CCNT system description: NIST SRE 2010

Yingchun Yang[1], *Zhenchun Lei*[2], *Ting Huang*[1], *Li Chen* [1], *Wenxiang Chen*[1], *Qi Yu*[1], *Zhenyu Shan*[1], *Rui Huang*[1].

[1]Zhejiang University, China
[2]Jiangxi Normal University, China

{yyc,huangting,stchenli,cwx,yuqi0103,shanzhenyu,hr}@zju.edu.cn,
zhenchun.lei@hotmail.com

1. INTRODUCTION

CCNT submitted three systems to NIST SRE 2010 evaluations, only to the core test. The primary system is a fusion of two sub-systems (also as the alternate systems) using two different cepstral features and 2 different classifiers. The Alternate System I is based on the generative GMM-UBM [1] approach, the Alternate System II is based on discriminative SVM techniques [2]. Only the Alternate System I was submitted to the later extended core test. We mainly outline in this document the CCNT system including feature extraction, classifiers, fusion methods. Some implementation details such as the processing speed and the utilized software are also introduced.

2. SUBMITTED SYSTEMS

2.1 NIST SRE 2010 core test

Three systems were submitted including primary system, Alternate System I and Alternate System II.

Our primary system is a fusion of the two subsystems:

- GMM LFA (Alternate System I)
  Gender dependent UBM GMM LFA system with MFCC13/LPCC18 features and gender dependent z-norm.

- CD UBM SVM GSV NAP (Alternate System II)
  Channel Dependent UBM (CD UBM) SVM GSV NAP with MFCC13/LPCC18 features and gender dependent t-norm.

2.2 NIST SRE 2010 extended core test

Only the Alternate System I was submitted.

- GMM LFA (Alternate System I)
  Gender dependent UBM GMM LFA system with MFCC13/LPCC18 features and gender dependent z-norm.

Fig.1 presents the framework of our CCNT system.

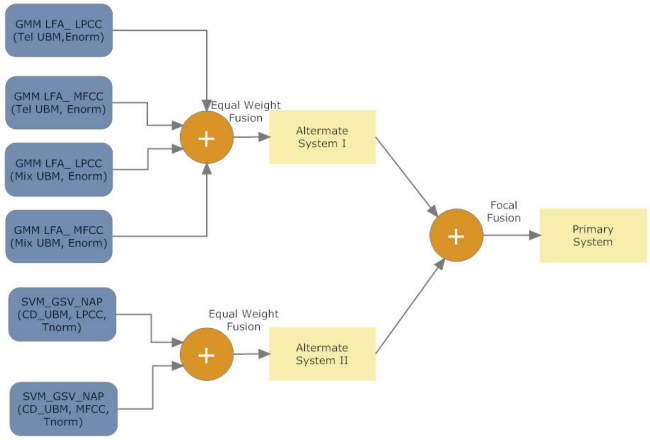

Fig. 1. CCNT system framework

3. FEATURE EXTRACTION

Two sets of features are used for recognition—MFCCs and LPCCs. For MFCCs, 13 cepstral coefficients and delta are computed to produce a 26 dimensional feature vector. For linear prediction (LP) based processing, 18 LPCCs are obtained from 18 LP coefficients extracted using the standard Levinson-Durbin recursion. Both MFCCs and LPCCs feature vector streams are processed through an energy-based voice activity detector (VAD) to eliminate non-speech vectors. CMS and variance normalization are then applied to the feature streams. The two features share the same parameters setting in the procedures including windowing , frame rate, CMS, and variance normalization[3].

4. CLASSIFIERS

4.1 GMM LFA

The GMM LFA (latent factor analysis) classifier was similar with the work presented in [4].

4.1.1 Universal background models (UBM)

The two gender-dependent UBMs are trained on NIST SRE08 data. Two sorts of UBMs named Tel-UBM and

Mix-UBM are adopted. For Tel-UBM, only the data from telephone channel are used for training, while for Tel-UBM, all the data are used for training, from channels including telephone ， microphone and interview. EM algorithm with 20 iterations is used for training 512 mixture GMM UBM[].

4.1.2. Latent Factor analysis

The Latent Factor analysis (LFA) [3-5] system is used with MFCC13 and LPCC18 features. First, for the LFA system, the channel matrix rank is set to 50 and the regulation factor is 14. Considering the interview channel data was provided only in NIST SRE 2008 data, the interview and microphone Eigen-channel matrices are constructed on all the data from both NIST SRE 2008 and follow up data selecting those speakers with more than 5 utterances. For telephone Eigen-channel, we choose NIST 2004, 2006, 2008 SRE as the training data and also select those speakers with more than 6 utterances. As the NIST SRE 2010 corpus is large, we only use the Eigen-channel on the enrollment data, not on the test trails.

4.1.4 Normalization

Z-norm was performed on the scores [6]. The Z-norm imposters are drawn from NIST SRE08 data. 121 males and 162 females segments are used for the interview and microphone channel score normalization. Then, 371 males and 400 females are chosen for telephone channel score normalization.

4.2 SVM GSV NAP

Our SVM GSV NAP classifier is SVM GMM supervector (GSV) using a linear kernel followed by NAP projection [2,3,7].

4.2.1 CD UBM (Channel Dependent Universal background models)

Gender dependent and channel dependent 512 mixture GMM UBMs are trained. The whole evaluation set is split into 10 different type according to gender and train-test type. UBM for each type is trained independently, and the number of utterances for each type is listed in Table 1.

	INT	INT_TEL	TEL	MIC	INT_MIC
female	2204	5023	3638	829	3204
male	1615	3158	2827	649	2255

Table 1: Number of utterances for training CD UBM under different channels. (INT, TEL and MIC means train and test type are all interview, telephone and microphone. INT_TEL means train type is interview and test type is telephone, INT_MIC means train type is interview and test type is microphone.)

All utterances in 2008 short2-short3 evaluation corpus are used to train CD UBM(Channel Dependent Universal background models). For mic-UBM and tel-UBM, 2010 tarball data with different vocal effort are also included. For each utterance in training corpus, GMM is adapted from different UBM according to the channel type of the utterance. In the evaluation step, GMM is chosen according to the train-test type.

4.2.2 SVM GSV

GSV (GMM supervector) is computed as follows. 2010 core evaluation data are split into five subsets according to train-test channel type. All of the training data with each utterance of about 3, 5 or 8 minutes are split into 8 parts evenly, with 20% overlap with adjacent parts. Each part is guaranteed to be more than 1000 frames. Each part is adapted to GMM with certain channel-UBM. The means of this GMM UBM are then stacked to form a GSV. For SVM GSV training, the previous GSVs from training data are positive samples while the GSVs constructed on the utterances in NIST 2004 1-side training corpus are selected as negative samples. The number of negative samples is 368 for female and 248 for male.

4.2.3 NAP (Nuisance Attribute Projection)

NAP is applied to GMM supervector to remove the unwanted channel session variability by removing the principal components of this variability. The training data for NAP matrix are selected form 2008 SRE core evaluation corpus. Those speakers with more than 5 utterances are selected. The number of utterances and speakers for each channel is shown in table 2. The figure outside the bracket is the number of utterances and the inside is the number of speakers. The number of principal components is set to 50.

	INT	INT_TEL	TEL	MIC	INT_MIC
female	2204(87)	3766(295)	1399(217)	820(75)	3024(160)
male	1616(63)	2446(169)	734(121)	640(58)	2255(116)

Table 2: Training data for NAP matrix under different channels.
(The meanings of the parameters are the same as Table 1.)

4.2.4 Normalization

T-norm is applied to score normalization [6]. We select 400 utterances (200 from telephone channel, 100 from interview channel and 100 from microphone channel) from SRE08 core evaluation corpus as cohorts. And these utterances are selected to cover speakers as many as possible.

5. FUSION

Two stages of fusion are adopted in our CCNT system. The first stage fusion is performed inside the two Alternative systems of GMM LFA and SVM GSV NAP. The second stage fusion integrates the two Alternative systems to generalize our primary system.

5.1 Stage 1 fusion

5.1.1 GMM LFA

The output score of the GMM LFA system (Alternate System I) is fusion of the following four subsystem scores:
GMM LFA_ LPCC(18+18 delta)_Tel-UBM,
GMM LFA_ MFCC(13+13 delta)_Tel-UBM,

GMM LFA_ LPCC(18+18 delta)_Mix-UBM,
GMM LFA_ MFCC(18+18 delta)_Mix-UBM.

The output score of the GMM LFA system (Alternate System I) is generalized by fusing the scores from four subsystems with equal weights.

5.1.2 SVM GSV NAP

We build two sub-systems for SVM GSV NAP (Alternate System II) as follows:
SVM GSV NAP_ LPCC(18+18 delta)
SVM GSV NAP_ MFCC(13 + 13 delta)

The output score of the SVM GSV NAP system (Alternate System II) is generalized by fusing the scores from two subsystems with equal weights.

5.2 Stage 2 fusion

Stage 2 fusion combines the output sore of the two Alternative systems with prior weights. The prior weights are obtained by FOCAL Toolkit [8] with the minimum equal-error-rate (EER) criterion on the NIST 2008 SRE data.

6. RESULTS ON DEVELOPMENT SET

This section reports the results on our development set on NIST SRE 2008 data for core test. The results are listed in table 3 and 4.

7. IMPLEMENTATION DETAILS

7.1 Processing Speed

The processing speed of the system is measured on a standard PC Intel Xeon E5420 processor with 4 Gb RAM.

For GMM LFA system, the real time factor is approximately 1.02×RT for training and 0.019×RT for testing.

For SVM GSV NAP system, the real time factor is approximately 0.43×RT for training and 0.11×RT for testing.

7.2 Utilized Software

The software we used are Alize [9] and FOCAL Toolkit [8].

Primary system						
	Int-int	int-tel	tel-tel	tel-int	tel-mic	all
EER	0.017774	0.055528	0.034727	0.038456	0.058919	0.037122
minDCF	0.0090087	0.029919	0.018588	0.015268	0.024183	0.019479
SVM GSV NAP						
EER	0.02014	0.071002	0.040434	0.090164	0.094185	0.06209
minDCF	0.011285	0.033057	0.020042	0.036102	0.045458	0.038477
GMM LFA						
EER	0.02722	0.099118	0.057714	0.041706	0.074831	0.051513
minDCF	0.029669	0.044077	0.028289	0.024495	0.029309	0.0259

Table 3: Results on NIST SRE 2008 data (Female).

Primary system						
	Int-int	int-tel	tel-tel	tel-int	tel-mic	all
EER	0.0088571	0.043283	0.027218	0.045337	0.040005	0.02734
minDCF	0.0037425	0.016974	0.012306	0.020511	0.016547	0.014494
SVM GSV NAP						
EER	0.0098159	0.029614	0.033467	0.08757	0.060936	0.044478
minDCF	0.0045662	0.012321	0.013963	0.030664	0.031142	0.030289
GMM LFA						
EER	0.014555	0.072002	0.041225	0.048367	0.049664	0.039571
minDCF	0.0066624	0.036083	0.017533	0.021532	0.023617	0.017958

Table 4: Results on NIST SRE 2008 data (Male).

8. ACKNOWLEDGMENTS

This work was supported by National Natural Science Foun dation of China (NSFC60970080) and the Special Funds for Key Program of the China(2009ZX01039-002-001-04)

9. REFERENCES

[1] D.A. Reynolds, T.F. Quatieri and R.B. Dunn, “Speaker Verification Using Adapted Gaussian Mixture Models”, *Digital Signal Processing*, 10(1):19--41,2000.
[2] W. M. Campbell, J. P. Campbell, D. A. Reynolds, E. Singer, and P. A. Torres-Carrasquillo, “Support Vector machines for

speaker and language recognition," *Computer Speech and Language*, vol. 20, no. 2-3, pp. 210-229, 2006.

[3] Sturim, D., Campbell, W., Karam, Z., Reynolds, D. A., Richardson, F., The MIT Lincoln Laboratory 2008 Speaker Recognition System, Interspeech 2009, Brighton, UK, Sept. 6, 2009

[4] Robbie Vogt, Brendan Baker, and Sridha Sriharan, "Modelling session variability in text-independent speaker verification," Proc. Interspeech, 2005, pp. 3117–3120.

[5] P. Kenny, G. Boulianne, and P. Dumouchel, "Eigenvoice modeling with sparse training data," IEEE Trans. Speech and Audio Processing, vol.13, no. 3, pp. 345–354, 2005.

[6] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, "Score normalization for text-independent speaker verification systems," Digital Signal Processing, vol. 10, no 1-3, pp. 42-54, Jan 2000.

[7] WM Campbell, DE sturim, Da Reynolds, A Solomonoff. "SVM based speaker verification using a GMM supervector kernel and NAP variability compensation". ICASSP 2006, p97-100.

[8] FOCAL Toolkit: http://niko.brummer.googlepages.com/ .

[9] Jean-Francois Bonastre, Nicolas Scheffer, Driss Matrouf, Corinne Fredouille, Anthony Larcher, Alexandre Preti, Gilles Pouchoulin, Nicholas Evans, Benoˆıt Fauve and John Mason, "ALIZE/SpkDet: a state-of-the-art open source software for speaker recognition", odyssey2008, Stellenbosch, South Africa,January 21-24, 2008