SVIST System Description: NIST SRE 2010

Changli, Chaoqunliu,Shenshenlin

Shanghai Voice Info Science and Technology Ltd.

lichang.nppl@gmail.com,chaoqunliu1@163.com, shenshenlin@live.cn

SVIST(Shanghai Voice Info Science and Technology Ltd.) participated in three tasks, core-core, 8conv-core, 10sec-10sec.

For core-core and 8conv-core task, the primary system is a fusion of three sub-systems(system1): one based on Joint Factor Analysis [1] channel compensation and PLP feature, the other is Total variability, Cosine Distance(TVCD) [2] and LPCC feature, the last one is GMM-SVM-NAP[3] and MFCC feature. Then a fusion of JFA-PLP system and GMM-SVM-MFCC sub-system is used as an alternate system(system2).

The primary system for 10sec-10sec task is a fusion of three sub-systems(system3): one sub-systems is based on eigenvoice modeling and MFCC feature, the second is about eigenvoice modeling and PLP feature,the third one is GMM-SVM with MFCC feature. Then a fusion of the second and third sub-system is used as an alternate system(system4).

Feature Extraction:

We use three types of features:

MFCC: Short time gaussianized MFCC (19 dimentions) + energy augmented with their delta and double delta coefficients, making 60 dimensional feature vector. The analysis window has 25 ms with shift of 10ms. A simply energy based VAD is used to remove silence.

LPCC: Short time gaussianized LPCC (18 dimentions) augmented with their delta coefficients, making 36 dimensional feature vector. The analysis window has 25 ms with shift of 10ms. The ETSI VAD are modified to be used.

PLP: Short time gaussianized PLP (18 dimentions) with their delta coefficients, making 36 dimensional feature vector. The analysis window has 30 ms with shift of 10ms. A simply energy based VAD is used to remove silence.

Static short time parameters of MFCC,LPCC,PLP are all extracted by HTK tools.

For interview data, interviewer's speech is removed by our vad tag by simply energe based VAD.

Joint Factor Analysis based sub-system

We trained two gender-dependent UBM's having 2048 Gaussians and gender-dependent factor analysis models having 300 eigenvoices estimated from telephone data, 100 eigenchannels estimated from telephone data, 50 eigenchannels estimated from microphone data and 100 eigenchannels estimated from interview data. PLP feature is used.

For UBM training, we used the NIST 2005, 2006 and 2008 SRE English telephone data. NIST 2005, 2006, 2008 SRE telephone data are used to train the eigenvoices, and NIST 2004 SRE data are used to train the diagonal matrix. As to the complex conditions in evaluation plan, we

combined three different eigenchannels to get the final eigenchannels: NIST 2006 and 2008 SRE telephone eigenchannels, NIST 2005, 2006 and 2008 microphone eigenchannels, NIST 2008 and followup interview eigenchannels.

Gender and channel dependent ztnorm is applied. We selected 500 tnorm speakers and 400 znorm utterances for every condition. All these imposters were taken from the same dataset in the JFA training.

Total Variability Cosine Distance based sub-system

The similiar UBM training and ztnorm as mentioned in JFA sub-system are used but with LPCC feature.

The total variability matrix was trained in NIST 2004, 2005, 2006, 2008 SRE telephone data. We used 400 total factors. The LDA matrices were trained on the same data as the total variability matrix. The within class covariance matrix were trained in NIST 2005, 2006 and 2008 SRE data. As to the cross-trials, we also selected telephone, microphone and interview data as we did in JFA system for balance. Finally we used cosine distance to get the original score before ztnorm.

GMM-SVM-NAP sub-system

The same UBM training mentioned in JFA sub-system is used. We get the GMM mean vectors using MAP adaptation from UBM. Normalized mean vectors are concatenated to form a supervector. Nuisance attribute projection (NAP) is used to remove the channel effect.

The NAP matrix is trained in NIST 2005, 2006 and 2008 SRE data. As to the cross-trials, we selected telephone, microphone and interview data as we did in JFA system for balance. the background imposters selected from the same dataset.

Gender and channel dependent tnorm is applied. We selected 500 tnorm speakers for every condition. All these imposters were taken from the same dataset in the NAP matrix training.

GMM-SVM sub-system

The same UBM training mentioned in JFA sub-system is used except that the mixtures for the gender dependent UBM are both 512 . We get the GMM mean vectors using MAP adaptation from UBM, as we did in GMM-SVM-NAP system. Normalized mean vectors are concatenated to form a supervector.

Gender and channel dependent tnorm is applied. We selected 500 tnorm speakers for every condition. All these imposters were taken from the same dataset in the NAP matrix training.

Eigenvoice sub-system

Eigenvoice[4] is used as the model training algorithm. NIST 2005, 2006, 2008 SRE telephone data are used to train the eigenvoices . MFCC and PLP features are used in two different sub-systems.

Gender dependent ztnorm is applied. We selected 500 tnorm speakers and 300 znorm utterances.

System fusion

Linear fusion of scores from different systems are used as the final score for the submitted system. A development database was carefully designed from NIST SRE 08 database, to simulate the training and test conditons in SRE 2010. And the weights are obtained from this developments database , and then used in SRE 2010. The scores can not be interpreted as log likelihood ratios.

1. Kenny, P., Ouellet, P., Dehak, N., Gupta, V., and Dumouchel, P. A Study of Inter-Speaker Variability in Speaker Verification *IEEE Transactions on Audio, Speech and Language Processing*, July 2008.
2. Dehak, N., Kenny, P., Dehak, R., Dumouchel, P and Ouellet, P. Front-End Factor Analysis for Speaker Verification *submitted to IEEE Transactions on Audio, Speech and Language Processing*, November 2009.
3. W. M. Campbell, D. E. Sturim, D. A. Reynolds, SVM based speaker verification using A GMM supervector kernel and NAP variability compensation. Proc. IEEE ICASSP'06, 2006, vol. 1, pp. 97–100.
4. Kenny P, Eigenvoice Modeling with sparse training data, IEEE Transactions on speech and audio processing. 2005, 13(3): 345-354.