ATVS-UAM System Description
NIST SRE 2006

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

Escuela Politecnica Superior
Universidad Autónoma de Madrid
c/ Tomas y Valiente 11
28049 Madrid, Spain
email: joaquin.gonzalez@uam.es

1. Introduction

ATVS-UAM submission to NIST SRE 2006 is a combination of both acoustic and higher-level systems ranging from GMM and SVM to Phone-trigram and Prosodic systems with SVM system combination and Pair Adjacent Violators (PAV) based Log-LR computation and calibration.

2. Overview of ATVS-UAM submission

		Test			
		10sec 2-channel	1 conv 2-channel	1 conv summed	1 conv aux mic
Train	10sec 2-channel				
	1 conv 2-channel	**GMM + SVM-Fw**	**LLR (SVMC (KL-GMM+SVM-Fw+SVM-Bckw))**		
	3 conv 2-channel				
	8 conv 2-channel		**LLR (SVMC (KL-GMM+SVM-Fw+SVM-Bckw + Prosodic + PhoneTrigram))**		
	3 conv summed				

KL-GMM: Feature Mapped UBM-MAP-GMM with KL-Tnorm
SVMC: SVM Classifier
SVM: GLDS-SVM with third order polynomial expansion (Fw: forward, Bckw: backward)
Prosodic: Four-Level Delta-based tokenization and interpolated (0.8Spk+0.2UBM) trigram target models with Tnorm
PhoneTrigram: Spk models interpolated (0.7UBPM+0.3Spk) from UBPM (Universal Background Phone Model) trigram language model with Tnorm
LLR: Transformation from log-likelihood scores to log-likelihood ratios via PAV-based monotonic transformation

3. KL-GMM-MAP-UBM System

The basic system is a likelihood ratio detector with target and alternative probability distributions modelled by Gaussian mixture models (GMMs) [1]. A Universal background GMM model is used as the alternative hypothesis model, and target models are derived using MAP adaptation from UBM.

3.1 Feature extraction and signal processing:

- 20 ms. window length, 10 ms. overlapped, Hamming .
- 20 mel-spaced (0-4000 kHz) magnitude filters.
- 38 coefficients per frame (19 MFCC + delta).
- Bandlimiting from 300 to 3300 Hz.
- CMN and Rasta filter. Feature Mapping [10] is performed for channel compensation, followed by a cepstral mean and variance normalization.
- A static, energy-based VAD was used for silence removal.

3.2 UBM:

A root UBM was trained using 5 hours of channel- and gender-balanced speech after silence removal. We used data from MIXER (NIST SRE 2004 and 2005), Switchboard I and Switchboard II. We trained the UBM using 1024 gaussian mixtures and ML estimation via EM algorithm.

3.2 Channel models for Feature Mapping:

14 channel models (7 per gender) were adapted from the UBM in order to perform Feature Mapping. An average value of 2 hours of speech was used for each channel model training.

3.3 Target model:

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

3.4 Likelihood normalization:

a.- Tnorm [2] (gender dependent). Used in all 10sec4w testing conditions. The generic Tnorm cohort consists of the total of target models from NIST SRE 2004 Evaluation sets.

b.- KL-Tnorm (gender dependent). An adaptive cohort selection algorithm for Tnorm [3] based on a fast estimation of Kullback-Leibler divergence for GMMs [4] was used for all 1conv4w testing conditions. The selected cohort consists of 60 impostor models selected from the generic, gender dependent Tnorm cohort. In all cases, Tnorm cohort and target model training speech length conditions and gender are matched. Tnorm cohorts in each training condition and gender consist of the total of target models from NIST SRE 2004 Evaluation sets.

4. GLDS-SVM System

The forward acoustic SVM system uses a explicit normalized three degree polynomial expansion [5] followed by a decomposed Generalized Linear Discriminant Sequence Kernel (GLDS) as described in [6]. SVMTorch [13] was used to train the target models.
Reverse acoustic SVM system is used in the 1conv4w-1conv4w task with the same configuration as for the forward system, but training the target models using the test segments. We get the reverse system scores testing those models with the training utterances.

4.1 Feature extraction and signal processing: the acoustic GLDS-SVM system uses the same feature extraction module as the acoustic GMM system.
4.2 Background model: two gender dependent and channel independent datasets have been used. These datasets consist of over 5 hours of channel-balanced speech extracted from the NIST SRE 2004, Switchboard I and Switchboard II databases.
4.3 Target model: SVMTorch [13] was used to train a linear SVM in the new expanded space. Scores are generated as a dot product in that space.
4.4 Score normalization: gender and training condition dependent Tnorm [2] cohorts were used. These cohorts were built using the target models from NIST SRE 2004.

5. Prosodic System

In order to capture prosodic differences among speakers in the realization of intonation, rhythm, and stress, the F0 and energy contours are converted into a sequence of tokens by means of a slope quantification process [7]. After that, the obtained sequences of tokens are modeled with n-grams to build speaker models that will be used, along with a UBM, to classify the distinctive token patterns using a LR test.

5.1. Four-Level Delta-based token computation

The token computation process converts the input speech signal into a sequence of tokens comprising prosodic information. This process is applied to all the speech utterances (defined as the period of time when one speaker is speaking and there is no silent gap for more than 0.5 seconds) within the input speech file. Once the F0 and energy contours have been computed, each utterance is segmented at inflection points of the temporal trajectories or at the start or end of voicing. In order to do that, first the inflection points for each trajectory at the zero-crossings of the derivative (using a +-2 frame span) are detected. Second, each utterance is segmented using the inflection points from both time contours and the start and end of voicing. Finally, each segment is converted into a set of tokens that describes the joint-dynamics of both temporal trajectories. This conversion is carried out using a four level quantization of the slopes of the F0 and energy contours (fast-rising, slow-rising, fast-falling, slow-falling). Since errors in the F0 and energy estimation are likely to generate small segments, all segments smaller than 30 ms are removed from the sequence of joint-state classes. The same token is assigned to all the unvoiced regions of each utterance since no F0 information is available for quantization. To avoid the modeling of classes across utterances, the token <s> is introduced between utterances. Next table shows all possible tokens used to describe the speech utterances.

TOKEN	DESCRIPTION	
	FO	E
1	Fast-rising	Fast-rising
2	Fast-rising	Slow-rising
3	Slow-rising	Fast-rising
4	Slow-rising	Slow-rising
5	Fast-falling	Fast-falling
6	Fast-falling	Slow-falling
7	Slow-falling	Fast-falling
8	Slow-falling	Slow-falling
9	Fast-rising	Fast-falling
10	Fast-rising	Slow-falling
11	Slow-rising	Fast-falling
12	Slow-rising	Slow-falling
13	Fast-falling	Fast-rising

14	Fast-falling	Slow-rising
15	Slow-falling	Fast-rising
16	Slow-falling	Slow-rising
17	UV	---
<S>	---	---

5.2. Modeling

HTK 3.2.1 n-gram modeling tools are used to train gender-dependent UBMs and the target-speaker models. Male and female training data from NIST 2005, NIST 2004, SWITCHBOARD I and SWITCHBOARD II Extended-data task is used to train the male and female UBMs, respectively. Trigram models are used for that purpose. The target-speaker models are created by linear interpolation of the corresponding UBM (gender-dependent) and the speaker training data. The interpolation coefficients are set to 0.8 for the speaker data and 0.2 for the UBM. By including the "general knowledge" provided by the UBM into the target-speaker models, the amount of data needed for a good estimation of the trigram models is reduced.

5.3. Scoring

The speaker detection score is computed using a conventional log-likelihood ratio test between the target-speaker model and the corresponding UBM averaged over all n-gram types. Tnorm [2] technique is applied for score normalization. Cohorts consist of 60 models from NIST SRE 2004 database.

6. Phone Trigram System

6.1 Feature extraction and signal processing:

Two schemes have been used for feature extraction:

- The Advanced Distributed Speech Recognition Standard Front-End defined in the standard ETSI ES 202 050 [8]. This standard includes mechanisms for protection against additive noise (a double Wiener filter) and against channel distortion (blind equalization). The standard includes a Voice Activity Detection algorithm to discard non-speech frames that have not been used in our system. Apart from the noise and channel robustness mechanisms the front-end is based on the typical 13 Mel-Frequency Cepstral Coefficients (MFCC) with delta and double delta coefficients. The C0 coefficient is combined with the log-energy per frame.
- Sphinx [12] feature extraction system. This system is based on 13 MFCC coefficients along with delta and double delta coefficients and C0. The built-in automatic gain control was also used. CMN technique was applied for channel compensation. Audio was filtered and frecuencies outside of the range 130Hz-3700Hz were discarded.

6.2 Acoustic Modelling

Our system makes use of three independet sets of gender and context-independent phone models (for English, Spanish and Basque) based on Hidden Markov Models (HMMs). The HMM topology is three-state left-to-right with no skips. The output *pdf*s of each state are modelled as GMMs. The number of Gaussians per state were adjusted on the NIST SRE'05 data task corpus to miminize speaker recognition EER.

The set of English phone HMMs was trained on the TIMIT corpus. Since this corpus is microphone speech sampled at 16 kHz, we filtered it to simulate the telephone channel and then downsampled it to 8 kHz. One gaussian/state was used to model output pdfs.

The set of Spanish phone HMM was trained on the Albayzin corpus. The same subsampling process as described above was applied for this case. Five gaussians/state were used to model output pdfs.

Both systems use the ETSI ES 202 050 parameteriser.

Basque SpeechDAT was used in order to train the Basque phone HMM set, modelling output pdfs with 20 gaussians per state. The parameterisation was performed using the Sphinx parameteriser.

All sets were trained using HTK v3.2.1.

6.3 Acoustic-Phonetic Decoding

Acoustic-phonetic decoding (phone recognition) was performed with every recogniser on Switchboard I, Switchboard II, NIST SRE'04, NIST SRE'05, NIST SRE'06 train and test files using HTK v3.2.1, the trained models and a null grammar. The only information used from the acoustic-phonetic decoding was the phone streams. The output phone streams were filtered to avoid repetitions of inter-word silences.

6.4 Training

The Universal Background Phone Model (*UBPM*) is a trigram language model trained with data from Switchboard I, Switchboard II, NIST SRE'04 and NIST SRE'05. Smoothing of unlikely trigrams was performed with absolute discounting. No cut-off factor was applied. A different UBPM was used for each phonetic decoder.

Speaker Phone Models (*SPM$_i$*) are created by linear interpolation of the 8 sides training material for each target speaker from NIST SRE'06 training data. The interpolation factor (weight of the UBPM) for this adaptation was adjusted on NIST SRE'05 extended data task and was found to be optimal for an UBPM weight of 0.7.

6.5 Scoring

Scoring of a test file against a target speaker *i* consists of producing the phone stream, *X*, from its acoustic-phonetic decoding and computing the log-likelihood ratio:

$$LLR_i = \frac{1}{N_i} \log \left(\frac{P(X \mid SPM_i)}{P(X \mid UBPM)} \right)$$

This log-likelihood ratio is normalized by the number of phones (N_i) in *X* to yield the final score. Both ngram training and scoring was performed using HTK Language Modelling Tools.

6.5 Score normalization

For score normalization Tnorm [2] was applied using as cohort a gender dependent set of 60 models extracted from NIST SRE'04.

7. Fusion Strategies

Linear SVMs are trained to separate the genuine and impostor distribution of scores obtained with the ATVS-UAM individual systems on data from NIST 2004 evaluation. The fused scores for NIST 2005 evaluation are obtained as signed distances to these separating hyperplanes [9].

8. Log-LR calibration and decision

The Pair Adjacent Violators (PAV) algorithm has been used for log-likelihood ratio (llr) calibration (see [11]). The score to likelihood ratio mapping has been trained using NIST SRE 2005 development data. The resulting non-parametric mapping function has been linearly interpolated in order to guarantee monotonicity.

The final decision has been taken assuming log-likelihood ratios and a prior-cost ratio (or ‘effective prior’) of 9.9 versus 1, as defined by NIST. Thus, the threshold is established at log(9.9) = 2.29.

For trials involving all training or test data detected as empty, the final score was set to Log-LR=0 (LR=1).

9. References

[1] D. Reynolds, T. Quatieri and R. Dunn, “Speaker Verification Using Adapted Gaussian Mixture Models," Digital Signal Processing, vol. 10, pp. 19--41, 2000.
[2] R. Auckentaler, M. Carey, H. Lloyd-Thomas, “Score Normalization for Text-Independent Speaker Verification Systems," Digital Signal Processing, vol. 10, pp. 42--54, 2000.
[3] D. Sturim, D. Reynolds, “Speaker Adaptive Cohort Selection for Tnorm in Text-Independent Speaker Verification,” in proc. of ICASSP 2005, Philadelphia, USA.
[4] M. N. Do, “Fast approximation of Kullback-Leibler distance for dependence trees and hidden Markov models,” Signal Processing Letters, IEEE, Volume 10, Issue 4, April 2003 Page(s):115 – 118.
[5] V. Wan, W.M. Campbell “Support vector machines for speaker verification and identification”. Neural Networks for Signal Processing X, 2000. Proceedings of the 2000 IEEE Signal Processing Society Workshop Volume 2, 11-13 Dec. 2000 Page(s):775 - 784 vol.2
[6] W.M. Campbell, “Generalized linear discriminant sequence kernels for speaker recognition”. Proceedings. (ICASSP '02). IEEE International Conference on Acoustics, Speech, and Signal Processing. vol.1 pp 161-164
[7] A. Adami, Modeling Prosodic Differences for Speaker and Language Recognition. PhD thesis, OGI, 2004.
[8] ETSI ES 202 050 (v1.1.3): “Speech processing, transmission and quality aspects (STQ); Distributed speech recognition; Advanced front-end features extraction algorithm; Compression algorithms.”
[9] D. Garcia-Romero, J. Fierrez-Aguilar, J. Ortega-Garcia and J. Gonzalez-Rodriguez, “Support Vector Machine Fusion of Idiolectal and Acoustic Speaker Information in Spanish Conversational Speech”, Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing, ICASSP, vol. 2, pp. 229-232, Hong Kong, April 2003.
[10] D. A. Reynolds, “Channel Robust Speaker Verification via Feature Mapping”, in Proceedings of ICASSP 2003, vol. 2 pp. 53-56.
[11] Niko Brümmer and Johan du Preez, “Application Independent Evaluation of Speaker Detection”, Computer Speech and Language vol. 20(2-3), pp. 230-275.
[12] Sphinx systems homepage: http://cmusphinx.sourceforge.net/html/cmusphinx.php
[13] R Collobert, S Bengio "SVMTorch: Support Vector Machines for Large-Scale Regression Problems" Journal of Machine Learning Research, vol. 1, pp. 143-160, 2001