IFLY SYSTEM FOR THE NIST 2008 SPEAKER RECOGNITION EVALUATION

Wu Guo, Yanhua Long, Yijie Li, Lei Pan, Eryu Wang, Lirong Dai

MOE-Microsoft Key Laboratory of Multimedia Computing and Communication,
University of Science and Technology of China (USTC),China
{guowu, lrdai}@ustc.edu.cn, {lyhsa,andylyj,leipan,eryuwan}@mail.ustc.edu.cn

ABSTRACT

The description of iFLY system submitted for NIST 2008 speaker recognition evaluation (SRE), which has achieved excellent performance in the 2008 SRE evaluation, is presented in this paper. Our primary system is a fusion of two subsystems GMM-UBM and GMM-SVM. For each sub-system, two kinds of short-time acoustic features PLP and LPCC are adopted. We focus on three key issues in this evaluation: channel compensation, multi-lingual or bi-lingual cues and the voice activity detection. We also point out that data selection and factor analysis play key roles in the system improvement.

Index Terms— speaker verification, joint factor analysis, NAP, GMM.

1. INTRODUCTION

The 2008 NIST speaker recognition evaluation is part of an ongoing series of evaluations conducted by NIST[1].The SRE has focused on the cross-session (channel) problem from the very beginning. But in the previous core test,the session mainly included conversational telephone speech recorded on different telephone channels (GSM, CDMA, Landline) with different handset types (Elect., Carbon).The 2008 evaluation tasks are distinguished by including in the training and test conditions not only conversational telephone speech but also interview speech recorded with different microphones involving an interview scenario.The primary evaluation condition is short2-short3,in which the training condition is called short2 and the test condition short3.The Short2 training data includes conversational telephone speech and interview speech.The Short3 testing data consists of not only the previous two which are of the same type as Short2 but also the telephone speech recorded over an ancillary microphone channel.

In recent years, factor analysis[2][3][4] has been applied to the GMM-UBM system[5] to address the channel or session variability problem with great success.To solve the complicated cross-session problem in NIST-2008,inter-speaker variability modeling[2] which can capture the distinct information of different speakers based on factor analysis is adopted in our GMM-UBM sub-system;and for our GMM-SVM sub-system, the nuisance attribute projection (NAP)[6] approach is adopted.

The interview data for the evaluation is the MIX 5 corpus[7].And the speech data in this corpus is recorded concurrently over several microphones with different locations.Because of the different distances of speakers from the microphones, the energy of some utterances is very low while that of others may be very high.So a robust voice activity detector (VAD) is necessary to remove the silence from all utterances.A robust VAD which is based on energy optimized for the evaluation task is reported.

Bi-lingual speaker is another session variability introduced in NIST SRE evaluation since 2004.The 2008 SRE bi-lingual evaluation is similar to the 2004 and 2006 evaluations, but different from the 2005 evaluation.There are about 83% English utterances and 17% non-English utterances in 2008 SRE evaluation. The language balance is also an important factor affecting the system performance. Thus we take this problem into account in UBM training, NAP, factor analysis and score normalization process.

2. ACOUSTIC FEATURE EXTRACTION, VAD

Two types of features are used in our system: the 39-dimensional PLPs and 36-dimensional LPCCs.

We use the HTK tool to extract PLP features. Speech is segmented into frames by a 20-ms Hamming window progressing at a 10-ms frame rate. Each speech frame is parameterized by the 13th order PLPs and their first and second derivatives (i.e., a 39-dimensional feature vector). Further processing including RASTA filtering, VAD, CMS and Gaussianization[8] are applied to all PLPs.

Meanwhile, the SPTK tool is used to extract the LPCC features. In this case, speech is segmented into frames by a 30-ms Hamming window progressing at a 10-ms frame rate. And Each speech frame is parameterized by the 18th order LPCCs and their first derivative (i.e., a 36-dimensional feature vector). LPCCs are also preprocessed by RASTA filtering,

This work is partly supported by Science Research Fund of MOE-Microsoft Key Laboratory of Multimedia Computing and Communication(Grant No.07122803)

VAD, CMS and Gaussianization.

In speech recognition, a phoneme "sil" is used to indicate the silence. However, there is no such phoneme in speaker recognition because of the GMM algorithms. The speaker recognition relies more on the VAD to remove silence or noise frames than the speech recognition does. In this year's evaluation, the VAD is especially important. Since the telephone speech has a normal energy distribution as the previous year's evaluation, it can be easily processed by traditional VAD. But the energy of the interview data and telephone ancillary microphone data gives a variable distribution and can not be processed by traditional VAD. Besides, there are two situations that are even hard to deal with. First, some utterances are too low for us to hear. Second, some of the utterances are contaminated by noises, resulting in very low SNR. They are more like a frequency modulation (FM) signal than a speech.

We take the VAD algorithm in [9] as the basis of our approach but make some important modifications.In [9], energy thresholds are pre-defined by the paper to decide whether a frame is speech or a silence frame. In our scheme, we take a different approach. For a five-minute utterance, the energy of the first 6 seconds is calculated to decide the average energy of the whole utterance, the SNR of the signal and thresholds of the VAD. Since the speech is stable, these thresholds can be used for the whole utterance. Moreover, because the energy thresholds are defined by the first 6 seconds, we can always detect some "speech" frames even if the speech can not be audible. As a result, we can remove about 10% to 70% silence frames from each interview utterance.

In [10], noise reduction algorithms such as Wiener-filtering are applied to speech signals. In our SRE 2006 auxiliary microphone task, the best result is achieved without any noise reduction algorithm. So we use the original speech to extract the acoustic feature instead of the signal after denoising. For some utterances like FM signal, all the frames are treated as speech signals.

3. GMM-UBM AND JFA

3.1. Universal Background Model

The gender dependent GMM-UBM system is adopted in the evaluation. NIST SRE2004 1side training corpus is used to train two gender-dependent UBMs with 1024 Gaussian components. There are totally 367 female utterances and 249 male utterances in NIST SRE 2004 1side training corpus. Though the number of these utterances is not large enough, these utterances are language balanced and channel balanced, which makes them a reasonable choice for UBM training.

3.2. Utilization of MIX 5 development corpus

The interview data is new to all participants. NIST has released 6 persons' (3 females and 3 males) development data to all participants. Each person has 6 sessions of 30-minute speech. There are 9 channels speech for each session; so we have 9 conversations that have the same phone sequence with different channel information. They are fit for channel matrix training. In our experiments, each session is averagely divided into 6 segments with about five minutes in each segment. After removing the silence parts according to the VAD tag provided by the NIST, the duration for each segment is about 3 minutes. It is similar to the training and testing condition in the evaluation. Therefore, for each person there are 6 sessions$\times$9 channels$\times$6 segments = 324 utterances in total. Though the utterances in the first channel are not used in the evaluation, we use all these utterances to train the channel matrix in the factor analysis.

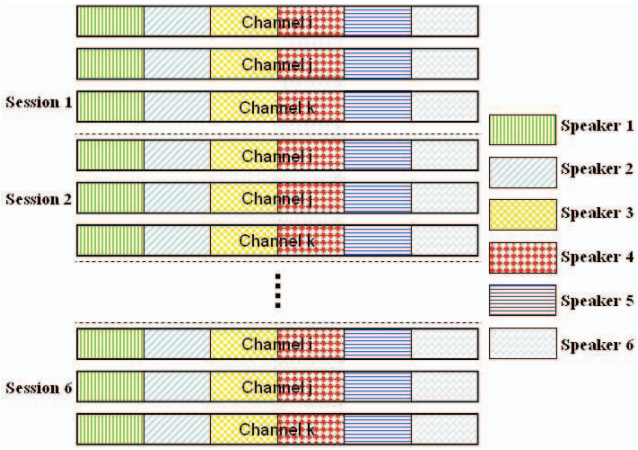

Fig. 1. Map 1 person speech from 6 sessions to 6 speakers

We have 3 persons for each gender. Each person has 324 utterances. It is not a good choice to put all the utterances from the same person together to train the channel matrix. Thus, we use a map trick to get more speakers. This procedure is shown in Fig. 1. For each session, we randomly select three channels from the nine channels. Then we select 1 segment from each channel to form the utterance of one fake speaker. There are 6 sessions for each person. So we get 3 channels$\times$6 sessions$\times$1 segment = 18 utterances. We use these 18 utterances to form a fake speaker; and we can fake 18 speakers from each person. After these steps, we can obtain 54 speakers for each gender. These 54 speakers are used for channel matrix training in factor analysis and NAP. Also, they are used as SVM negative samples.

3.3. Joint Factor Analysis

We use the large joint factor analysis model proposed by Patrick Kenny[2].The channel-dependent speaker mean supervector $\boldsymbol{M}$ can be represented as

$$\boldsymbol{M} = \boldsymbol{m} + \boldsymbol{vy} + \boldsymbol{dz} + \boldsymbol{ux} \quad (1)$$

where $\boldsymbol{m}$ is the UBM super-vector, $\boldsymbol{v}$ is the speaker loading matrix, $\boldsymbol{d}$ is the diagonal loading matrix and $\boldsymbol{u}$ is the channel

loading matrix. $\boldsymbol{y}$, $\boldsymbol{z}$ and $\boldsymbol{x}$ are the speaker factors, diagonal factors and channel factors respectively. The factor number of $\boldsymbol{v}$ is 300, and that of $\boldsymbol{u}$ is 200 in the evaluation.

The Switchboard II and Switchboard Cellular corpus are used to train the speaker loading matrix with 300 speaker factors. And the SRE 2004 corpus is used to train the diagonal matrix. We have only adopted the maximum likelihood algorithm to train the loading matrix, which is different from the training procedure in [2]. For channel loading matrix, a telephone loading matrix with 100 channel factors is trained based on the telephone data from NIST SRE2004, NIST SRE2005 and NIST SRE2006 corpus, which are multilingual. A microphone loading matrix with 50 channel factors is trained based on the ancillary microphone data from NIST SRE2005 and NIST SRE2006 corpus with English as their language. Finally, the MIXER5 six persons' development corpus is used to train an interview loading matrix with 50 channel factors. The language of MIX5 is also English. After that, the full channel loading matrix with 200 channel factors is formed by appending the above three loading matrixes.

3.4. ZTnorm

Gender-dependent ZTnorm is applied to the log-likelihood ratio scores. There are five sub-sections in the evaluation. The score distributions of these five parts are varied because of their different channel and language information. We can not get channel information in the evaluation; but the language information has been released by NIST. The telephone data is a multi-language corpus; the ancillary microphone and interview data are all English. So different cohorts' speech are used for ZTnorm according to the training and testing language information. And the channels of these cohorts speech are balanced.

For Znorm, we select about 1000 utterances from NIST SRE 2005 1side training and testing corpus for telephone data, and 1000 utterances from NIST SRE 2005 ancillary microphone data for interview data.

For Tnorm, we select 1000 utterances from NIST SRE 2006 1side training and testing corpus for telephone data, and 1000 from NIST SRE2006 ancillary microphone data for microphone and interview data.

4. GMM-SVM AND NAP

In GMM-SVM speaker recognition system, gender dependent UBM is trained and the GMM component number is 512. The GMM super-vector is adapted to the UBM with a relevance factor 8. NIST SRE 2005, 2006 and MIX 5 development corpus are used for NAP training procedure. 125 bi-lingual telephone speakers from NIST 2006, 34 speakers from NIST SRE 2005, 32 speakers from NIST SRE 2006 recorded both on the telephone and on the ancillary microphone channel and 3 persons' data in the MIX 5 development corpus are selected for training the female NAP matrix. As for the male NAP matrix training, we select the following data: 105 bi-lingual telephone speakers from NIST 2006, 44 speakers from NIST SRE 2005 and 30 speakers from NIST SRE 2006 which are recorded both on the telephone and on the ancillary microphone channel, and the 3 persons in the MIX 5 development corpus.

Because some speakers in NIST SRE 2006 also appeared in this year's evaluation, the utterances of SRE 2006 are not selected as a negative sample in SVM system. We select SRE 2004, 2005 and MIX 5 development data as SVM negative samples. And the SRE 2006 corpus is selected as the Tnorm cohorts. Gender-dependent Tnorm is applied to the scores of SVM. Moreover, about 400 utterances from each gender are selected as Tnorm cohorts.

5. RESULTS

We build four sub-systems in the evaluation, which are
(1) PLP, GMM-UBM, factor analysis, ZTnorm.
(2) LPCC, GMM-UBM, factor analysis, ZTnorm.
(3) PLP, GMM-SVM, NAP, Tnorm.
(4) LPCC, GMM-SVM, NAP, Tnorm.

The equal error rate (EER), minimum detection cost function (minDCF) and DET curves are used to evaluate the system performance. NIST has separated the trials according to their channel type. We list the results according to different training and testing conditions.

Table 1 lists the performance of the GMM-UBM system with factor analysis. The scores of PLP and LPCC systems are fused with equal weights.

Table 1. Fusion of the PLP and LPCC in the GMM-UBM

Training	Testing	EER	minDCF
Interview	Interview	3.3%	0.118
Interview	Telephone	5.1%	0.219
Telephone	Interview	5.0%	0.227
Telephone	Microphone	5.3%	0.200
Telephone	Telephone	5.0%	0.247

Table 2 lists the performance of the GMM-SVM system with NAP. Again, the scores of PLP and LPCC systems are fused with equal weights.

Table 2. Fusion of the PLP and LPCC in the GMM-SVM

Training	Testing	EER	minDCF
Interview	Interview	3.6%	0.149
Interview	Telephone	5.4%	0.247
Telephone	Interview	5.4%	0.216
Telephone	Microphone	5.2%	0.200
Telephone	Telephone	6.5%	0.350

Table 3 lists the fusion of the GMM-SVM and GMM-UBM systems. Because there are PLP and LPCC acoustic features, the scores of the four systems are fused with equal

weights. The final fused system is our primary system for NIST submission. The fusion of GMM-UBM and GMM-SVM can improve the performance for most sub-section trials except for the telephone training-testing trials, for which the GMM-UBM sub-system has much better results than the GMM-SVM sub-system.

Table 3. Fusion of GMM-UBM and GMM-SVM

Training	Testing	EER	minDCF
Interview	Interview	2.8%	0.105
Interview	Telephone	4.3%	0.178
Telephone	Interview	4.3%	0.171
Telephone	Microphone	4.1%	0.170
Telephone	Telephone	5.3%	0.280

The DET curves of the primary system are also depicted in Fig. 2.From the EER, minDCF and DET curves we can see that the interview data and microphone data can get better results than the telephone data, which is different from the results in [10]. Maybe this can be explained by the following two reasons. First, the microphone data and interview data has not passed the transmission channel or audio codec. Second, there are no bi-lingual speakers in these two conditions. Though the SNR of these two conditions is not as good as that of the telephone data, the SNR of most of the utterances is above 15db [10] and that will not affect the performance much in the GMM algorithms.

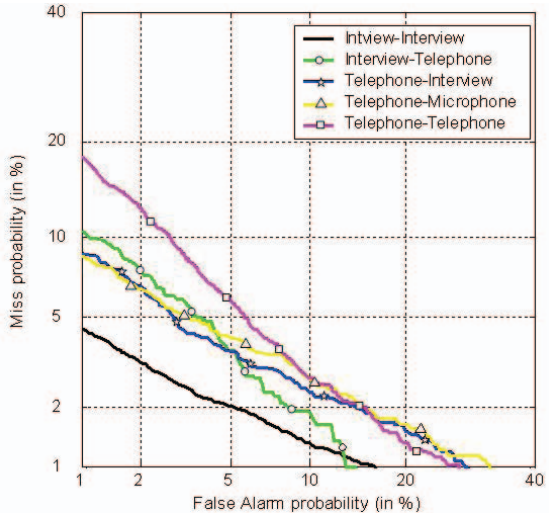

Fig. 2. The DET curves of five sub-sections trials

6. DISCUSSION

NIST provides microphone data besides telephone data, so there are more channel types in this year's evaluation. From the results we can see that as long as the development data matches the evaluation, the factor analysis and NAP algorithms can remove the channel bias effectively. Furthermore, the noise also affects the performance in speaker recognition, but not so much as in the speech recognition.

Bi-lingual problem is one of the important factors that affect the performance. But if we choose data properly in the process of UBM, factor analysis, NAP and score normalization, we can avoid this problem to some extent.

7. REFERENCES

[1] NIST, "The NIST Year 2008 Speaker Recognition Evaluation Plan", [Online],Available:http://www.nist.gov/speech/tests/sre/2008/sre08_evalplan_release4.pdf

[2] P. Kenny, P. Ouellet, N. Dehak, V. Gupta and P. Dumouchel, "A Study of Inter-Speaker Variability in Speaker Verification", [Online], Available: http://www.crim.ca/perso /patrick.kenny/.

[3] C. Vair, D. Colibro, F. Castaldo, E. Dalmasso and P. Laface, "Channel Factors Compensation in Model and Feature Domain for Speaker Recognition", *Proc. IEEE Odyssey 2006*, San Juan, June 2006, pp.1-6.

[4] R. Vogt, S. Sridharan, "Experiments in Session Variability Modeling for Speaker Verification", *Proc. ICASSP 2006*,Toulouse, May 2006, pp. 897-900.

[5] D.A. Reynolds, T.F. Quatieri and R.B. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models", *Digital Signal Processing*, vol. 10, pp. 19–41, 2000.

[6] W. M. Campbell, D. Sturim, and D. A. Reynolds, "SVM Based Speaker Verification Using a GMM Supervector Kernel and NAP Variability Compensation", *Proc. ICASSP 2006*, Toulouse, May 2006, pp. I-97- I-100.

[7] C. Cieri, L. Corson,D Graff, K. Walker, "Resources for New Research Directions in Speaker Recognition:The Mixer 3, 4 and 5 Corpora", *Proc. Interspeech 2007*, Antwerp, Aug. 2007, pp.950-953.

[8] B. Xiang, U. V. Chaudhari, J. Navratil, G.N. Ramaswamy and R.A. Gopinath, "Short-time Gaussianization for Robust Speaker Verification", *Proc. ICASSP 2002*,Orlando, FL, May 2002, pp. I-681- I-684.

[9] L. Lamel, L. Rabiner, A. Rosenberg, J. Wilpon, "An Improved Endpoint Detector for Isolated Word Recognition", *IEEE Trans. on Acoustics, Speech, and Signal Processing*,vol.29, no. 4, pp.777-785,Aug. 1981.

[10] D.E. Sturim, W.M. Campbell, D.A. Reynolds, R.B. Dunn and T.F. Quatieri, "Robust Speaker Recognition with Cross-Channel Data: MIT-LL Results on the 2006 NIST SRE Auxiliary Microphone Task", *Proc. ICASSP 2007*, Honolulu, HI, Apr. 2007, pp. IV-49-IV-52.