UBM}$:

$$\begin{aligned} \log P(X|\lambda_L) \quad &\simeq \quad \log P(X|\lambda_{UBM}) + \\ &+ \bigtriangledown_\mu \log P(X|\lambda_{UBM})'(\overline{\mu} - \overline{\mu}_{UBM}) \end{aligned} \tag{15}$$

Several advantages of this approach with respect to classical scoring, arise by carefully analizing the above Equation (15). First of all, the need to compute term $logP(X|\lambda_{UBM})$ is removed, being cancelled as easily shown substituting Equation (15) into Equation (14) as follows:

$$\begin{aligned} score_{X,\lambda_L} \quad &= \quad \log P(X|\lambda_{UBM}) + \bigtriangledown_\mu \log p(X|\lambda_{UBM})' \\ &\quad (\overline{\mu} - \overline{\mu}_{UBM}) - log(P(X|\lambda_{UBM})) \\ &= \quad \bigtriangledown_\mu \log p(X|\lambda_{UBM})'(\overline{\mu} - \overline{\mu}_{UBM}) \end{aligned} \tag{16}$$

Further, term $(\overline{\mu} - \overline{\mu}_{UBM})$ is just the offset in a classical MAP adaptation in which only a EM iteration is done. Taking advantage of this fact, target models can be expressed in FA-GMM-LS as the *offsets* in MAP adaptation, $m = (\tau I + N)^{-1}\overline{x}$ (see Equation (6)), since the need of using a UBM is removed from this step on.

Moreover, it can be shown that term $\bigtriangledown_\mu \log p(X|\lambda_{UBM})'$ is the first-order statistics $\overline{x}$ but normalized by the diagonal covariance matrix [17]. Thus, the scoring function is reduced to a dot product between the MAP offset model m and the first-order statistics calculated from X with respect to the UBM and normalized by the diagonal covariances matrix.

Summarizing the previous analysis, the score between a model λ_L generated from sufficient statistics N_{train} and $\overline{x}_{train}$ and a test dataset X represented by its first-order statistic $\overline{x}_{test}$ is defined by:

$$\begin{aligned} score_{X,\lambda_L} \quad &= \quad (\overline{\mu} - \overline{\mu}_{UBM}) \cdot (\Sigma^{-1}\overline{x}_{test}) \\ &= \quad (\tau I + N_{train})^{-1}\overline{x}_{train} \cdot (\Sigma^{-1}\overline{x}_{test}) \end{aligned} \tag{17}$$

Note that in order to apply session variability compensation in both train and test phases, first order statistics $\overline{x}_{train}$ and $\overline{x}_{test}$ must be replaced by compensated stats $\overline{y}_{train}$ and $\overline{y}_{test}$ following Equation 12.

B. SVM Working on Session Variability Compensated Supervectors

The ATVS SVM supervector (SVM-SV) system is based on the work proposed in [18] where a GMM mean supervector is considered a point in the high-dimensional transformed space where the SVM works. Each GMM mean supervector represents a mapping between an utterance and a high-dimensional vector and thus, the need for explicitly performing a mapping from a lower dimensional space as in GLDS approach [8] is avoided. Then, an hyperplane is estimated in this SVM subspace to discriminatively separate a target class from non-target classes.

A modification to the work in [18] was introduced into our system by employing a session variability compensation scheme within the statistics domain, by using the channel compensated first-order statistics from the FA-GMM-LS system. Then, a single MAP adaptation was applied in order to obtain compensated GMM supervectors.

Even though others channel compensated techniques applied to SVM have been proposed in the literature [19][6][20], as far as author's knowledge, none of them have been designed to work at this level, where its application implies some advantages. On one hand, although session variability compensation techniques applied to the feature domain such as feature Nuissance Attribute Projection (fNAP) [21] or

feature Latent Factor Analysis (fLFA) [6][21] have the prime advantage of allowing any type of posterior modeling, its application implies a frame-by-frame compensation over the set of features rather than a single compensation in model or statistics domain. This becomes a major drawback when large amounts of data must be processed, as in language recognition. On the other hand, once first-order statistics are channel compensated, no other FA techniques applied at model domain such as [20] or NAP [19] were necessary. This turned out in a major saving of computational time in our acoustic systems as well as a significant benefits in terms of recognition performance.

III. High Level Systems

Even though the ATVS submissions to recent LRE's have also included a prosodic system, in LRE'09 all our high-level systems were based on phonotactic systems. Among high-level systems, phonotactic systems are one of the most successful and classic approaches in the field of language recognition [22]. Phonotactic systems try to model the sequences of phonemes that are characteristic of a particular language by processing speech with a Phonetic Recognizer (PR) that transforms speech into a sequence of phonetic tokens. Systems can use a single PR or many different PRs in different languages (Parallel PR, or PPR) for better performance. The set of languages of the PRs does not need to meet with those to be recognized, which is highly desirable because otherwise it would be necessary to train a new PR for each new language to recognize.

The sequence of recognized phonetic tokens can be used in different ways for language recognition. The most classical approach is to use statistical Language Modelling (LM) techniques to model the frequencies of phones and phone sequences (n-grams) for each particular language. The combination of a single PR and LM gives the Phone Recognition Language Modelling (PRLM) approach [22]. The language model (LM_i) is previously trained on the phonetic sequences obtained by the PR from utterances known to be of language i. It is common to use also a Universal Background Model with a structure similar to the language models but trained on phonetic sequences obtained from many languages to represent the generality of all languages through a PR. Once these two models are available, the first step to verify the language of the utterance is to process it with the PR to produce the phonetic sequence, X. Then, the phonetic decoding of the test utterance, X, and the statistical models (LM_i, UBM) are used to compute the likelihoods of the phonetic decoding, X, given the language model LM_i and the background model UBM. The recognition score is the log of the ratio of both likelihoods, normalized by the number of phonemes in the phonetic sequence. Global scheme of this process is shown in Figure 1. As different PRs can be used for the same task, it is common to use a combination of several PRs and LM in an approach known as Parallel-PRLM (PPRLM) [22]. This approach dominated the field of language recognition for years and is still, with some evolutions and improvements, one key subsystem of state-of-the-art language recognition systems.

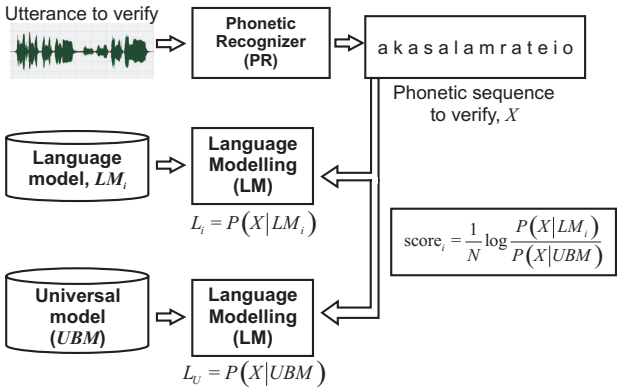

Fig. 1: *Verification process in PRLM scheme.*

One of the most important recent improvements in terms of performance is the use of SVMs for classifying the whole n-gram probability matrices [9], instead of using them in a likelihood ratio framework. This last type of system is usually referred to as Phone-SVM and is the type of system used in ATVS submission to LRE'09.

IV. Fusion and Calibration

As previously stated, a complete language recognition system is usually a combination of many individual subsystems. Combining this information by efficiently using the complementary information of every subsystem involved is known as fusion. The back-end/fusion strategy presented in this work and used in the LRE'09 evaluation is based on the use of an anchor models scheme [23].

Recently, the anchor models approach has been successfully used for both speaker verification and language identification [24][25] but not with the goal of fusion. The idea behind this approach is not only modelling the distribution of the scores for a target language with the scores for every utterance belonging to this language but to take advantage of the distribution of these scores against non-target models as well. By using anchor models, each utterance is mapped into a model space, called anchor model space, where the relative behaviour of the speech utterance with respect to other models can be learned. A point in this space is built by simply stacking scores obtained for testing an utterance over the cohort of pre-trained model as shown in Figure 2 (a). Once the set of stacked scores vectors are obtained for each language, these are used as inputs of a SVM system for discriminative purposes. Incorporating new subsystems to this fusion scheme is trivial as can be shown in Figure 2 (b).

In order to take the actual detection decision we have followed a per-language detection approach to calibrate the output log-likelihood-ratios (log-LR). Each score for each of the 23 target languages in the evaluation has been mapped to a logLR assuming a target-language-vs-rest configuration (one-vs-all). Therefore, each score can be interpreted as follows:

$$s_{cal} = \log(LR) = \log \frac{p(s|\theta_0)}{p(s|\theta_1)} \tag{18}$$

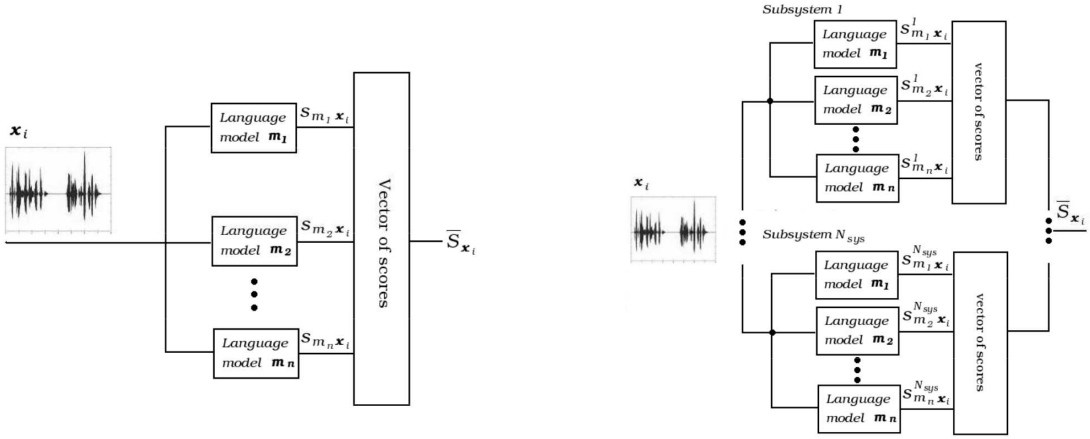

Fig. 2: *a) n-class parallel language detection problem where $\overline{S}_{x_i}$ stacks the similarities of x_i (input signal) over the set of models m_j, b) Generation of features (scores) in the anchor model space.*

where s_{cal} is the calibrated score, s is the score to be calibrated, and the hypotheses are defined as follows:

- θ_0: the language in the test utterance is the target language.
- θ_1: the language in the test utterance is not the target language.

Thus, a different score-to-log-LR mapping is performed per target language, and therefore the calibration strategy has been conducted independently for each target language. Linear logistic regression [11] has been trained, using the FoCal toolkit [2], on the complete development set of scores for each language.

After calibrating log-LR values, the logarithm of the Bayes threshold has been used in order to take decisions, defined as:

$$\log(\tau_B) = \log \frac{P_{nt} C_{fa}}{P_t C_{fr}} \tag{19}$$

where $P_{nt} = P_t = 0.5$ and $C_{fa} = C_{fr} = 1$ as defined by NIST, and therefore $\log(\tau_B) = 0$. If the calibration process is correctly performed, this is equivalent to choosing the minimum-cost threshold for each target language detection sub-system. Thus, after the log-LR transformation, both objective functions to optimize, namely C_{llrAvg} and C_{avg} as defined by NIST [26] tend to be as best as possible. However, a per-language one-vs-all calibration approach as this one will be slightly sub-optimal due to the fact that is does not take into account that this is actually a multiclass problem [27].

V. Databases, protocol and performance metric

LRE'09 evaluation included, for the first time, data coming from two very different audio sources. Besides CTS, used in past evaluations, telephone speech belonging to broadcast news was used for both train and test purposes. Broadcast data was obtained via an automatic acquisition system from "Voice of America" news (VOA) where telephone and non-telephone speech is mixed. Up to 2 terabytes of speech, automatically labeled in language and type, were distributed to participants. Further, around 80 audited segments for each target language (of aproximately 30 seconds duration each) was provided too for development purposes.

Both closed and open-set modes were defined as tasks in this evaluation each one tested with duration segments of 3, 10 and 30 seconds. We refer to closed-set as the task when only target languages are included in the test trials set, and to open-set when other non-target languages (unknown to participants) are also included. In this evaluation, 23 target languages were involved in closed-set as it was shown in Table I and 40 in open-set. More detailed information can be found in the LRE'09 evaluation plan [26].

In order to face this new challenge, where database mismatch play and important role [28], an ATVS development dataset was set up, ATVS-Dev09 onwards. This dataset was built to reproduce in the most accurately possible way, blind evaluation conditions by using different sets of CTS and VOA data provided by NIST. ATVS-Dev09 covered all target evaluation languages and test evaluation duration segments (3, 10 and 30 seconds). Table I shows the 23 evaluation target languages along with ATVS available data type per language. Specifically, the CTS training material (ATVS-DevTrain09) consisted of the "Callfriend" database, the full-conversations of LRE'05 and development data of LRE'07. For Russian data we used also "RuSTeN"[3]. Telephone broadcast data was obtained from speech segments (minimum length 30s.) extracted from VOA long files using telephone labels provided by NIST. The test material (ATVS-DevTest09) was obtained from the test part of LRE'07 (for target languages in both LRE'07 and LRE'09), and from manually labeled data from VOA provided by NIST. Finally, about 15,000 segments, balanced in segments of 3, 10 and 30 seconds, while LRE'09 evaluation included about 15,000 segments per duration (∼45,000 segments) and therefore about 1 million trials since every segment is tested against every target language.

In order to assess performance, two different metrics were

[2]Available at http://niko.brummer.googlepages.com/

[3]LDC 2006S34 ISBN 1-58563-388-7, www.ldc.upenn.edu

Language	Abbreviation	Data Type (VOA/CTS)
Amharic	$amha$	$VOA/-$
Arabic	$arab$	$-/CTS$
Bengali	$beng$	$-/CTS$
Bosnian	$bosn$	$VOA/-$
Chinese (Cantonese)	$cant$	$VOA/-$
Chinese (Mandarin)	$mand$	VOA/CTS
Creole	$creo$	$VOA/-$
Croatian	$croa$	$VOA/-$
Dari	$dari$	$VOA/-$
English (Indian)	$inen$	$-/-$
English (American)	$usen$	VOA/CTS
Farsi	$fars$	VOA/CTS
French	$fren$	$VOA/-$
Georgian	$geor$	$VOA/-$
German	$germ$	$-/CTS$
Hausa	$haus$	$VOA/-$
Hindi	$hind$	VOA/CTS
Japanese	$hind$	$-/CTS$
Korean	$kore$	VOA/CTS
Pashto	$pash$	$VOA/-$
Portuguese	$port$	$VOA/-$
Russian	$russ$	VOA/CTS
Spanish	$span$	VOA/CTS
Tamil	$tami$	$-/CTS$
Thai	$thai$	$-/CTS$
Turkish	$turk$	$VOA/-$
Ukranian	$ukra$	$VOA/-$
Urdu	$urdu$	$VOA/-$
Vietnamese	$viet$	VOA/CTS

TABLE I: *Alphabetical list of available languages. In bold, LRE'09 target languages.*

used, both evaluating the capabilities of one-vs.-all language detection. On the one hand, DET curves measure the discrimination capabilities of the system. On the other hand, C_{avg} which is a measure of the cost of taking bad decisions, and therefore it considers not only discrimination, but also the ability of setting optimal thresholds (i. e., calibration). In this work, while DET and C_{avg} results are shown, all our development process was based on C_{avg}, showing now also DET's just to visually observe the discrimination ability of the systems.

VI. Submitted Systems and Notes On Implementation Details

Different combinations of systems presented in Sections II and III were submitted leading to a total of four different systems built under different criteria:

- ***ATVS4*** is a phonotactic-only submission, fusion of the 10 PhoneSVM systems in use (seven from ATVS plus three from BUT)
- ***ATVS3*** is a fast and reliable acoustic-only submission with just the FA-GMM-LS system, designed to optimize the computational time but with a high level of recognition performance.
- ***ATVS2*** consisted of a fusion of all our acoustic (FA-GMM and SVM-SV) and phonotactic (PhoneSVM) systems, as shown in figure 3.
- ***ATVS1*** (primary) is a fusion of ATVS2 with primary system from other participant (TNO), where the latter consisted of a fusion of six acoustic systems: three GMM-SVM and three FA-GMM linear scoring as in [4].

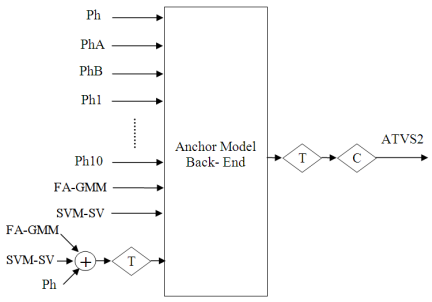

Fig. 3: *Fusion scheme for ATVS2 submitted system.*

A design decision was to generate language models for every target language in both VOA and CTS data where possible depending on data availability, using as well available data on other non-target languages. In that sense 14 CTS and 22 VOA front-end models were trained for every system (VOA IndianEnglish was trained only in the back-end due to data scarcity) as shown in table I. This was done with the goal of later fusing information provided for each model type. Figure 3 shows the fusion scheme for all our systems (ATVS2), remaining fusion systems following similar schemes.

Implementation details for each type of system as well as fusion and calibration notes are shown in the rest of this section.

A. Spectral Systems

A parameterization consisting of 7 MFCCC with CMN-Rasta-Warping [29] concatenated to 7-1-3-7 SDC-MFCCs was used [30] for spectral systems.

According to the data type, two UBMs namely UBM_{CTS} and UBM_{VOA} with 1024 gaussians were trained. Data from CallFriend, LRE'05 and train part of LRE'07 was used for training UBM_{CTS}, while the training of UBM_{VOA} was composed by VOA development data provided by NIST. Distribution per hours of this training is as follows. A total of 38.5 hours was used in UBM_{CTS} training, incluiding about 2.75 hours per 14 available languages. For UBM_{VOA} a total number of 31.2 hours balanced on 1.42 hour per 22 languages was used (IndianEnglish was not included due to data scarcity for this language).

Further, two different FA-GMM-LS systems were developed by using above UBMs. Two session variability subspaces matrices were trained from CTS and VOA data respectively, U_{CTS} and U_{VOA}. We found this approach to outperform the approach where mixed data (CTS,VOA) is processed to train a unique session variability subspace. In this work, session variability subspaces were trained via EM algorithm after a PCA initialization based on [31][7] and only top-50 eigenchannels were taken into account turns out in a $CDx50$ dimension matrix. In order to train the session variability subspaces, a large amount of data was used. U_{CTS} was trained with a total number of 350 hours by using 600 segments of about 150 seconds per the 14 languages available; while U_{VOA} was trained with 550 hours, using 600 segments of

about 150 seconds as well but of the 22 languages available. Data distribution for training UBMs and session variability subspaces is summarized in Table II.

Compensated statistics via Factor Analysis by using U_{CTS} and U_{VOA} as described in II-A3 were also used on the SVM-SV system.

B. High Level Systems

The phonotactic ATVS system is a fusion of 10 different Phone-SVM subsystems (Ph1 to Ph10) as described in Section III. Ph1 to Ph7 use phonetic tokenizers developed by ATVS and Ph8 to Ph10 use phonetic tokenizers trained with Hungarian, Czech and Russian data respectively [4]. The ATVS phonetic tokenizers are based on Hidden Markov Models (HMMs), trained with HTK [32] and later transformed to be used by the SPHINX [33] speech recognition engine for faster recognition. The phonetic HMMs are three-state left-to-right models with no skips, and the output pdf of each state is modeled as a weighted mixture of 20 Gaussians. The acoustic processing is based on 13 Mel Frequency Cepstral Coefficients (MFCCs) (including $C0$) and velocities and accelerations for a total of 39 components, computing a feature vector each 10 ms and performing Cepstral Mean Normalization (CMN). The languages of the phonetic decoders from Ph1 to Ph6 and the corresponding corpora used for training are English (with the corpus with ELDA catalogue number S0011), German (S0051), French (S0185), Arabic (S0183 + S0184), Basque (S0152) and Russian (S0099)[5]. Ph7 uses a phonetic decoder in Spanish trained on Albayzin spanish speech database [34] downsampled to 8 kHz, which contains about 4 hours of high-quality phonetically labelled speech. Once the speech segment has been transformed into a sequence of recognized phonetic tokens (with any of the phonetic decoders), this sequence is used to estimate count-based 1-grams, 2-grams and 3-grams, pruned with a probability threshold, resulting in about 40,000 n-grams. These are rearranged as a feature vector, which is taken as the input of an SVM that classifies the test segment as corresponding (or not) to one language. PhoneSVMs are combined in different ways to obtain different front-end systems. Each PhX system consists of 22 VOA and 14 CTS models trained separately. Channel dependent t-norm is the last stage of those phonotactic front-ends.

C. Fusion and calibration

Input vectors to our fusion systems anchor model based back-end had dimension 216 (36 ATVS models - 14CTS+22VOA- x 6 component systems) while primary was 438 adding scores output of other site. Back-end t-norm was design as channel-independent (VOA+CTS), while calibration was duration-dependent. Anchor model training was 90/10 bootstrapped while calibration training was bootstrapped with 80/20 using available training data. A channel independent T-Norm (models from VOA and CTS) stage was applied for scoring normalization.

LRE'09 considered three different nominal durations for the test segments: 3, 10 and 30 seconds of speech. The same individual subsystems were used to perform language recognition tests for the different durations. However, calibration has been trained specifically for the estimated different durations and an automatic voice activity detector has been used to classify test segments. As the calibration was applied after the back-end, a single score for each test segment was used, and scores from all the speech types (VOA, CTS) were pooled for training. Thus, all the available scores for each duration from each target language were used to train logistic regression, and the linear transformation obtained was used to calibrate the scores from testing data.

VII. Development and evaluation Results

The performance of ATVS submitted systems is summarized in Figure 4 for development (ATVSDev09) and evaluation (LRE'09) tests. Here, the discrimination per each system (ATVS1-4) and test segment duration (3, 10 and 30 seconds) is showed in a pooled DET curve. Several global observations can be inmediately extracted. Firstly, the good behaviour of the anchor models fusion scheme introduced is justified as being ATVS1 (fusion of systems) the system with lower error rates. The effect of test segment duration in system performance is also highlighted and it affects in a similar manner to both, acoustic and high level systems. Further, a slight degradation in the evaluation results with respect to development ones is showed. This degradation performance, common to all participants, is usually due to the database mismatch among the development and testing databases, and is a common effect in LRE's. Table III summarizes this information in terms of *meanCavg* (mean of Cavg per language) per system, evaluation dataset and test segment durations. It is also worth pointing out that acoustic systems outperform phonotactic ones except for short durations, and this with a much smaller computational complexity, but fusion of both kind of systems improve results, which encourages the use of multilevel approaches for language recognition.

In more detail, Figure 5 compares systems performance per target language. Again, results are presented on both, development and evaluation, but only for 30s test segment duration. Analysis shows the varying degrees of recognition difficulty among the different target languages (or better said, among the data available from those target languages). In the same way, Figure 6 presents in detail the effect of test segment duration per language for our primary system (ATVS1).

The need of proper session variability compensation is showed in Figure 7 where both spectral systems, FA-GMM-LS and SVM-SV are assesed with and without compensation via factor analysis on ATVSDev09. Results shows that channel compensation via FA is crucial in GMM modelling performance, getting an improvement of about 82% in $meanC_{avg}$ terms. Also, system SVM-SV take advantage of this compensation but to a lesser extent (4%). This effect appears

[4]These have been developed and made available for research purposes by the Speech Processing Group at Faculty of Information Technology, Brno University of Technology.

[5]www.elda.org.

Prior model	Databases	#Languages	#Hours/language	Total
UBM_{CTS}	$CallFriend, LRE05, TrainLRE07$	14	2.75	38.5
U_{CTS}	$CallFriend, LRE05, TrainLRE07$	14	25	350
UBM_{VOA}	VOA	22	1.42	31.2
U_{VOA}	VOA	14	25	550

TABLE II: *Distribution of data used for training Universal Background Models and Session Variability Subspaces.*

	ATVS Systems Performance					
	ATVS-Dev09			LRE'09		
	03s	10s	30s	03s	10s	30s
$ATVS1$	16.50	**6.48**	**1.56**	17.97	**7.87**	**3.71**
$ATVS2$	**16.17**	7.25	2.02	**17.92**	8.39	4.26
$ATVS3$	20.37	10.30	3.25	21.93	10.65	5.67
$ATVS4$	18.80	9.41	3.73	20.87	10.81	6.55

TABLE III: *ATVS submitted systems performance (meanCavg x 100) on development and evaluation datasets.*

	ATVS1 on LRE09		
	$03s$	$10s$	$30s$
$closed-set$	17.97	7.87	3.71
$open-set$	18.69	8.80	4.58

TABLE IV: *ATVS1 performance (meanCavg x 100) on LRE'09 closed- and open-set.*

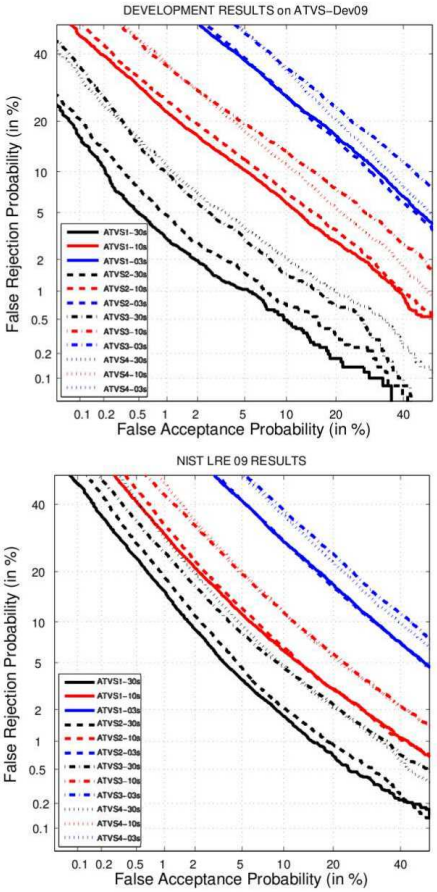

Fig. 4: *Pooled DETs per ATVS submitted systems on development (ATVS-Dev09) and evaluation (LRE'09) per all target test segment durations (3, 10 and 30 seconds)*

due to differences in SVM and GMM modelling. In GMM, target languages models, trained with huge amount of data, are far shifted with respect UBM reference model after even a single MAP adaptation. This mean shifting includes not only information belonging to the language but session variability found in the training database which it is mainly independent of the languages. This leads to models that are growing strongly affected by session variability effects. On the contrary, the SVM exhibits a higher robustness to this problem due to its ability to estimate an hyperplane separating target single utterances models against all non-target ones. However, once session variability compensation is applied, GMM outpeforms SVM-SV system.

Table VII presents the system performance of our primary system on the closed- and open-set where a total of 40 languages were involved (23 target + 17 non-target). Results for the core condition (closed-set, 30s) are comparable to the best systems in the evaluation. It is worth highlighting the excellent performance of the ATVS primary system in the open-set condition, where a second rank position was obtained. Results in that task prove the robustness of anchor models working under unseen languages.

VIII. Conclusion and Future Work

In this article we have described the ATVS-UAM submission to the 2009 NIST Language Recognition Evaluation. This submission was particularly successful since our systems achieved the 2^{nd} position in the open-set condition with speech segments of 30 seconds. The article has discussed and presented the state-of-the-art technologies used in our systems, with emphasis on the two main research innovations introduced. Firstly, anchor models based fusion has been proposed and has proven to be an excellent scheme for fusion of a set of different subsystems. Secondly, session variability compensation has been applied on statistics domain and has shown to outperform the SVM-SV system, thus avoiding the need for a frame by frame compensation and allowing statistics extracted from the linearized FA-GMM system to be reused. Besides these innovations, the LRE'09 task included several new research challenges with respect to former evaluations, as huge amount of data to process and a larger number of target languages (23). A special mention deserves the broad session variability due to the use of telephone data from two different sources, broadcast news (extracted from Voice of America news -VOA-) and conversational telephone speech (CTS). ATVS acoustic and high level systems were built taking into account all these factors and achieved good performance

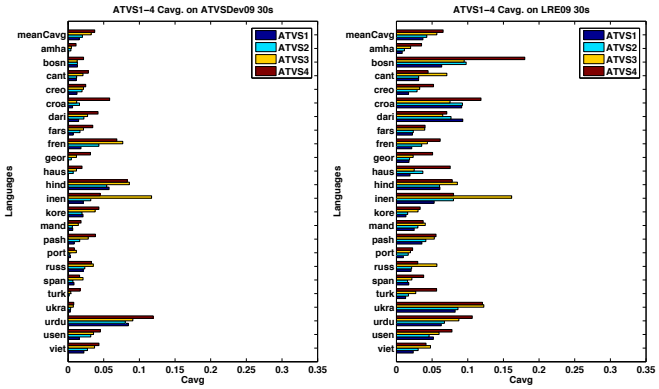

Fig. 5: *Comparision of ATVS submitted systems on both, development (ATVS-Dev09) and evaluation (LRE'09) datasets for 30 seconds test duration segments.*

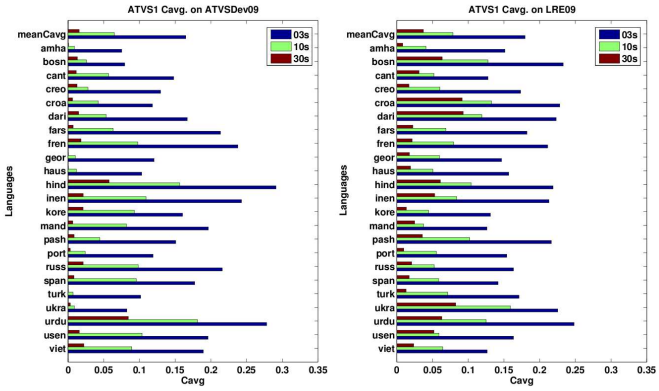

Fig. 6: *ATVS primary system performance on both, development (ATVS-Dev09) and evaluation (LRE'09) datasets (3, 10 and 30s).*

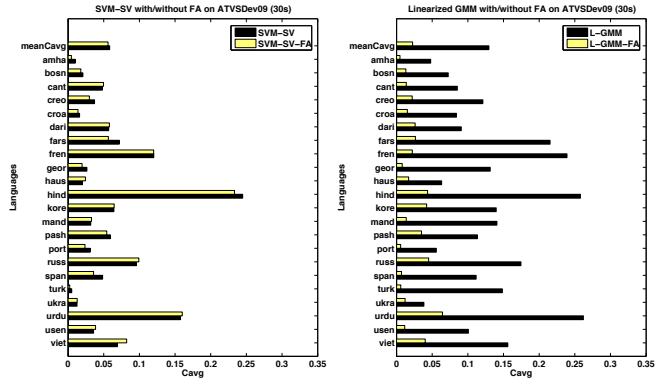

Fig. 7: *Effect of session variability compensation on SVM-SV and FA-GMM-LS systems. Results on ATVS-Dev09 using VOA models and* U_{VOA}.

in the task with remarkable results in all submitted tasks. To achieve this goal, the use of a powerful session variability compensation scheme via Factor Analysis have demonstrated to be crucial for acoustic systems performance, obtaining significant improvements in both the SVM-SV and the FA-GMM-LS models submitted. Future work includes several lines such as to explore new accurate ways to better extract and combine complementary information from different systems; to build systems more independent to the effects of test duration and to explore new techniques for fast adaptation to new channel conditions in session variability compensation when a limited set of unseen background data is available.

ACKNOWLEDGMENT

This work has been supported by the Spanish Ministry of Education under project TEC2006-13170-C02-01. Javier Gonzalez-Dominguez also thanks Spanish Ministry of Education for supporting his doctoral research under project TEC2006-13141-C03-03. Special thanks are given to Dr. David Van Leeuwen from TNO Human Factors (Utrech, The Netherlands) for his strong collaboration, valuable discussions and ideas. Also, authors thank to Dr. Patrick Lucey for his final support on (non-target) Australian English review of the manuscript.

REFERENCES

[1] K. Atkinson, "Language Identification from Nonsegmental Cues," *The Journal of the Acoustical Society of America*, vol. 44:378A, 1968.

[2] National Institute of Standards and Technology, "NIST LRE website," http://www.nist.gov/speech/tests/lang, (accessed 04 July 2008).

[3] F. Bimbot, J.-F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-Garcia, D. Petrovska-Delacretaz, and D. A. Reynolds, "A Tutorial on Text-Independent Speaker Verification," *Journal on Applied Signal Processing*, vol. 2004, no. 4, pp. 430–451, 2004.

[4] N. Brümmer, A. Strasheim, V. Hubeika, P. Matejka, L. Burget, and O. Glembek, "Discriminative Acoustic Language Recognition via Channel-Compensated GMM Statistics," in *Proc. of Interspeech*, 2009.

[5] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Factor Analysis Simplified," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 1, 2005, pp. 637–640.

[6] F. Castaldo, D. Colibro, E. Dalmasso, P. Laface, and C. Vair, "Compensation of Nuisance Factors for Speaker and Language Recognition," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 7, pp. 1969–1978, 2007.

[7] R. Vogt and S. Sridharan, "Explicit Modeling of Session Variability for Speaker Verification," *Computer Speech & Language*, vol. 22, no. 1, pp. 17–38, 2008.

[8] W. Campbell, J. Campbell, D. Reynolds, E. Singer, and P. Torres-Carrasquillo, "Support Vector Machines for Speaker and Language Recognition," *Computer Speech & Language*, vol. 20, no. 2-3, pp. 210–229, 2006.

[9] W. Campbell, J. Campbell, D. Reynolds, D. Jones, and T. Leek, "High-level Speaker Verification with Support Vector Machines," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2004, pp. 73–76.

[10] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, "Score Normalization for Text-Independent Speaker Verification Systems," *Digital Signal Processing*, vol. 10, no. 1/2/3, pp. 42–54, 2000.

[11] N. Brümmer, L. Burget, J. Cernocky, O. Glembek, F. Grezl, M. Karafiat, D. A. van Leeuwen, P. Matejka, P. Scwartz, and A. Strasheim, "Fusion of Heterogeneous Speaker Recognition Systems in the STBU Submission for the NIST Speaker Recognition Evaluation 2006," *IEEE Transactions on Audio, Speech and Signal Processing*, vol. 15, no. 7, pp. 2072–2084, 2007.

[12] J. Gonzalez-Rodriguez, P. Rose, D. Ramos, D. T. Toledano, and J. Ortega-Garcia, "Emulating DNA: Rigorous Quantification of Evidential Weight in Transparent and Testable Forensic Speaker Recognition," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 7, pp. 2072–2084, 2007.

[13] O. Glembek, L. Burget, N. Dehak, N. Brummer, and P. Kenny, "Comparison of Scoring Methods Used in Speaker Recognition with Joint Factor Analysis," in *ICASSP '09: Proc. of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. Washington, DC, USA: IEEE Computer Society, 2009, pp. 4057–4060.

[14] P. Kenny, G. Boulianne, P. Oullet, and P. Dumouchel, "Speaker and Session Variability in GMM-based Speaker Verification," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 15, no. 4, pp. 1448–1460, 2007.

[15] D. Reynolds, T. Quatieri, and R. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models," *Digital Signal Processing*, vol. 10, no. 1/2/3, pp. 19–41, 2000.

[16] J. Gauvain and C. Lee, "Maximum a Posteriori Estimation for Multivariate Gaussian mixture Observations of Markov chains," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 2, pp. 291–298, 1994.

[17] V. Wan and S. Renals, "Speaker Verification Using Sequence Discriminant Support Vector Machines," *Speech and Audio Processing, IEEE Transactions on*, vol. 13, no. 2, pp. 203–210, March 2005.

[18] W. M. Campbell, D. Sturim, and D. Reynolds, "Support Vector Machines Using a GMM Supervectors for Speaker Verification," *IEEE Signal Processing Letters*, vol. 13, no. 5, pp. 308–311, 2006.

[19] A. Solomonoff, W. Campbell, and I. Boardman, "Advances in Channel Compensation for SVM Speaker Recognition," in *ICASSP*, vol. I, 2005, pp. 629–632.

[20] D. Matrouf, N. Scheffer, B. Fauve, and J.-F. Bonastre, "A Straightforward and Efficient Implementation of the Factor Analysis Model for Speaker Verification," in *Proc. of Interspeech*, 2007, pp. 1242–1245.

[21] W. M. Campbell, D. Sturim, P. Torres-Carrasquillo, and D. Reynolds, "A Comparision of Subspace Feature-Domain Methods for Language Recognition," in *Proc. of Interspeech 2008*, September 2008.

[22] M. Zissman, "Comparison of Four Approaches to Automatic Language Identification of Telephone Speech," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 4, no. 1, pp. 31–44, 1996.

[23] I. Lopez-Moreno, D. Ramos, J. Gonzalez-Rodriguez, and D. T. Toledano, "Anchor-model Fusion for Language Recognition," in *Proc. of Interspeech 2008*, September 2008.

[24] M. Collet, Y. Mami, D. Charlet, and F. Bimbot, "Probabilistic Anchor Models for Speaker Verification," in *Proc. of Interspeech*, vol. 1, 2005, pp. 211–214.

[25] E. Noorl and H. Aronowitz, "Efficient Language Identification Using Anchor Models and Support Vector Machines," in *Speaker Odyssey*, 2006, pp. 1–6.

[26] "The 2009 NIST Language Recognition Evaluation Plan," http://www.itl.nist.gov/idad/mig/tests/lre/2009/LRE09_EvalPlan_v6.pdf (accessed 20 July), 2009.

[27] N. Brümmer and D. van Leeuwen, "On Calibration of Language Recognition Scores," in *Proc. of Odyssey*, San Juan, Puerto Rico, 2006.

[28] D. Ramos, J. Gonzalez-Rodriguez, and J. Gonzalez-Dominguez, J. Lucena, "Adressing Database Mismatch in Forensic Speaker Recognition with Ahumada III: a Public Real-Casework Database in Spanish," in *Proc. of Interspeech*, 2008.

[29] J. Pelecanos and S. Sridharan, "Feature Warping for Robust Speaker Verification," in *A Speaker Odyssey, The Speaker Recognition Workshop*, 2001, pp. 213–218.

[30] A. Rosenberg, C. Lee, and F. Soong, "Approaches to Language Identification Using Gaussian Mixture Models and Shifted Delta Cepstral Features," in *ICSLP*, vol. 1, 2002, pp. 89–92.

[31] P. Kenny, G. Boulianne, and P. Dumouchel, "Eigenvoice Modeling With Sparse Training Data," *IEEE Trans. on Speech and Audio Processing*, vol. 13, no. 3, pp. 345–354, 2005.

[32] "Hidden Markov Model Toolkit (HTK)," available on http://htk.eng.cam.ac.uk/.

[33] "Carnegie Mellon University SPHINX speech recognizer," available on http://sourceforge.net/projects/cmusphinx/.

[34] A. Moreno, D. Poch, A. Bonafonte, E. Lleida, J. Llisterri, J. Mario, and C. Nadeu, "ALBAYZN Speech Database: Design of the Phonetic Corpus," in *European Conference on Speech Communication and Technology, Eurospeech*, vol. 1, 1993, pp. 175–178.

Javier Gonzalez-Dominguez received his M.S. degree in Computer Science in 2005 from Universidad Autonoma de Madrid, Spain. In 2007 he obtained the postgraduate Master in Computer Science and Telecommunication Engineering from U.A.M. He has formed part of other international research groups and has participated in the development of the ATVS speaker and language recognition systems for several NIST speaker and language recognition evaluations, he has also participated on several related projects since 2005.

Ignacio Lopez-Moreno received his M.S. degree in Telecommunication Engineering in 2009 from Universidad Politecnica de Madrid (UPM). Currently he is a PhD student with the Biometric Recognition Group - ATVS -, where he is working as an assistant researcher since 2004. He has participated in several national projects and technology evaluations, and NIST speaker and language recognition since 2005. He has been recipient of several awards and distinctions, such as the IBM Research Best Student Paper in 2009.

Javier Franco-Pedroso Bachelor degree in Telecommunication Engineering (image and speech intensification) in 2005 from Universidad Politecnica de Madrid (UPM). Since september 2004 he is as an assistant researcher for the Biometric Recognition Group - ATVS - at Universidad Autonoma de Madrid. Currently he is an Electrical Engineering master degree student at UPM. His research interests include speaker verification, speaker tracking, pattern recognition and speech processing.

Daniel Ramos received his M.S in Telecommunication Engineering in 2001 from Universidad Politecnica de Madrid, Spain; and Ph.D. in Telecommunication Engineering in 2007 from Universidad Autonoma de Madrid, Spain. He is, since 2003, with ATVS - Biometric Recognition Group -, and since 2006 works as an Assistant Professor at Universidad Autonoma de Madrid. Dr. Ramos has participated in the development of the ATVS speaker and language recognition systems since 2004. He has been part of several organizing and scientific committees in the field, and has been recipient of several awards and distinctions, such as the IBM Research Best Student Paper Award at Odyssey 2006.

Dorote Torre Toledano M.S. Telecommunication Engineering degree from the Universidad Politecnica de Madrid, Spain, in 1997, obtaining the best academic records of his class. Ph.D. degree in telecommunication engineering from the same university, receiving a Ph.D. Dissertation Award from the Spanish Association of Telecommunication Engineers. He is with ATVS Biometric Recognition Group at the Universidad Autonoma de Madrid (Spain) where he is currently Associate Professor. He has served as a member of the scientific committee of several international conferences as well as a reviewer for several journals in the field.

Joaquin Gonzalez-Rodriguez , received the M.S. degree in 1994; and the Ph.D. degree "cum laude" in 1999, both in electrical engineering, from Univ. Politecnica de Madrid (UPM), Spain. Dr. Gonzalez-Rodriguez is co-director of the Biometric Recognition Group - ATVS -. After 15 years of research and lecturing at UPM, he is since May 2006 an Associate Professor at the Computer Science Department at Univ. Autonoma de Madrid, Spain. He has led ATVS participations in NIST Speaker and Language Recognition Evaluations since 2001. He is a member of ISCA and the Signal Processing Society of IEEE, and is also a member of the Program Committee of the ISCA Odyssey conferences on Speaker and Language Recognition.