The L^{2}F - UPC Speaker Recognition System for NIST SRE 2010

Alberto Abad[1]*, Jordi Luque*[13]*, Isabel Trancoso*[13] *and Javier Hernando*[2]

[1]L^{2}F - Spoken Language Systems Lab, INESC-ID Lisboa, Portugal
[2]TALP research center, Universitat Politècnica de Catalunya, Spain
[3]IST, Lisboa, Portugal

{alberto.abad,isabel.trancoso}@l2f.inesc-id.pt
{luque,javier}@tsc.upc.edu

Abstract

This document describes the joint submission of the INESC-ID's Spoken Language Systems Laboratory (L^{2}F) and the TALP Research Center from the Technical University of Catalonia (UPC) to the 2010 NIST Speaker Recognition evaluation. The L2F-UPC primary system is composed by the fusion of five individual sub-systems. Speaker recognition results have been submitted only for the *core-core* condition.

1. Introduction

The National Institute of Standards and Technology (NIST) has organized in the last years a series of evaluations in some relevant speech processing topics aimed at encouraging language research activities.

The 2010 NIST Speaker Recognition Evaluation (SRE2010) task consists of determining whether a specified speaker is speaking during a given segment of speech. The gender of speakers in train and test segments is known. Different common test conditions are defined depending on the characteristics of the training and test segments involved. Detailed information on the SRE10 campaign can be found in the evaluation plan document [1].

This report presents the speaker recognition (SR) system jointly developed by the INESC-ID's Spoken Language Systems Laboratory (L^{2}F) and the TALP Research Center from the Technical University of Catalonia (UPC) for the SRE10 campaign. The primary system is composed by the fusion of five individual SR sub-systems of very different characteristics. Two of the sub-systems are based on Joint Factor Analysis (JFA) with different speech features: (I) the JFA-spectral is based on Perceptual Linear Prediction (PLP) features with log-RelAtive SpecTrAl (log-RASTA) processing and (II) the JFA-prosodic uses prosodic features. Two additional sub-systems are based on Gaussian Supervectors (GSV) using also PLP features with log-RASTA processing: (III) the GSV-SVM is the standard supervector approach, combining Gaussian mixture models (GMM) with Support Vector Machines (SVM), and (IV) the GSV-GMM is the pushing-back version of the supervector approach. Finally, the (V) Transformation Network features with SVM modelling (TN-SVM) system is a new approach based on features obtained from the adaptation transforms applied to the Multi-Layer Perceptrons (MLP) that form a connectionist speech recognizer. The TN-SVM sub-system is the only one that makes use of the automatic transcripts provided by NIST.

In addition to the primary system, two alternative systems consisting of different system combinations have been submitted. The first alternative submission consists of the fusion of the two JFA sub-systems, that is JFA-spectral + JFA-prosodic (I+II). The second one is the combination all the sub-systems that do not depend on the automatic transcriptions provided by NIST, that is JFA-spectral + JFA-prosodic + GSV-SVM + GSV-GMM (I+II+III+IV).

2. Common characteristics

In this section some common characteristics shared by various sub-systems of the L^{2}F-UPC submission are described.

2.1. Development and training data

The data used for the development and the training of the systems comes from previous NIST evaluations. The NIST SRE 2004, 2005 and 2006 telephone data sets were used in this work for the systems training. Different subsets were selected for training the Universal Background Models (UBM), performing score normalization, modelling the background impostor set in SVM based sub-systems (III and V) or applying speaker/channel variability compensation techniques. The performance of the individual sub-systems and several other tested SR approaches was assessed in the NIST SRE 2008 *telephonic-telephonic* test sub-set. The SRE 2008 core test condition, the so called *short2-short3* task condition, with around one hundred thousand trials was used for system calibration and fusion of the final submission.

Notice that some of the tools used by the SR system and developed at the L^{2}F during the last years have been trained with additional data. For instance, the MLP speech-non-speech detector of next section has been trained mainly with down-sampled broadcast news (BN) data, augmented with music and sound effects data. The MLP acoustic models of the hybrid speech recognizer described in section 2.6 were trained on 140 hours of manually transcribed HUB-4 data.

2.2. Speech/non-speech segmentation (I,II,III,IV)

The output of the MLP speech-non-speech detector is combined with the alignment generated by a simple bi-Gaussian model of the log energy distribution computed for each speech segment to detect low-energy and highly likely non-speech frames. This speech/non-speech segmentation is used by sub-systems I through IV, particularly in the feature extraction process.

Segmentation of the interview segments was additionally post-processed. In order to obtain a better target speaker segmentation, the speech/non-speech segmentation of the interviewer (non-target) channel was obtained. Then, regions with

simultaneous speech activity in the interviewee and the interviewer channels were removed from the target speaker segmentation.

2.3. Spectral features (I,III,IV)

The spectral features used in sub-systems I, III and IV consist of 19 PLP features with log-RASTA processing and the frame energy, from a sliding window of 20 ms with a step size of 10 ms. First and second derivatives are concatenated to form 60 element feature vectors. Low-energy and highly likely non-speech frames are removed according to the speech segmentation previously described. Finally, mean and variance feature normalization is applied with mean and variance being computed independently for every speech utterance.

2.4. Prosodic features (II)

The prosodic features used in sub-system (II) are aimed at modelling the prosodic contours (both energy and pitch) of syllable-like regions [2]. We use the Snack toolkit [3] to extract the log-pitch and the log-energy of the voiced speech regions of every utterance. Log-energy is normalized on an utterance basis. The prosodic contours are segmented into regions by splitting the voiced regions wherever the energy signal reaches a local minimum (the minimum length of the regions is 60 ms). For each region, the log-energy and log-pitch contours are approximated with a Legendre polynomial of order 5, resulting in 6 coefficients for each contour. The final feature vector is formed by the two contour coefficients and the length of the syllable-like region, which results in a total of 13 elements.

2.5. GMM-UBM (I,II,III,IV)

Gender-dependent Universal Background Models (UBMs) were trained on NIST SRE 2004, 2005 and 2006 telephone data. The Audimus software package [4] and its utilities developed at the L^{2}F Laboratory were used for GMM modelling. A total of 72 hours from 870 male speakers and 100 hours from 1200 female speakers were used. The two gender-dependent UBMs were incrementally trained up to 1024 Gaussians, doubling the number of Gaussians at each iteration up to 25 iterations of the EM algorithm.

Two sets of UBMs were trained: one with the spectral features described in 2.3 and used for the development of sub-systems I, III, and IV; and the other trained with the prosodic features of 2.4 used for the development of sub-system II.

2.6. Automatic speech recognition (V)

Sub-system V uses a set of novel features extracted from adaptation techniques applied to the Multi Layer Perceptrons that form a connectionist speech recognizer.

2.6.1. The Audimus hybrid speech recognizer

The Audimus [4] ASR module uses MLP networks that act as phoneme classifiers for estimating the posterior probabilities of a single state Markov chain monophone model. The baseline system combines three MLP outputs trained with PLP features (13 static + first derivative), log-RASTA features (13 static + first derivative) and Modulation SpectroGram features (MSG, 28 static). When applied to narrow band recordings, the advanced Font-End from ETSI features (13 static + first and second derivatives) are also used. The number of context input frames is 13 for the PLP, RASTA and ETSI networks and 15 for the MSG network. The system adopted in this work models only monophone units, resulting in MLP networks of 40 softmax outputs for English.

2.6.2. Narrow-band acoustic models

The lack of conversational telephone speech (CTS) orthographically labelled data prevented us from developing an ASR system matched to the characteristics of the NIST Speaker Recognition Evaluation data sets. Consequently, a simple narrow-band speech recognizer with acoustic models trained with down-sampled BN data was used for this evaluations. The MLP acoustic models were trained on the same 140 hours of manually transcribed HUB-4 speech used for our American English BN transcription system [5].

2.6.3. Generation of phonetic alignments

Word-level automatic transcriptions provided by NIST were forced aligned using the narrowband acoustic networks to obtain phonetic alignments. Then, the alignments were used for training the speaker dependent transformation networks. Whenever the NIST transcriptions were not available, the narrow-band speech recognizer with the BN language model was used to generate a (weak) automatic transcription.

2.7. ZT-norm (I,II,IV)

Raw scores are ZT-normalized [6] in sub-systems I, II and IV. In the case of sub-systems III and V, which are based on SVM classifiers, a significant impact of score normalization strategies was not observed and hence these strategies were not applied in the submitted version.

Gender-dependent sets were defined for score normalization. We used 400 speech segments (200 male and 200 female) for modelling the impostor set of speakers and a total of 400 speech segments (200 male and 200 female) for modelling the impostor score distribution per each target speaker. Both sets were randomly selected from the SRE2004 and SRE2005 data. No care was taken to avoid overlapping with the data used for UBM training.

3. The L^{2}F-UPC SR sub-systems

The complete L^{2}F-UPC speaker recognition system is the result of the fusion of five speaker verification scores generated by 5 individual sub-systems. Particularities of the sub-systems are described next.

3.1. (I) The JFA-spectral system

Our JFA based submission consists of a Universal Background Model generation and JFA itself. The UBMs are the ones described in section 2.5. For JFA, the cookbook developed by Ondrej Glembek at Brno University of Technology [7] was used. This approach has become one of the most successful compensation techniques for speaker verification. Our JFA system closely follows the description of "Large Factor Analysis model" in paper [8]. The speaker model is represented by the mean supervector:

$$M = m + \mathbf{V}y + \mathbf{D}z + \mathbf{U}x \quad (1)$$

where m is the speaker independent mean supervector, V is a subspace with high speaker variability whose columns are referred to as eigenvoices, U is a subspace with high inter-session/channel variability whose columns are referred to as

eigenchannels, and D is a diagonal matrix describing remaining speaker variability not covered by V. Speaker factors y, z and channel factors x are assumed to be normally distributed random variables. This representation constrains all supervectors m to lie in an affine subspace which is spanned by the columns of V.

The UBMs were used to collect zero and first order statistics for training two gender-dependent JFA systems. The mean m and the variances of Gaussian components were set to the UBM mean and UBM variance respectively and not retrained in the training of JFA.

For speaker and channel modelling, 300 eigenvoices were trained on the NIST SRE 2004 and 2005 sets using speakers with at least 8 recordings or sessions (totalling 372 male and 519 female speakers). MAP estimates of speaker factors were obtained and they were fixed for the following training of eigenchannels. A set of 80 eigenchannels were trained on NIST SRE 2004 and 2005 telephone data (1806 recordings from 184 different male speakers and 2301 segments from 245 different female speakers). The diagonal matrix D in the JFA equation was estimated on all eigenvoices and eigenchannels. A set of NIST SRE 2006 speakers composed of 2384 and 3215 recordings of 298 male and 402 female speakers respectively is used for this purpose. MAP estimates of speaker and channel factors are fixed for estimating this diagonal matrix. The speaker factor y was jointly estimated with the channel factor x from the enrolment data. The common factor z was also estimated from training data.

In the testing stage, zero and first order statistics were extracted from the trial data. The channel's shift from UBM, i.e. the channel factor x, was estimated from the testing sentence, fixing it for all the speaker models, following the UBM point estimate assumption [9]. A linear scoring was performed to obtain the scores. Finally, factor analysis likelihood ratios were ZT-normalized, as described in section 2.7.

3.2. (II) The JFA-prosodic system

The JFA-prosodic system shares the same architecture of the previous JFA-spectral system, but relies on a complete different set of features for speech representation. Instead of the classical spectral coefficients, the prosodic features described in section 2.4 are used in this system. The data sets for UBM modelling, estimation of speaker parameters in equation 1 and for score normalization remain the same as in the JFA-spectral sub-system. ZT-norm score normalization is also applied.

3.3. (III) The GSV-SVM system

Combining Gaussian mixture models with Support Vector Machines [10], the so-called Gaussian supervector approach, is known to be a high performance speaker recognition approach.

For this evaluation, we have built a GSV system based on mean supervectors. First Gaussian Mixture Models for each target speaker are obtained with MAP adaptation of the Gaussian means of the UBM based on spectral features. UBM means are adapted with 20 MAP iterations with a relevance factor of 16 to obtain the speaker models.

The Gaussian Super Vector (GSV) system concatenates the mixture means of the MAP adapted Gaussian speaker models to obtain super vectors of every speech segment. The linear SVM kernel of [11] is used for training the speaker models with the libSVM tool [12]. The background set used as negative examples for SVM training is formed by 874 male, and 1204 female speech segments from the SRE2004, SRE2005 and SRE2006 1 side training corpora. The trained speaker SVM models are used for scoring the test supervectors.

Due to time constraints, we did not implement Nuisance Attribute Projection (NAP) for this sub-system, which is known to provide additional benefits.

3.4. (IV) The GSV-GMM system

The GSV-GMM sub-system is based on the GSV-SVM speaker recognition system of the previous section, but uses the alternative scoring approach of [13]. In contrast to the conventional GSV, each speaker SVM model is *pushed back* to a *positive* and a *negative* speaker GMM model, which are used in testing to calculate log-likelihood ratio scores. In certain situations, especially on short utterances, this approach provides improved performances. In this sub-system score normalization is applied. However, at the time of the submission all the necessary trials for performing the complete ZT-norm with 200 files per normalization and per gender were still not available. For that reason, Z-norm with only 100 Z-segments per gender was applied to the scores generated by the GSV-GMM subsystem.

3.5. The TN-SVM-NAP system

The Transformation Network features with SVM modelling system is a novel approach [14] that makes use of adaptation transforms employed in speech recognition as features for speaker recognition. However, in contrast to [15], the automatic speech recognizer that we rely on for computing the "differences" between the speaker independent and the speaker dependent model is the connectionist hybrid artificial neural network/hidden Markov model (ANN/HMM) system described in 2.6. Our approach uses a method known as Transformation Network [16] to train a linear input network that maps the speaker-dependent input vectors to the speaker independent system, while keeping all the other parameters of the neural network fixed.

The necessary phonetic alignments for network adaptation are obtained as described in section 2.6.3. For each MLP network that composes the acoustic models described in 2.6.2 the TN adaptation method is applied and a set of adaptation weights is obtained. A single TN feature vector of total size 3895 is formed with the linear transformation weights of the four MLP networks, and with the mean and variance statistics of the features data.

Additionally, nuisance attribute projection is applied to the TN features. Gender-dependent NAP projections were trained with the multisession conversational telephone speech training sets of SRE2004, SRE2005 and SRE2006 (7195 recordings from 921 different female speakers and 5226 recordings from 670 male speakers). We used a nuisance space of dimension 32.

The resulting TN features with NAP are used for training SVM speaker models. Gender-dependent negative examples for SVM training are obtained from the 1 side conversation training corpus of SRE2004, SRE2005 and SRE2006. In total, 867 and 1201 male and female segments are used for the background. Score normalization was not applied to the TN-SVM-NAP system. Additional implementation details can be found in [14].

4. Calibration and fusion

4.1. About data used

The SRE2008 *short2-short3* evaluation condition data set has been used for adjusting calibration and fusion of the sub-systems that compose the L^2F-UPC submission. Unfortunately, this set is known to be small and not adequate to the particularities of the new cost function considered in SRE2010. We are quite confident that we can improve the quality of our calibration and fusion stage using a larger number of trials.

4.2. Linear Logistic Regression with FoCal

Linear logistic regression tools provided by the FoCal Toolkit [17] have been used for both calibration and fusion. In a first stage, each sub-system was independently calibrated. In some cases, some of the sub-systems were not able to produce a score for a concrete trial. In that case, a score of 0 was given to the trial for that sub-system after the first calibration stage. Then, with all the scores of the five sub-systems, a second linear logistic regression was trained to obtain the final scores. The decision threshold was set in accordance to the new SRE2010 cost function.

4.3. Configurations

Three different calibration and fusion configurations were trained depending on the characteristics of the training and testing segments involved in a given trial. The "mic-mic" configuration was trained with the *interview-interview* subset of the *short2-short3* data set. The "mic-tel" configuration was obtained with the *interview-phonecall/telephone* trials. The *phonecall-phonecall/telephone* trials were used for estimating the calibration and fusion weights of "tel-tel" configuration.

In testing, the "tel-tel" configuration was used for the trials with both the training segment and the test segment identified as *phonecall telephone* data segments. The "mic-tel" calibration and fusion is used for trials that involve speaker models trained with *interview* data (both *3min* and *8min*) and test segments with *phonecall/telephone* data. Finally, the "mic-mic" configuration is used for the rest of the test trials: trials with *interview* data segments in both training and testing (independently of their length), trials with models trained with *interview* data and tested with *phonecall/microphone*, and trials with both *phonecall/microphone* data in train and test.

5. Summary and conclusions

The speaker recognition teams of L^2F (from Lisbon, Portugal) and UPC (from Barcelona, Spain) have presented a joint primary submission at the core condition of the NIST SRE 2010 campaign, consisting of the fusion of five different sub-systems. Additionally, two different combinations of the sub-systems that form the primary system have been presented as alternative contrastive systems. Time constraints made it impossible for us to submit results for the other evaluation conditions. We expect to evaluate our primary system in some of the alternative conditions as part of our post-evaluation work.

Our main objective in participating in this evaluation was to introduce ourselves to the speaker recognition community, to explore the recently proposed methods and to learn as much as possible. In this sense, independently of the final results, our participation was already quite successful. Additionally, the collaboration between two research groups from different countries was a nice achievement and we hope that can produce future fruitful collaborations.

Since it is our first participation at NIST SRE, most of our work during the last months was focused on the development and assessment of SR algorithms and methods. As a consequence, we could not devote enough attention to the new challenges proposed in this year campaign. For instance, no special attention was given to "low vocal effort" challenge or to the problems introduced as a consequence of the new cost function.

One important limitation of the submitted system is that cross-channel problems have been little or not studied during the development. Most of the data used for development is telephonic (background, UBM, eigenchannels, eigenvoices, NAP...). In fact, in most cases SR experiments during the development of the sub-systems were performed only in the *tel-tel* condition of SRE2008. Thus, we can expect a considerable better performance in the *tel-tel* condition compared to the other evaluation conditions.

Some sub-systems could have been significantly improved. In fact, some modules were removed at the very last minute due to time problems and implementation difficulties. For instance, NAP was not applied to sub-system III, although it was in our initial plans. Neither were we able to submit zt-norm scores of sub-system IV, having just applied z-norm. We also believe that significant improvements could be potentially obtained just by selecting a better calibration and fusion development set. We are confident that we will be able to improve the performance of our primary submission in the post-evaluation experiments, taking care of some of the problems that were just commented.

6. Acknowledgments

The authors would like to thank to David Matos and Tiago Luís for their support with the processing machines and data management issues. We also would like to thank to the organizers of the 2010 NIST speaker recognition evaluation for their availability for solving our doubts and problems. This work was partially funded by the European project I-DASH and by the Spanish project SAPIRE (TEC2007-65470).

7. References

[1] "The 2010 NIST Speaker Recognition Evaluation Plan (SRE10)", URL: http://www.itl.nist.gov/iad/mig//tests/sre/2010/

[2] L. Ferrer, N. Scheffer, E. Shriberg, "A Comparison Of Approaches For Modeling Prosodic Features In Speaker Recognition", In Proc. of ICASSP 2010.

[3] Snack Toolkit v2.2.10, KTH Royal Institue of Technology, Department of Speech, Music and Hearing. http://www.speech.kth.se/snack/.

[4] Meinedo, H., Caseiro, D., Neto, J. and Trancoso, I., "Audimus.media: a broadcast news speech recognition system for the European Portuguese language", in Proc. PROPOR 2003, Faro, Portugal, 2003.

[5] Pellegrini, T. and Trancoso, I., "Error detection in automatic transcriptions using Hidden Markov Models", In Proc. of Language and Technology Conference, 2009.

[6] Zheng, R., Zhang, S., and Xu, B., "A Comparative Study of Feature and Score Normalization for Speaker Verification", Lecture Notes in Computer Science, vol. 2832/2005, pp. 531-538, Springer Berlin/Heidelberg, 2005.

[7] Joint Factor Analysis Matlab Demo, Speech Processing Group, Brno University of Technology, Faculty of Information Technology http://speech.fit.vutbr.cz/en/software/joint-factor-analysis-matlab-demo

[8] P. Kenny, “Joint Factor Analysis of Speaker and Session Variability: Theory and Algorithms”, technical report CRIM-06/08-13, CRIM, 2005

[9] O. Glembek, L. Burget, N. Dehak, N. Brmmer, P. Kenny , “Comparison of scoring methods used in speaker recognition with joint factor analysis”, in Proceedings of the International Conference on Acoustics, Speech and Signal Processing, ICASSP09, 2009, pp. 4057-4060

[10] Campbell, W. M., Campbell, J. R., Reynolds, D. A., Singer, E. and Torres-Carrasquillo, P. A. “Support vector machines for speaker and language recognition”, Computer Speech and Language, vol. 20, pp. 210-229, 2006.

[11] Campbell, W. M., Sturim, D. E. and Reynolds, D. A., “Support vector machines using GMM supervectors for speaker verification” IEEE Signal Processing Letters, vol. 13(5), pp. 308-311, 2006.

[12] Chang, C.-C. and Lin, C-J, “LIBSVM - A Library for Support Vector Machines”, URL: http://www.csie.ntu.edu.tw/ cjlin/libsvm/index.html.

[13] Campbell,W.M, ”A covariance kernel for SVM language recognition,“ In Proc. of ICASSP 2008]

[14] Abad, A. and Luque, J., “Connectionist Transformation Network Features for Speaker Recognition”, in Proc. of Odyssey The Speaker and Language Recognition Workshop 2010, Brno, 2010.

[15] Stolcke, A., Ferrer, L., Kajarekar, S., Shriberg, E. and Venkataraman, A., “MLLR transforms as features in speaker recognition”, in Proc. Eurospeech 2005, pp. 2425-2428, Lisbon, 2005.

[16] Abrash, V., Franco, H., Sankar, A. and Cohen, M., “Connectionist Speaker Normalization and Adaptation”, in Proc. of Eurospeech 1995, pp. 2183-2186, Madrid, 1995.

[17] Brummer, N., “Tools for Fusion and Calibration of automatic speaker detection systems”, URL: http://www.dsp.sun.ac.za/ nbrummer/focal/.