I3A NIST SRE2010 System Description

Jesús Villalba, Carlos Vaquero, Eduardo Lleida, Alfonso Ortega, Antonio Miguel

Aragon Institute for Engineering Research (I3A)
University of Zaragoza, Spain
{villalba,cvaquero,lleida,ortega,amiguel}@unizar.es

1. Introduction

I3A has submitted several systems to NIST SRE2010 for the core-core, core-10sec, 10sec-10sec and core-summed conditions. The systems are different variants of the GMM-UBM with Joint Factor Analysis approach: JFA-LLR, JFA-SVM, iVectors and the fusion of them. All the systems share the same feature extraction and UBM. The speaker segmentation for the summed condition is performed using a speaker factors based system.

2. Feature Extraction

The front end extracts feature vectors of 20 MFCC including C0 (C0-C19) over a 25 ms hamming window every 10 ms (15 ms overlap), and first and second order derivatives are computed over the feature vector sequence.

Voice Activity Detection (VAD) is performed computing the Long-Term Spectral Divergence (LTSD) of the signal every 10 ms, and comparing it against a threshold as in [1]. For phone calls, where two channels are available, namely channel of interest and reference channel, the reference channel is used for crosstalk removal. For interview segments, ASR labels are used for removing the interviewer.

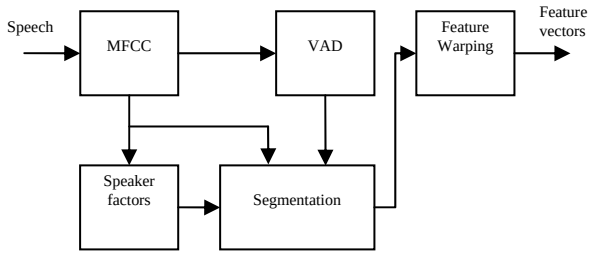
After frame selection, features are short time gaussianized with a 3 seconds window as in [2].

3. Speaker Segmentation

For the core-summed condition we used a segmentation system to generate two speaker dependent feature vector streams for every test segment.

To perform speaker segmentation, first, we compute a stream of speaker factors of dimension 20 for the given recording. These factors are computed using 12 MFCC with no derivatives (C1-C12) and a GMM of 256 Gaussians. Then, we model the stream with two Gaussians and, considering that every Gaussian belongs to a single speaker, we segment the stream using Viterbi decoding. The system is very similar to that proposed in [3], but we perform the algorithm on the whole segment, rather than in one minute segments. After this first segmentation, we apply two Viterbi re-segmentations using 12 MFCC as features, and GMM as speaker models, using soft-clustering in our second re-segmentation [4].

Segmentation is done over speech frames only. Then, the feature vector stream is separated into two different streams (one per speaker). After that, every stream is warped separately. The Feature extraction with the segmentation system is shown in the following figure:



4. JFA-LLR System

This system is a simplified version of [5] with the following configuration.

4.1. Universal Background Models

Gender Dependent (GD) and Gender Independent (GI) Universal Background Models (UBM) of 2048 Gaussians are trained by EM iterations. For this purpose we have used all the telephone signals in SRE2004, SRE2005 and SRE2006 databases (649 male speakers with 7412 signals and 801 female speaker with 9889 signals).

4.2. JFA Training

JFA Hyperparameters are trained from the previous background models. 300 eigenvoices (v) and 100 telephone eigenchannels (u_{phn}) are trained using telephone data from all the speakers of SRE2004, SRE2005 and SRE2006 databases having, at least, 8 recordings by speaker (530 male speakers with 7398 signals and 731 female speakers with 9938 signals). Another 100 eigenchannels (u_{mic}) are trained using all signals from speakers having far field microphone data in SRE2005 and SRE2006 and 50 speakers with interview data from SRE2008 (106 male speakers with 6244 signals and 119 female speakers with 6919 signals). Both eigenchannel matrices are stacked together for the core-core condition, for the other conditions, only the telephone eigenchannel matrix is

used. Finally, the remaining speaker variability matrix (d) is trained from the speakers of SRE2004, SRE2005 and SRE2006 having least than 8 recordings by speaker (201 male speakers with 547 signals and 152 female speakers with 668 signals). MAP estimates of speaker and channel factors are fixed for estimating d matrix to speed up the system. The d matrix is used in all the conditions but the 10sec-10sec. All the matrices are trained using ML+MD iterations.

4.3. Speaker Enrollment and Scoring

Speakers are enrolled into the system using MAP estimates of their speaker and remaining variability factors (y,z) so speaker means super vector is given by:

$$M_s = m_{UBM} + vy + dz \quad (1)$$

Trial scoring is performed using first order Taylor approximation of the LLR between the target and the UBM Models like in [6].

$$LLR \approx (vy_{trn} + dz_{trn})^t \Sigma^{-1} (F_{tst} - N_{tst} u\, x_{tst}) \quad (2)$$

ZTNorm score normalization is applied using telephone data from SRE2004, SRE2005 and SRE2006 (628 male speakers and 858 female speakers with 4 segments by speaker).

For the summed condition the maximum score of the two automatic segmented speakers is chosen.

5. JFA-SVM System

5.1 JFA Training

This system shares the same UBM and JFA matrices from the previous one.

5.1. SVM Scoring

Speaker enrollment and scoring is done by an SVM with the following kernel:

$$k(x_1, x_2) = (vy_1 + dz_1)^t \Sigma^{-1} (F_2 - N_2 u\, x_2) + (vy_2 + dz_2)^t \Sigma^{-1} (F_1 - N_1 u\, x_1) \quad (3)$$

Background signals for SVM training are chosen from SRE2004, SRE2005, SRE2006 and the 50 speakers from SRE2008 used in u_{mic} matrix training. For telephone background segments we have 653 male speakers and 883 female speakers with 4 signals by speaker. For microphone background segments we use 119 males and 106 female speakers with 2 signals by speaker and type of microphone.

ZTNorm is applied to the SVM score using the same SVM background segments.

For SVM training we use libsvm [7].

6. iVectors System

6.1. Total Variability Space Training

We follow the approach taken in [8]. Total variability space of dimension 400 is training using the same data as for JFA eigenvoices matrix by ML+MD iterations. LDA and WCCN are applied to the total variability factors (iVectors) for inter-session compensation reducing the iVector dimension to 200. For this purpose, we use the same data as for telephone eigenchannels matrix.

6.2. Scoring

Speakers are enrolled into the system using MAP estimates of their channel compensated iVectors. Trials are evaluated by cosine product of training and testing iVectors.

We apply ZTNorm using the same segments as in the JFA-LLR system.

7. Calibration and Fusion

7.1. Core-core

7.1.1. Calibration

For the core condition we have build a trial list using data from SRE2008 including most of the common conditions included in SRE2010 evaluation. Our development list includes all trials that can be done using all training short, long and follow-up versus all testing short, long and follow-up English SRE2008 segments. We keep out the 50 speakers used in JFA training. In total, we have around 4M male trials and 10M female trials. This big amount of trials is needed for robust calibration in the new operating point. Calibration has been done by linear logistic regression using FoCal package [9].

We do multiple condition calibration iteratively until convergence:

1. Gender dependent (male, female)
2. Channel dependent (mic-mic same channel, mic-mic different channel, mic-phn, phn-phn)
3. Length dependent (short-short, long-long, long-short, short-long)

7.1.2. Same-Different Microphone detection

As far as this information is not provided by NIST, we do automatic same/different microphone detection. For that, we use a iVectors like system. We use the microphone speaker factors corresponding to the u_{mic} matrix as features. We train LDA and WCCN using SRE2008 data from the 50 speakers used for JFA training and keeping the 12 more discriminative directions. The scores are normalized using SNorm [10]

with microphone segments again from the 50 speakers used for JFA training.
The system score is calibrated using FoCal with a same channel prior probability of 0.1. This way, for each mic-mic trials we get a probability of being same or different microphone trial. We use this probability together FoCal bilinear for training the calibration.

7.1.3. Fusion

The calibrated systems are channel dependent fused using again FoCal package.

7.2. Non core-core

For non core-core conditions we do gender dependent calibration and fusion using the det7 matching conditions of SRE2008.

8. Submitted Systems

The systems submitted for each of the conditions are:

- core-core:
    1. Fusion of JFA-SVM and JFA-LLR, both GD
    2. JFA-SVM GD
    3. JFA-LLR GD
- core-10sec:
    1. Fusion of JFA-LLR and iVectors, both GD
    2. JFA-LLR GD
- 10sec-10sec:
    1. Fusion of JFA-LLR and iVectors, both GD.
    2. JFA-LLR GD
- core-summed:
    1. JFA-LLR GI

9. Computational Resources

The system has been run on Intel Xeon E5520 2.27 GHz like Machines. The resources are given by trial and are approximated.

System	CPU	Memory
Speaker Segmentation	0.2 RT	3MB
JFA-LLR	0.25 RT	500 MB
JFA-SVM	0.4 RT	500 MB
iVectors	0.15 RT	400 MB

10. References

[1] Ramirez, J. et al., "Voice Activity Detection with Noise Reduction and Long-Term Spectral Divergence Estimation", in Proc ICASSP, II: 1093-1096, Montreal, Canada, 2004.

[2] Pelecanos, J. and Sridharan, S., "Feature Warping for Robust Speaker Verification", Odissey 2001.

[3] Castaldo, F. et al, "Stream Based Speaker Segmentation Using Speaker Factors and Eigenvoices", in Proc ICASSP, 4133-4136, Las Vegas, Nevada, 2008.

[4] Reynolds, D. et al "A Study of New Approaches to Speaker Diarization", in Proc Interspeech, 1047–1050, Brighton, UK, 2009

[5] Kenny, P., Joint factor analysis of speaker and session variability : Theory and algorithms - Technical report CRIM-06/08-13 Montreal, CRIM, 2005.

[6] Glembek, O., Burget, L., Dehak, N., Brummer, N., and Kenny, P., "Comparison of Scoring Methods used in Speaker Recognition with Joint Factor Analysis" In Proc ICASSP 2009, Taipei, Taiwan, April 2009

[7] Chih-Chung Chang and Chih-Jen Lin, LIBSVM: a library for support vector machines, 2001.

[8] Dehak, N., Dehak, R., Kenny, P., Brummer, N., Ouellet, P and Dumouchel, P., "Support Vector Machines versus Fast Scoring in the Low Dimensional Total Variability Space for Speaker Verification" In Proc Interspeech 2009, Brighton, UK, September 2009.

[9] http://sites.google.com/site/nikobrummer/focalbilinear

[10] Brummer, N., Strasheim, A., "AGNITIO's Speaker Recognition System for Evalita 2009". In Proc Evalita 2009, Reggio Emilia, Italy, 2009.