MIT-CSAIL Spoken Language Systems and Lincoln Labs NIST 2010 SRE Submission

Najim Dehak, Stephen Shum and Jim Glass

William Campbell, Douglas Sturim, Fred Richardson, Alan McCree, Pedro Torres-Carrasquillo,

Zahi Karam, Douglas Reynolds

In the NIST 2010 speaker recognition evaluation, we collaborated with MIT Lincoln Laboratory (MITLL) and LRDE-EPITA.

Feature extraction:

Our cepstral system operated on 19 mel-frequency cepstral coefficients together with the zeroth cepstral value every 10ms using a 25 ms Hamming window. Delta and double delta coefficients were then calculated over a 5-frame window using a first order polynomial fit. This 60-dimensional feature vector was subjected to feature warping [4] using a 3s sliding window. This MFCC configuration was provided by the MIT-LL. The silence detector for telephone data was provided by Brno University. It corresponds to the Hungarian speech recognizer labels (for more details, please see Brno 2010 submission). The speech activity detection for microphone data was obtained from CRIM (for more details, please see CRIM 2010 submission [6]).

Core condition:

We propose two systems for the core condition. Both systems are based on the total variability space [1][2][3].

- The primary system is composed of two separate subsystems; the first one is used when only telephone speech is present in both training and testing steps of the 2010 evaluation data. It is based on 600 total factors trained only on telephone speech [1]. The dimensionality is subsequently reduced to 250 via Linear Discriminant Analysis (LDA), and then Within Class Covariance Normalization (WCCN) is applied to carry out channel compensation in the total variability space [2]. Table 1 shows the list of corpora and their respective roles in the creation of our system. Similar to [1], we used the cosine scoring and zt-normalization to make the final decision. As with everything else so far, the impostors for zt-norm were entirely selected from telephone speech data.

The second subsystem is used when we have microphone and interview data in training or in testing. This system is based on the total variability space and its 600 total factors estimated in telephone speech along with an additional 200 total factors trained on microphone and interview data. We then use Probabilistic LDA [7] to project all microphone and telephone total factors of dimension 800 into a speaker space of dimension 600. The PLDA consists of a mean vector of dimension 800 estimated from telephone data, an eigenvoices matrix of dimension 800x600 trained on telephone speech, an eigenchannels matrix of dimension 800x200 trained solely on microphone and interview speech, and a full covariance matrix trained from telephone speech. After the projection using PLDA, we applied LDA to reduce the 600 dimensions to 250 as well as WCCN projection techniques to carry out a second channel compensation in the speaker space. These channel compensation matrices were estimated using telephone, microphone and interview data pooled together. As before, the decision scores were computed using cosine scoring, but the final scores were obtained after s-normalization [5]. The impostors used for s-norm are taken from NIST 2005, 2006 SRE telephone and microphone data, as well as some interview data from NIST 2008 SRE.

Table 1: Corpora used to estimate the UBM, total variability matrix (T), LDA and WCCN

	UBM	T	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

- The second system was a blind system, independent of training and test conditions, which means that exactly the same system was applied for both telephone and interview data. The only required conditioning is in the score calibration level. This system corresponds exactly to the second subsystem of the primary system which was, in the previous case, used only in the presence of microphone and interview data. The motivation for the development of this blind system suited for all data conditions came from the fact that all the systems components were trained in both telephone and interview data. The score calibration was done by MIT-LL (For more details please read MIT LL NIST 2010 SRE submission).

10 sec – 10 sec condition

We proposed two systems for the 10sec-10sec condition:

- The primary system is a fusion of two systems, both of which are based on total variability space of 600 total factors trained in telephone data. The only difference between these two systems is in the MFCC features. The first system used the same features as previously described and is exactly the same system as the one described for the primary core condition. The second system used similar MFCC features except feature warping was applied before the computation of delta and double delta coefficients. This system was built in collaboration with LRDE-EPITA, and the fusion of both systems was done using logistic regression from the FoCal toolkit [8].
- The alternate system corresponds to the first subsystem of the primary core system in which all the parameters were trained on telephone speech only.

Reference:

[1] Najim Dehak, Réda 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 Proc INTERSPEECH 2009, Brighton, UK, September 2009.

[2] Dehak, N., Kenny, P., Dehak, R., Ouellet. P., Dumouchel. P., «Front end factor analysis for speaker verification» submitted to IEEE Transactions on Audio, Speech and Language Processing.

[3] N, Dehak «Discriminative and generative approaches for long- and short-term speaker characteristics modeling : Application to speaker verification.» PhD thesis at École de Technologie Supérieure de Montreal.

[4] J. Pelecanos and S. Sridharan« Feature Warping for Robust Speaker Verification,» Proc. Speaker Odyssey, Crete, Greece, pp 213-218, jun 2001.

[5] P. Kenny, «Bayesian speaker verification with heavy tailed priors,» in Proc. Odyssey 2010: The speaker and Language Recognition Workshop, Brno, Czech Rebublic, June 2010.

[6] P. Kenny, P. Ouellet, M. Senoussaoui «The CRIM System for the 2010 NIST Speaker Recognition Evaluation» NIST 2010 SRE workshop

[7] S. J. D. Prince and J. H. Elder, «Probabilistic linear discriminant analysis for inferences about identity» in Proc. 11 International Conference on Computer Vision, Rio de Janeiro, Brazil, Oct. 2007.

[8] Niko Brummer, Lukas Burget, Jan Honza Cernocky, Ondrej Glembek, Frantisek Grezl, Martin Karaat, David A. van Leeuwen, Pavel Matejka, Petr Schwarz, and Albert Strasheim, "Fusion of Heterogeneous Speaker Recognition Systems in the STBU Submission for the NIST Speaker Recognition Evaluation 2006," in IEEE Trans. On Audio, Speech and Language Processing, September 2007.