Multilevel and channel-compensated language recognition: ATVS-UAM systems at NIST LRE 2009

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

ATVS Biometric Recognition Group, Universidad Autonoma de Madrid, Spain

{javier.gonzalez, ignacio.lopez, javier.franco, daniel.ramos, doroteo.torre, joaquin.gonzalez} @uam.es

Abstract

This paper presents the systems submitted by ATVS – Biometric Recognition Group at 2009 language recognition evaluation, organized by the National Institute of Standards and Technology of United States (NIST LRE'09). Apart from the huge size of the databases involved, two main factors turn the evaluation into a very difficult task. First, the number of languages to be recognized was the biggest of all NIST LRE campaigns (23 different target languages). Second, the database conditions were strongly variable, with telephone speech coming from both broadcast news, extracted from Voice Of America (VOA) broadcast system, and conversational telephone speech (CTS). ATVS participation consisted of state-of-the-art acoustic and high-level systems incorporating session variability compensation via Factor Analysis. Moreover, a novel back-end based on anchor models was used 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 for most of the languages (among them, Iberian languages such as Spanish and Portuguese)

1. Introduction

Language recognition has been an increasing research area in the last years, mainly due to its interest in applications such as audio segmentation and indexing or information retrieval. This interest is also motivated by the availability of the technology to yield acceptable performance, which has fostered the deployment of real-world applications. Among the driving factors of this rapid performance improvement of state-of-the-art technologies, the efforts of the American National Institute of Standards and Technologies (NIST) deserve special mention [1]. Due to the organization and funding of the Language Recognition Evaluations (LRE), the ability of the technology to successfully face challenging problems has achieved a remarkable increase. Moreover, NIST LRE have settled the foundations for the establishment of common protocols for experimental evaluation, from valuable and rich publicly available databases to well-defined evaluation methodologies. Therefore, it has become a highly valuable forum for scientific researchers and technology developers who aim at adapting their systems to real-world challenges.

Following such objectives, ATVS – Biometric Recognition Group of the Universidad Autonoma de Madrid (hereafter, ATVS) has been participating in NIST LRE since 2005, submitting systems at the spectral and higher levels for blind and public competition. The aim of this work is to describe the system presented by ATVS to NIST LRE in its 2009 edition, consisting 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. The first system was built according to the FA-GMM linear scoring framework [2] and the second one is a SVM whose inputs are model supervectors adapted from the first-order compensated sufficient statistics [3]. The phonotactic components are PhoneSVM composed of seven ATVS tokenizers and three tokenizers made available by Brno University of Technology (BUT). The systems work in a front-end-back-end configuration: first, dual models are obtained in the front end for VOA (22 models, indian-english was not trained in the front-end because of data scarcity) and CTS (14 models) data. Second, an anchor model back-end (23 VOA+CTS models, indian-english learned from other 22 model scores) was used for fusion. Front-end scores were channel-dependent (22 VOA/14 CTS) t-normalized while back-end scores are channel-independent (23 VOA+CTS) t-normalized. A calibration stage was finally used for transforming output scores into log-likelihood ratios (logLR) in order to allow the use of Bayes thresholds for decision making. The same logLR sets were submitted to the closed- and open-set conditions of the evaluation.

The development process prior to the blind submission was carried out by the construction of a corpus, which we have called ATVS-Dev09, using Callfriend, LRE'07 and VOA databases, and including the 23 languages of LRE'09.

The paper is organized as follows. First, ATVS individual spectral and high-level systems at the front-end are described in Section 2. Section 3 presents the fusion and calibration back-end. Finally, section 4 presents the results for ATVS-Dev09 set-up and also for the blind NIST LRE 09 evaluation dataset.

2. ATVS Systems

2.1. Spectral systems

2.1.1. DS-CS: FA-GMM linear scoring system

ATVS DS-CS (DotScoring with Compensated Statistics) GMM-FA linear scoring system is based on the work presented in [2]. In this work a complete acoustic system based on generative modelling GMM-FA framework is introduced, adding a new scoring approach based on a linear approximation to log-likelihood ratios. System shows a great

performance in both computational burden and recognition performance.

Feature extraction is shared among acoustic systems, consisting in 7 MFCCC with CMN-Rasta-Warping concatenated to 7-1-3-7 SDC-MFCCs. Given a UBM, zero and first order sufficient statistics are extracted for every utterance (train and test); then, first order statistics are session-variability compensated using FA, and models are generated from the compensated statistics. Finally, scores are obtained via dot product between test first compensated statistics and model supervector.

The model for session variability compensation is as follows:

$$m' = m + Ux$$

where the low rank U matrix defines the session variability subspace (U), and x represents the channel factors estimated from the training data used to build the model m. U was trained via EM algorithm after a PCA initialization based on [4][5]. Only top-50 eigenchannels were taken into account.

Two different GMM-FA linear scoring systems were developed according to the two different types of data presented in the evaluation. In that sense two UBMs and U matrices were trained from CTS and VOA data respectively. We found this approach to outperform the approach where mixed data (CTS, VOA) is processed to train a unique session variability subspace.

2.1.2. SV-CS: SVM chanel compensated supervector

ATVS supervector approach is also based on the statistics computed in 2.1.1 which are adapted from the UBM model (trained with the same data as 2.1.1 but having only 512 mixtures). Therefore, we obtain a single adapted statistic per utterance that summarises its information. Difference between the standard supervector, and statistics-based supervector is that in the latter we replace the vector of means of the adapted GMM by the utterance-adapted statistics.

2.2. High level systems

2.2.1. PhX: Phone-SVMs

Each of the seven different ATVS Phone-SVM subsystems (Ph1-Ph7) is based on the following steps. First a voice activity detector segments the test utterance into speech and non-speech segments. The speech segments are recognized with one open-loop phonetic decoder. The best decoding 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 per recognizer. 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 [3].

The process described above is repeated for the seven different open-loop phonetic recognizers used. In particular these subsystems use six phonetic decoders trained on SpeechDat-like corpora, each of which contain over 10 hours of training material covering hundreds of different speakers. The languages of these phonetic decoders and the corresponding corpora used are English (with the corpus with ELDA catalogue number S0011), German (S0051), French (S0185), Arabic (S0183 + S0184), Basque (S0152) and Russian (S0099) (www.elda.org). We have also included a 7th phonetic decoder in Spanish trained on Albayzin [6] downsampled to 8 kHz, which contains about 4 hours of speech for training. All these decoders are based on Hidden Markov Models (HMMs) trained using HTK and used for decoding with SPHINX. The phonetic HMMs are three-state left-to-right models with no skips, being the output pdf of each state modeled as a weighted mixture of 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 10ms and performing Cepstral Mean Normalization (CMN).

For each test utterance, the systems make n-grams with the 1-best solution produced by the phonetic decoders. Support Vector Machines (SVMs) take the n-grams as input vectors [3].

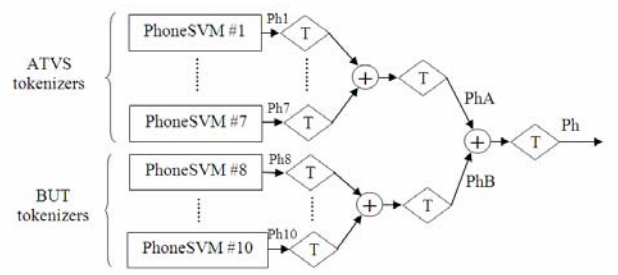

Fig. 1: *Hierarchical combination of phonotactic systems. T stands for t-norm, performed in a channel dependent way (VOA/CTS) in front-end systems.*

Additionally, three speech recognizers (Hungarian, Czech and Russian) from BUT (Speech@FIT, Speech Processing Group at Faculty of Information Technology, Brno University of Technology - FIT BUT, Czech Republic) have been used as additional high-quality tokenizers (Ph8-Ph10). The PhoneSVM systems are built then in the same way as with ATVS tokenizers. PhoneSVMs are combined in different ways to obtain different Front-end systems, as shown in figure 1. 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.

3. Fusion and calibration

Our back-end/fusion strategy was based on the use of anchor models [7], where high-dimensionality input vectors are classified in a single SVM per target model (23) both for VOA and CTS data. Recently, the anchor models approach has been successfully used for speaker verification and language identification too [8, 9, 7]. By using anchor models, each utterance is mapped into a model space where the relative behaviour of the speech utterance with respect to other models can be learned. The mapping function consists of testing every single utterance over a cohort of reference models, known as anchor models. The feature vector is the concatenation of all the scores. A channel independent T-Norm (models from VOA and CTS) stage was applied for scoring normalization.

In order to take the actual decision we followed a one-vs.-all detection approach to calibrate the output log-likelihood-ratios (logLR). Each score for each of the 23 target languages in the evaluation was mapped to a logLR assuming a target-language-vs.-rest configuration. Thus, a different score-to-logLR mapping was performed per target language. Linear logistic regression [10] was trained on the complete development set of scores for each language and for each given duration (3s, 10s and 30s) separately. The FoCal toolkit

has been used in order to train logistic regression (http://niko.brummer.googlepages.com/focal). After calibrating logLR values, the logarithm of the Bayes threshold has been used in order to take decisions.

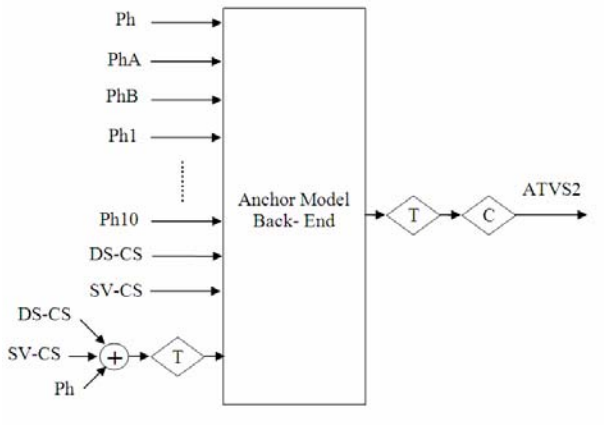

Fig. 2: *ATVS Fusion Scheme. T stands for t-norm, and C for calibration.*

Different combinations of systems presented in section 2 were submitted leading to a total of four different systems built under different criteria:

- **ATVS4** was a fusion of the 10 PhoneSVM systems used (7 from ATVS, 3 using BUT freely-available recognizers) and it evaluates the performance of our high-level technology.
- **ATVS3** only included the acoustic DS-CS system, which was designed to optimize the computational burden but with a high level of recognition performance.
- **ATVS2** consisted of a fusion of all our systems, as shown in figure 2. This system illustrates the performance reached by fusing ATVS systems.
- **ATVS1** (primary) consisted of a fusion of ATVS2 and the primary system of another participant in NIST LRE**.** This shows how our systems can take advantage of other different sources of language recognition information.

4. Development and evaluation Results

4.1. Databases, protocol and performace metric

A closed-set development dataset, known as ATVS-Dev09, composed of portions or all of LRE'05, Callfriend, LRE'07 and VOA data (different portions and/or selection criteria for train and test and for each language) was used to test the submitted systems in the 23 languages of LRE'09. We refer closed-set as the task where only target languages are included in the test stage, opposite to open-set where other non-target languages can be included. Detailed information can be found in the NIST evaluation plan [11].

The training material (ATVS-DevTrain09) for the CTS language models consisted of the Callfriend database, the full-conversations of NIST LRE 2005 and development data of NIST LRE 2007. For Russian data we used also RuSTeN (LDC 2006S34 ISBN 1-58563-388-7, www.ldc.upenn.edu). VOA models are obtained from speech segments (minimum length 30s.) extracted from VOA2 and VOA3 long files (except manually labeled files, used for testing) using telephone labels provided by NIST.

The test material (ATVSDevTest) was obtained from LRE07Test (for target languages in both LRE07 and LRE09), and from manually labeled data from VOA2 and VOA3. A total number of about 15000 segments (30s, 10s and 3s) were used. The evaluation included about 15000 segments per duration (~45000 segments) and therefore about 1 million trials are defined, because every utterance is faced against every language model (23 languages). Details about the protocol can be found in the NIST LRE'09 evaluation plan [11].

In order to assess performance, two different metrics are used in this paper, 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} is a measure of the cost of taking bad decisions, and therefore it considers not only discrimination, but the ability of setting optimal thresholds (i. e., calibration). In this paper we also show the C_{avg} value of our calibrated systems for the Bayes threshold. Details about NIST performance measures can be found in [11].

4.2. Development Results

Development results in ATVS-Dev09 for all durations (30s, 10s, 3s) and all submitted systems are presented in figure 3 while figure 4 shows the C_{avg} after calibration of the ATVS primary system.

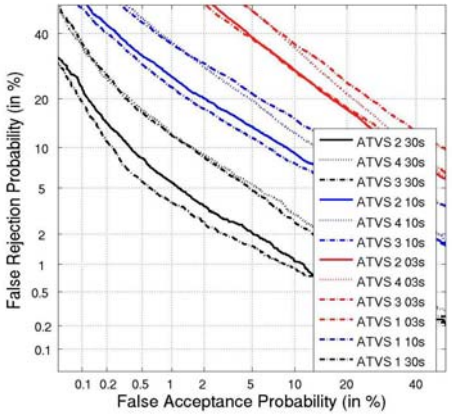

Fig. 3: *Pooled DETs (EERs in %) of submitted systems on ATVS-DevTest09.*

Results show the performance achieved for every submitted system. It is worth pointing out that acoustic systems outperform phonotactic ones, but fusion of both kind of systems improve results, which encourages the use of multilevel approaches for language recognition. Performance degradation due to duration of test segments is also showed.

The effect of using a session variability compensation scheme based on factor analysis is presented in figure 5. Here, the DS-CS system is evaluated on the ATVS-dev09 with and without session variability compensation A relative improvement on the EER of about of 56% is obtained when compensation is applied.

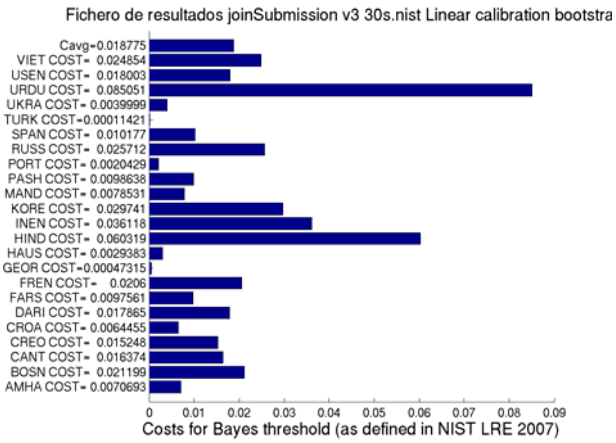

Fig. 4: *C_{avg} of ATVS1 on ATVS-DevTest09 set for 30s test segments.*

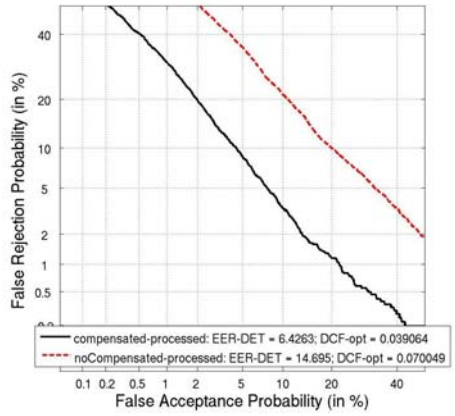

Fig. 5: *Pooled DETs (EERs in %) with acoustic dot-scoring system with and without FA channel compensation on ATVS-Dev09 prior to t-nomalization. 30s test segments.*

4.3. LRE09 Evaluation Results

Although a degradation of the performance of systems was observed in the evaluation with respect to development test, the behaviour of the systems in both experimental scenarios is consistent. This degradation performance, common to all participants, is due to the database mismatch among the development and testing databases, and is a common effect in NIST LRE. Moreover, the evaluation database exhibited a higher variability in terms of number of speakers.

Figures 6 and 7 show the ATVS primary system evaluation results for the closed and open set tasks respectively. Results for the core condition (closed-set, 30s) are comparable to the best systems in the evaluation. Moreover, 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.

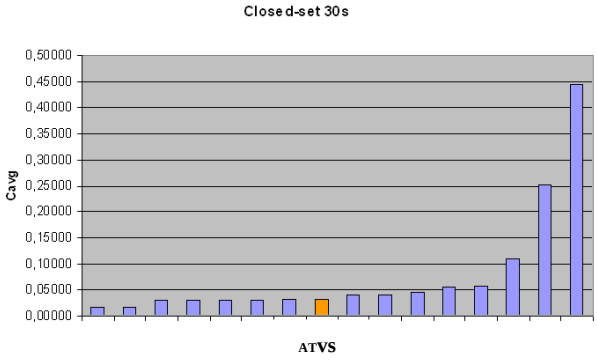

Fig. 6: *Official results on closed-set 30s task*

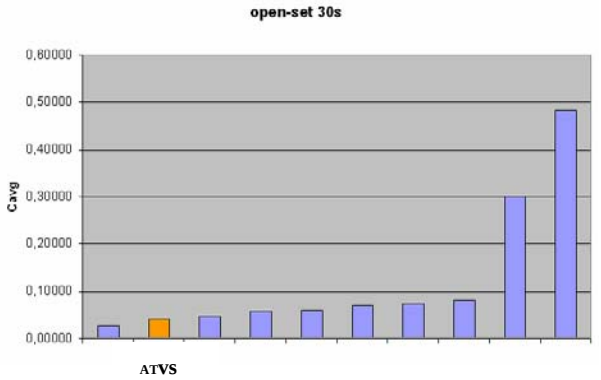

Fig. 7: *Official results on open-set 30s task*

5. References

[1] NIST LRE Website on http://www.nist.gov/speech/tests/lang/ (Accessed 04 June 2008).

[2] N.Brümmer A. Strasheim, et al. "Discriminative Acoustic Language Recognition via Channel Compensated GMM Statistics". Interspeech 2009, Brighton, U.K. Accepted.

[3] W. M. Campbell, J. P. Campbell et al. "Support vector machines for speaker and language recognition," Computer Speech and Language, vol. 20, no. 2-3, pp. 210–229, 2006.

[4] Kenny, P. and Boulianne, G. et al, "Eigenvoice Modeling With Sparse Training Data", IEEE Trans.~on Speech and Audio Processing, vol. 13, no. , pp 345-354.

[5] R. Vogt and S. Sridharan, "Explicit modelling of session variability for speaker verification," Computer Speech & Language, vol. 22, no. 1, pp. 17–38, 2008.

[6] A. Moreno, D. Poch et al, "ALBAYZÍN Speech Database: Design of the Phonetic Corpus," in poceedings of EUROSPEECH. Berlin, Germany, 21-23 September 1993. Vol. 1. pp. 175-178.

[7] I. Lopez-Moreno, D. Ramos et al. "Anchor-model fusion for language recognition", in *Proceedings of Interspeech 2008*, Brisbane, Australia, September 2008.

[8] M. Collet, Y. Mami et al. "Probabilistic Anchor Models Approach for Speaker Verification", in INTERSPEECH 2005.

[9] E. Noor1, H. Aronowitz "Efficient Language Identification using Anchor Models and Support Vector Machines", in Odyssey 2006 ISBN: 1-4244-0472-X pp 1-6.

[10] N.Brümmer, L. Burget et al. "Fusion of Hetereogeneous speaker recognition systems in the STBU submission for the NIST speaker recognition evaluation 2006} IEEE Transactions on Audio, Speech and Signal Processing, 2007. Vol 15. pp 2072-2084

[11] The 2009 NIST language recognition evaluation plan "www.itl.nist.gov/iad/mig/tests/lre/2009/LRE09_EvalPlan_v6.pdf."