The LRDE-EPITA and MIT-CSAIL System Description for NIST 2010 Speaker Recognition Evaluation

Réda Dehak[1] and Najim Dehak[2]

[1] Laboratoire de Recherche et Développement de l'EPITA (LRDE-EPITA)
[2] MIT-CSAIL Spoken Language Systems
reda.dehak@lrde.epita.fr, najim@csail.mit.edu

For NIST 2010 speaker recognition evaluation, we submitted in collaboration with MIT-CSAIL Spoken Language Systems one system for the core-core condition and an equivalent for the 10sec-sec condition. These two systems are based on cepstral parameters and perform in the same way for the telephone data. For the core-core condition, interview and microphone data were not processed in the same way as telephone data. The two systems are based on total variability method [2, 1, 3].

1 Feature Extraction

We used cepstral features extracted using HTK toolkit. We extract 19 Mel Frequency Cepstral Coefficients together with log energy every 10ms within each 25ms hamming window. We use only speech part of audio files. We used Brno University Silence detector (for more details, see Brno 2010 submission) for telephone data of core-core condition and CRIM voice activity detection (see CRIM 2010 submission [4]) for microphone data. In the 10sec-10sec condition, we used NIST ASR transcription to remove silence segment. We apply a feature warping [5] using a 3s sliding window. Delta and double delta coefficients were computed and appended to the initial vector. The final vector had a 60 dimension.

2 core-core condition

The proposed system is composed of two separated subsystems; the first one is used for telephone-telephone speech trials. The second subsystem is used when we had microphone or interview speech for the target or test segment. We used a score calibration of the two subsystems at the end of each system. This calibration was done by the MIT-LL (see MIT-LL description for more details)

2.1 Telephone-telephone speech system:

This system is used when we had telephone speech in both target and test segment. It is based on the same system described in [2]. A 600 total factors was trained on telephone speech to define the total variability space. We apply a dimension reduction to 250 using Linear

Table 1: Speech dataset used to train the UBM, total variability space, LDA and WCCN

Speech dataset	UBM	Total Factors Space	LDA	WCCN
Switchboard II, Phases 1, 2 and 3	X	X	X	
switchboard Cellular, Parts 1 and 2	X	X	X	
Fisher English database Part 1 and 2		X		
NIST 2004 SRE	X	X	X	X
NIST 2005 SRE	X	X	X	X
NIST 2006 SRE	X	X	X	X

Discriminate Analysis (LDA) in combination with Within Class Covariance Normalization (WCCN) techniques to carry out channel compensation in total variability space. Table 1 describes the dataset used for each step of system building.

We compute the 250th dimensional vector for the target and test segment, and use a cosine distance for the score. Finally, we use a zt-norm to normalize the score.

2.2 Microphone or interview speech system:

This system is the same one as MIT-CSAIL Spoken Language primal system for microphone and interview data. It is based on 600th dimensional total factors space estimated using telephone speech. We append a 200th dimensional total factor space trained on microphone and interview data. A probabilistic LDA [6] was used to reduce the 800th dimensional total factors into 600th dimensional speaker space. The PLDA is composed by 800th dimensional mean vector estimated on telephone data, 800x600 eigenvoice matrix trained on telephone speech, 800x200 eigenchannels matrix trained on microphone and interview speech and full covariance matrix trained on telephone speech. Finally, we train an LDA similar to the first subsystem to reduce the 600th dimensional space into a 250th dimensional space except for the training data. We use here telephone, microphone and interview data to compute channel compensation matrices. The decision score is computed using a cosine distance between the target and test reduction vectors. We use s-norm to compute the final score. The s-norm impostors are chosen from NIST 2005 and 2006 SRE telephone and microphone data, as well as some interview data from NIST 2008 SRE.

3 10sec-10sec condition

This system is similar to the telephone-telephone system for the core condition, It was fusion with another system from MIT-CSAIL Spoken Language Systems submission the build the final MIT-CSAIL Spoken Language Systems primary system.

References

[1] Najim Dehak. *Discriminative and Generative Approaches for Long- and Short-Term Speaker Characteristics Modeling : Application to Speaker Verification.* PhD thesis, ETS, Montreal, CA, 2009.

[2] Najim Dehak, Reda Dehak, Patrick Kenny, Niko Brummer, Pierre Ouellet, and Pierre Dumouchel. Support Vector Machines versus Fast Scoring in the Low-Dimensional Total Variability Space for Speaker Verification. In *INTERSPEECH*, Brighton, UK, 2009.

[3] Najim Dehak, Patrick Kenny, Reda Dehak, Pierre Ouellet, and Pierre Dumouchel. Front end factor analysis for speaker verification. *submitted to IEEE Transactions on Audio, Speech and Language Processing*.

[4] P. Kenny, P. Ouellet, and M. Senoussaoui. *The CRIM Systems for the 2010 NIST Speaker Recognition Evaluation*, 2010.

[5] J. Pelecanos and S. Sridharan. Feature Warping for Robust Speaker Verification. In *IEEE Odyssey*, pages 213–218, Crete, Greece, June 2001.

[6] S. J. D. Prince and J. H. Elder. Probabilistic Linear Discriminant Analysis for Inferences about Identity. In *11^{th} International Conference on Computer Vision*, Rio de Janeiro, Brazil, 2007.