The IIR NIST 2006 Speaker Recognition System: Fusion of Acoustic and Tokenization Features

Rong Tong[1,2], Bin Ma[1], Kong-Aik Lee[1], Changhuai You[1], Donglai Zhu[1], Tomi Kinnunen[1], Hanwu Sun[1], Minghui Dong[1], Eng Siong Chng[2] and Haizhou Li[1,2]

[1]Institute for Infocomm Research
21 Heng Mui Keng Terrace, Singapore 119613
tongrong@i2r.a-star.edu.sg
[2]School of Computer Engineering,
Nanyang Technological University, Singapore 639798
asechng@ntu.edu.sg

Abstract. This paper describes our recent efforts in exploring effective discriminative features for speaker recognition. There is an obvious trend in the past few years that the information fusion from multiple resources is critical to improve the performance of speaker recognition system. The extracted information for speaker recognition varies from acoustic features to high dimensional vectors with different levels of tokenization. In the IIR NIST 2006 Speaker Recognition System, we integrated cepstral GMM modeling, cepstral SVM modeling and tokenization at both phone level and frame level. The experimental results on both NIST 2005 SRE corpus and NIST 2006 SRE corpus are presented.

Keywords: speaker recognition, cepstral feature, phonotactic feature, Gaussian mixture model, support vector machine, tokenization, fusion

1 Introduction

Automatic speaker recognition is the task of identifying a speaker corresponding to a given voice. In the past decade, much progress has been made in text-independent speaker recognition by using acoustic features, such as Gaussian Mixture Modeling (GMM) on amplitude spectrum based features [1] and Support Vector Machine (SVM) on Shifted Delta Cepstral (SDC) [2]. In recent years, some tokenization methods with higher level information have attracted great interests. These tokenization methods convert the speech into different levels of tokens, such as words, phones and GMM tokens. For example, lexical features based on word n-grams has been studied in [3] for speaker recognition; Parallel Phone Recognition followed by Language Modeling (PPRLM) [4] has been extensively adopted in language and speaker recognition; Gaussian Mixture Model Tokenization [5], [6] has been used with the tokens at the frame level for language and speaker recognition.

It is generally agreed that the integration with different degrees of discriminative information can improve the performance of speaker recognition system. The

extraction and organization of multiple resource information has become a critical task to the success of speaker recognition.

The acoustic features, such as MFCC features, adopted in speech recognition systems are the natural choices for speaker recognition. The Gaussian mixture modeling (GMM) based on MFCCs has demonstrated a great success for text-independent speaker recognition [1]. To model out-of-set speakers, Universal background model (UBM) is used to normalize the likelihood scores from different speakers. The model of a specific speaker is obtained with Bayesian adaptation based on UBM by using the training data of that speaker [1]. Test normalization (Tnorm) [7] is another good choice to make score normalization by calculating the mean and variance parameters from multiple non-target speaker models. In this paper, a new proposed acoustic feature termed temporal discrete cosine transform (TDCT) feature [8] is also used to capture the long time dynamic information in the GMM framework.

Support vector machines (SVMs) have been a powerful classifier in many pattern classification tasks. An SVM is a discriminative classifier to separate two classes with a hyperplane in a high-dimensional space. In [2], the generalized linear discriminant sequence kernel (GLDS) is proposed for speaker and language recognition. The feature vectors extracted from an utterance are expanded to a high-dimensional space by calculating all the monomials. To simplify the computation, it is assumed that the kernel inner product matrix is diagonal. It also shows that the front-end with linear prediction cepstral coefficients (LPCCs) give better performance than the front-end with MFCCs. We construct two SVM subsystems based on both MFCCs and LPCCs.

In the passed several years, phonotactic features have proven to provide effective complementary cues for speaker and language recognition. The phonotactic features are extracted from an utterance in the form of tokens. The tokens may be at different levels, words, phones and even frames. PPRLM [4] uses multiple parallel phone recognizers to convert the input utterance into phone token sequence. It is followed by a set of phone *n*-gram language models that impose constraints on phone decoding and provide language scores. Instead of *n*-gram phone language models, we proposed to use vector space modeling (VSM) as the backend classifier [9]. For each phone sequence generated from the multiple parallel phone recognizers, we count the occurrences of phone *n*-grams. A phone sequence is then represented as a high-dimensional vector of *n*-gram occurrences. SVM is used as the classifier on the concatenated vectors from multiple phone sequences, named as Bag-of-Sounds (BOS) vectors.

The tokenization can also be made at the frame level, such as Gaussian Mixture Model Tokenization [5] for language identification. It captures another aspect of acoustic and phonetic characteristics among the languages and the speakers, and provides more tokens than the phone recognizers from the limited speech data. Same as PPRLM, multiple parallel GMM tokenizers can be used to improve speaker coverage in speaker recognition. We propose to use speaker cluster based GMM tokenization as one of the subsystems in our speaker recognition system that multiple GMM tokenizers are constructed according to the speaker characteristics.

This paper is organized as follows. In Section 2, we introduce the speech corpora we used, including the training data, development data and the NIST 2006 Speaker Recognition System (SRE) corpus. In Section 3, we describe our six subsystems and the score fusion strategy. In Section 4, we present the experimental results on the

development data (NIST 2005 SRE) as well as on the NIST 2006 SRE data. We make discussions in Section 5.

2 IIR Submission and Speech Corpora

The NIST 2006 SRE evaluation task is divided into 15 distinct and separate tests. Each of these tests involves one of the five training conditions and one of four test conditions [10]. The five training conditions are 10-second speech excerpt from a two-channel/4-wire (10sec4w), one conversation side of approximately five minutes total duration from a two-channel/4-wire (1conv4w), three conversation sides (3conv4w), eight conversation sides (8conv4w) and three conversation sides from a summed-channel/2-wire (3conv2w). The four test conditions are 10-second speech excerpt from two-channel/4 wire (10sec4w), one conversation side from a two-channel/4-wire (1conv4w), one conversation side from a summed-channel/2-wire (1conv2w) and 1 conversation side recorded by auxiliary microphone (1convMic).

The performance of the NIST speaker recognition system is evaluated by the detection cost function. It is defined as a weighted sum of miss and false alarm error probabilities:

$$C_{Det} = C_{Miss} \times P_{Miss|Target} \times P_{Target} + C_{FalseAlarm} \times P_{FalseAlarm|NonTarget} \times (1 - P_{Target}) \quad (1)$$

In NIST 2006 SRE, $C_{Miss} = 10$, $C_{FalseAlarm} = 1$ and $P_{Target} = 0.01$. The experiment results presented in this paper are reported in Equal Error Rate (EER) and DET curves. EER is used to decide the operating point when the false acceptance rate (FAR) and false rejection rate (FRR) are equal.

2.1 IIR Submission for the NIST 2006 SRE

IIR's speaker recognition system participating seven tests that involve 4 training conditions and 2 test conditions: 10sec4w-10sec4w, 1conv4w-10sec4w, 3conv4w-10sec4w, 8conv4w-10sec4w, 1conv4w-1conv4w, 3conv4w-1conv4w and 8conv4w-1conv4w. Table 1 shows the total 15 tests of the NIST 2006 SRE and the 7 tests that IIR participated.

There are six subsystems in the IIR speaker recognition system for the NIST 2006 SRE. These subsystems fall into three categories: (i) spectral features with SVM modeling including MFCC feature based spectral SVM (Spectral MFCC-SVM) and LPCC feature based spectral SVM (Spectral LPCC-SVM); (ii) spectral feature with GMM modeling including MFCC feature based GMM (MFCC-GMM) and TDCT feature [8] based GMM (TDCT-GMM); (iii) tokenization features with vector space modeling (VSM) including parallel phone recognizers based tokenization: Bag-of-Sounds (BOS) and speaker clustering based multiple GMM tokenizers (GMM token). The first four subsystems capture the characteristics of spectral features while the last

two tokenization subsystems capture the phonotactic information. Fig. 1 shows the system framework of IIR submission. The score fusion is conducted with the six subsystems to make the final decision.

Table 1. The seven tests that IIR participated in the NIST 2006 SRE

		Test segment condition			
		10sec 2-chan	**1conv 2-chan**	**1conv summed chan**	**1 conv aux mic**
Training condition	**10sec 2-chan**	10sec4w-10sec4w			
	1conv 2-chan	1conv4w-10sec4w	1conv4w-1conv4w	N.A.	N.A.
	3conv 2-chan	3conv4w-10sec4w	3conv4w-1conv4w	N.A.	N.A.
	8conv 2-chan	8conv4w-10sec4w	8conv4w-1conv4w	N.A.	N.A.
	3conv summed-chan		N.A.	N.A.	

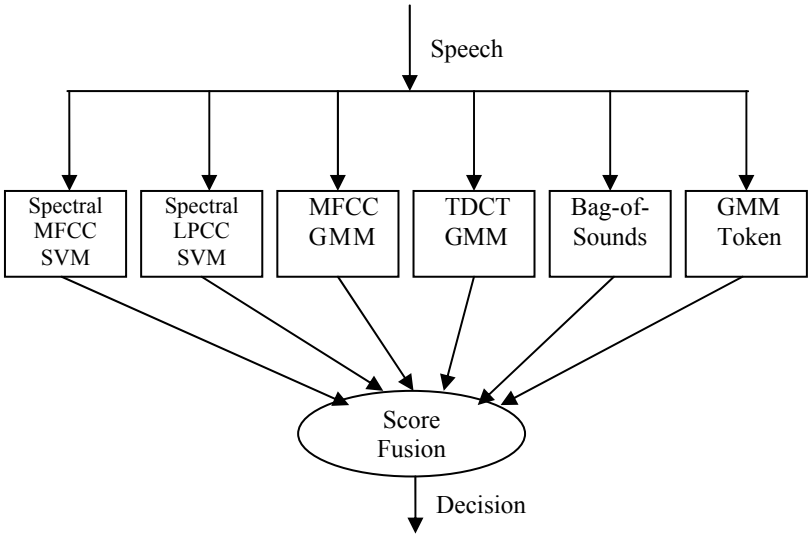


Fig. 1. System framework of IIR submission for the NIST 2006 SRE

2.2 Speech Corpora

Table 2 shows the training and development data for each subsystem. The tokenizer training data are used to model the parallel phone recognizers or to model the parallel

GMM tokeniers. Background speaker data are used to train UBM models or to train speaker background models. Cohort data are used to make the Test normalization (Tnorm). The NIST 2005 SRE corpus is used to evaluate the performance of individual systems. The output scores of the six subsystems are used to train the score fusion to facilitate the final decisions on the NIST 2006 SRE data. We will describe each of these speech corpora in next section together with the subsystems.

Table 2. Speech corpora for the training and development

	Spectral MFCC SVM	**Spectral LPCC SVM**	**MFCC GMM**	**TDCT GMM**	**Bag-of-Sounds**	**GMM Token**
Tokenizer training data	N.A.				1. IIR-LID corpus 2. LDC Korean corpus 3. MAT corpus 4. OGI-TS corpus	NIST 2002 SRE corpus
Background speaker data (UBM)	Switchboard corpora: sw3p1, sw3p2, sw2p2 and sw2p3		NIST2004 1side training files	NIST2004 1side training files	NIST 2002 SRE corpus	NIST 2004 SRE corpus
Cohort data (Tnorm)	Evaluation set of the NIST 2004 SRE corpus		N.A.			
Development /Test	NIST 2005 SRE corpus					

3 System Description

For the spectral SVM and GMM subsystems, an energy-based voice activity detector (VAD) is applied after feature extraction to remove non-speech frames. We train two GMMs with 64 mixture components to model the energy distributions of the speech frames as well as the non-speech frames by using the development set of the NIST 2001 SRE corpus. With such a VAD algorithm, about 38% speech and 62% non-speech frames were detected in the NIST 2006 SRE corpus.

For the Bag-of-Sounds and GMM token subsystems, the VAD algorithm chunks the long conversations into smaller utterances so that the tokenization methods can be applied to create phone or GMM token sequence for each of the utterances. The utterance based cepstral mean subtraction is performed to filter off the channel distortion.

3.1 Spectral LPCC-SVM and Spectral MFCC-SVM subsystems

Support vector machine (SVM) is a two-class classifier. In speaker recognition, it can be used to model the boundary between a speaker and a set of background speakers. The background speakers represent the population of imposters expected during recognition. We follow the work reported in [2] and [11] in which generalized linear

discriminant sequence kernel (GLDS) is proposed for speaker and language recognition.

Two kinds of acoustic spectral features, MFCC features and LPCC features, both with a dimension of 36, are used in the two SVM subsystems. For the MFCC front-end, we use a 27-channel filterbank, and 12MFCC + 12Δ + 12ΔΔ coefficients. For the LPCC front-end, 18LPCC + 18Δ coefficients are used.

The feature vectors extracted from an utterance is expanded to a higher dimensional space by calculating all the monomials up to order 3, resulting in a feature space expansion from 36 to 9139 in dimension. The expanded features are then averaged to form an average expanded feature vector for each of the utterances under consideration. In the implementation, it is also assumed that the kernel inner product matrix is diagonal for computational simplicity.

During enrollment, the current speaker under training is labeled as class +1, whereas a value of -1 is used for the background speakers. The set of background speaker data is selected from Switchboard 3 Phase 1 and Phase 2 (for Cellular data) and Switchboard 2 Phase 2 and Phase 3 (for landline telephone). We randomly select 2000 utterances from each of the 4 datasets to form the background speaker database of 8000 utterances, with roughly equal amounts of male and female speakers. Each utterance in the background and the utterance of the current speaker under training is represented with an average expanded feature vector b_{av}. These average expanded features are used in the SVM training. The commonly available toolkit SVMTorch [13] is used for this purpose. The result of the training is a vector w of dimension 9139 which represents the desired target speaker model [11]. During evaluation, an average expanded feature vector b_{av} is formed for each of the input utterances, and the score is taken as the inner product between these two vectors, i.e, $w^T b_{av}$.

Test normalization (Tnorm) method [7] is adopted in the two subsystems. The NIST 2004 training data is used to form the cohort models. In particular, the speaker models in the NIST 2004 are used as the cohort models. The training condition of the cohort models and evaluation corpus are matched. For example, the trained models in the 1side of NIST 2004 are used as the cohort models for the target models in the 1conv4w training condition of the NIST 2005 and 2006 SRE corpus. Similar concept is applied to 10sec4w, 3conv4w, and 8conv4w training conditions.

3.2 MFCC-GMM and TDCT-GMM subsystems

Two kinds of spectral features are used in the two GMM modeling subsystems. One is the MFCC features same as those adopted for spectral SVM subsystem. Another use the temporal discrete cosine transform (TDCT) features [8].

Conventional MFCC features characterize the spectral character in a short-time frame of speech (typically 20~30 ms). Psychoacoustic studies [13] suggest that the peripheral auditory system in humans integrates information from much larger time spans than the temporal duration of the frame used in speech analysis. Inspired by this finding, the TDCT feature is aiming at capturing the long time dynamic of the spectral features. Fig. 2 illustrates the TDCT feature computation procedure. Each cepstral coefficient is considered as an independent signal which is windowed in blocks of

length B. Discrete cosine transform (DCT) is applied on each block, and the lowest L DCT coefficients, which contain most of the energy, are retained. Suppose we have M coefficients in the MFCC feature vector, the DCT coefficients can be stacked to form a long vector with the dimensionality of $M \times L$. The next TDCT vector is computed by advancing the block by one frame. Experimental results show that a block size of B = 8, and L = 3 for the DCT, give the best performance on the NIST 2001 SRE dataset [8]. The resulting TDCT feature vector has a dimension of 36×3 = 108, and corresponds to a total time span of 250ms.

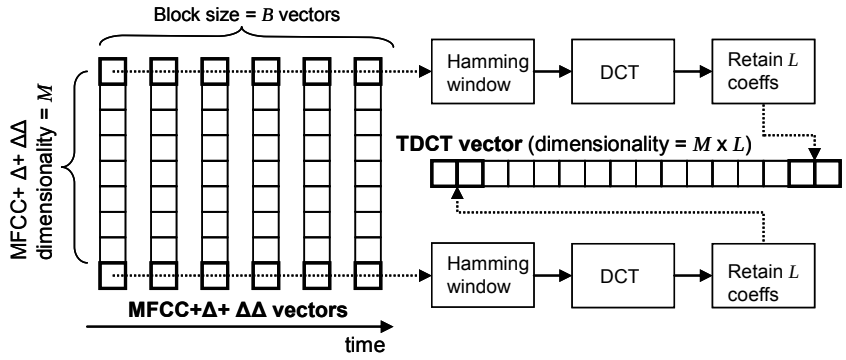


Fig. 2. Illustration of the TDCT feature computation

For both the two GMM subsystems, the gender-dependent background models with 256 Gaussian mixtures are trained by using the NIST2004 1-side training data subset. The background model having the same gender with the target is used for adaptation. In the evaluation, the background model of the same gender with the target speaker is used to give the background score.

3.3 Bag-of-Sounds subsystem

This approach uses parallel phone tokenizers as the front end, and vector space modeling as the back end classifier [9].

Seven phone tokenizers are constructed, including English, Korean, Mandarin, Japanese, Hindi, Spanish and German. English phone recognizer is trained from IIR-LID database [14]. Korean phone recognizer is trained from LDC Korean corpus (LDC2003S03). Mandarin phone recognizer is trained from MAT corpus [15]. Other four phone recognizers are trained from OGI-TS corpus [16]. Each phone is modeled with a three-state HMM, and 39-dimensional MFCC features are used. Each HMM state in English, Korean and Mandarin are modeled with 32 Gaussian mixtures, while the states in other languages are with 6 Gaussian mixtures considering the availability of training data. Phone recognition is performed with the Viterbi search using a fully connected null-grammar network of phones.

For a given speech utterance, the tokenizers yield seven phone sequences. They are converted to a vector of weighted terms in three steps. Firstly, we compute unigram and bigram probabilities for each phone sequence, and then organize the probabilities

into a vector. Secondly, each entry in the vector is multiplied by a background component [17]. Finally, we concatenate the seven vectors to form a long vector.

In the training process of the SVM training, a single vector of weighted probabilities is derived from each conversation side. We use a one-versus-all strategy to train the SVM model for a given speaker. The conversation side of the target speaker is labeled as class +1, while all the conversation sides in the background are labeled as class -1. NIST 2002 SRE corpus is used as background data. During the evaluation, the input utterance is converted to the long vector and a score is produced from the SVM model. The toolkit SVMTorch [12] with a linear kernel is used.

3.4 GMM Token subsystem

This approach uses multiple GMM tokenizers as the front end, and vector space modeling as the back end classifier [6]. Each GMM tokenizer converts the input speech into a sequence of GMM token symbols which are indexes of the Gaussian components scoring highest at every frame in the GMM computation. The GMM token sequences are then processed in the same way as the process of phone sequences in the bag-of-sounds approach, i.e., the sequences are converted to a vector of weighted terms and then recognized by a speaker's SVM model.

Inspired by the finding of PPRLM in language recognition where multiple parallel single-language phone recognizers in the front-end enhance the language coverage and improve the language recognition accuracy over single phone recognizer, we explore multiple GMM tokenizers to improve speaker characteristics coverage and to provide more discriminative information for speaker recognition [6]. By clustering all the speakers in the training set into several speaker clusters, we represent the training space in several partitions. Each partition of speech data can then be used to train a GMM tokenizer. With each of these parallel GMM tokenizers, a speech segment is converted to the feature vector of weighted terms. The multiple feature vectors are then concatenated to form a composite vector for SVM modeling. We use the NIST 2002 SRE corpus for the training of speaker cluster based GMM tokenizers, and use the NIST 2004 SRE corpus as the background data. 10 parallel GMM tokenizers, each having 128 mixtures of Gaussian components, are constructed.

3.5 Score fusion of subsystems

The six subsystems described above are combined together. We use SVM classifiers again for the final decision as shown in Fig. 3. For a given speech utterance and the reference speaker, a 6-dimensional score vector is derived from the six subsystems. The score vectors are first normalized to zero mean and unit variance. Then the polynomial expansion of order 1, 2 and 3 are applied to the normalized score vectors. Three sets of expanded score vectors with dimension 7, 28 and 84 are obtained. Each set of the expanded score vectors are used to train a SVM model. The final decision is made according to the averaged value of three SVM scores.

The NIST 2005 SRE evaluation corpus is used as the training data for these three SVMs. The score vectors generated from the genuine utterances are labeled as class

+1, and the score vectors generated from the impostor utterances are labeled as class -1. The thresholds estimated from the NIST 2005 SRE corpus are used for final True/False decision on the NIST 2006 SRE. The toolkit SVMTorch [12] with a radial kernel is used.

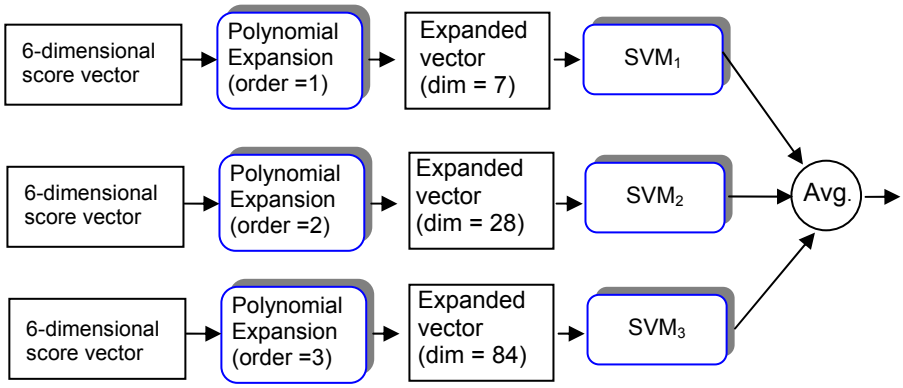


Fig. 3. Score fusion of the six subsystems

4 Experiment Results

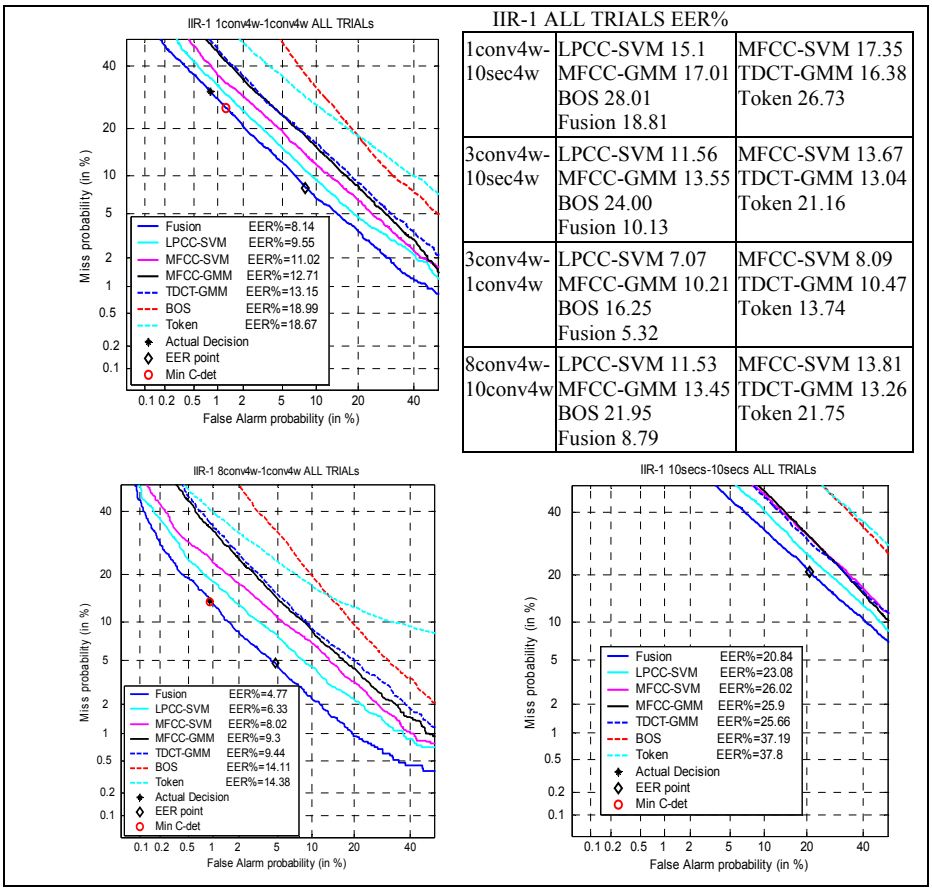
The NIST 2005 SRE evaluation set is used to evaluate the performance of the six subsystems before the whole system is submitted to the NIST 2006 SRE. It is also used as the development set to estimate the thresholds of the score fusion which provides the genuine/impostor decision for all the trials in the NIST 2006 SRE. Table 3 shows the equal error rates (EER%) of the six subsystems as well as the score fusion on seven test conditions in the NIST 2005 SRE.

Table 3. EER% of subsystems and fusion on the NIST 2005 SRE evaluation set

Test / System	LPCC SVM	MFCC SVM	MFCC GMM	TDCT GMM	BOS	GMM Token	Fusion
10sec4w-10sec4w	29.41	31.28	28.72	30.39	41.80	40.35	24.62
1conv4w-10sec4w	18.74	19.92	19.78	18.76	28.96	31.05	13.80
1conv4w-1conv4w	10.55	11.32	13.55	13.81	19.31	22.38	7.82
3conv4w-10sec4w	14.40	16.02	16.16	15.60	24.93	25.26	11.32
3conv4w-1conv4w	6.87	8.07	10.26	9.97	14.32	16.11	5.67
8conv4w-10sec4w	13.05	14.00	14.54	14.45	22.29	24.34	9.76
8conv4w-1conv4w	5.73	7.17	9.42	9.11	12.22	17.27	4.56

Among the six subsystems, four acoustic feature based subsystems outperform the two tokenization subsystems and spectral SVM method with LPCC features gives the best performance in all the seven test conditions. The score fusion combines the scores from both the acoustic feature modules and the tokenization feature modules. It improves the overall accuracies significantly.

With the six subsystems and the thresholds of the score fusion obtained from the NIST 2005 SRE corpus, we can now process the NIST 2006 SRE data. Fig. 4 shows the performance of the seven test conditions of the NIST 2006 SRE. We show both the DET curves and the EER% for the three test conditions, 10sec4w-10sec4w, 1conv4w-1conv4w and 8conv4w-1conv4w. The other four test conditions have only EER% included. In the DET curves, the points of Min C-det denote the best results we can achieve from all possible thresholds for the final decision. The points of Actual Decision denote the results on our actual designed thresholds.



IIR-1 ALL TRIALS EER%

1conv4w-10sec4w	LPCC-SVM 15.1 MFCC-GMM 17.01 BOS 28.01 Fusion 18.81	MFCC-SVM 17.35 TDCT-GMM 16.38 Token 26.73
3conv4w-10sec4w	LPCC-SVM 11.56 MFCC-GMM 13.55 BOS 24.00 Fusion 10.13	MFCC-SVM 13.67 TDCT-GMM 13.04 Token 21.16
3conv4w-1conv4w	LPCC-SVM 7.07 MFCC-GMM 10.21 BOS 16.25 Fusion 5.32	MFCC-SVM 8.09 TDCT-GMM 10.47 Token 13.74
8conv4w-10conv4w	LPCC-SVM 11.53 MFCC-GMM 13.45 BOS 21.95 Fusion 8.79	MFCC-SVM 13.81 TDCT-GMM 13.26 Token 21.75

Fig. 4. Performance on the NIST 2006 SRE

To study the contribution of each subsystem category to the final fusion, we use spectral LPCC-SVM subsystem as the baseline. Three other subsystems, MFCC-SVM, MFCC-GMM and Bag-of-Sounds (BOS), will be combined with LPCC-SVM subsystem individually. The experiments are conducted on three test conditions of NIST 2006 SRE, 10sec4w-10sec4w, 1conv-1conv4w and 8conv4w-1conv4w that involve three training segment conditions and two test segment conditions. Table 4 shows the results. The numbers in the bracket are relative EER reduction compared with the baseline system, the spectral SVM with LPCC features only.

Since more information is provided, the combinations generally give us better performance. For the short test segment (10sec4w-10sec4w), the MFCC-GMM subsystem brings the larger error reduction. Although both MFCC-GMM and LPCC-SVM use acoustic features, they model the spectral features with different method and can make good use of more discriminative information. Bag-of-Sounds subsystem uses phonotactic features that provide complementary information to acoustic features for the speaker recognition task. A relative EER reduction of 22.6% has been achieved based on the LPCC-SVM subsystem on the 8conv4w-1conv4w. The combination of LPCC and MFCC features with SVM method also produce better results in all the three test conditions.

Table 4. Performances (EER%) of different subsystem combinations on the NIST 2006 SRE

Test/ System	**LPCC-SVM**	**LPCC-SVM MFCC-SVM**	**LPCC-SVM MFCC-GMM**	**LPCC-SVM BOS**
10sec4w-10sec4w	23.08	22.94 (0.6%)	21.27 (7.8%)	23.47 (-1.7%)
1conv4w-1conv4w	9.55	8.72 (8.7%)	8.44 (11.6%)	8.78 (8.1%)
8conv4w-1conv4w	6.33	5.18 (18.2%)	5.07 (19.9%)	4.90 (22.6%)

5 Summary and Discussion

We present the IIR speaker recognition system for NIST 2006 SRE. The system consists of six subsystems that capture both acoustic features and phonotactic information. For the acoustic features, both GMM modeling and spectral SVM modeling are adopted. Besides the conventional features, such as MFCCs and LPCCs, we propose to use TDCT features to model the long time dynamic of the spectral information. To capture speaker discriminative information from the higher level, tokenization methods are used to create phone token sequence and GMM token sequence from each of the utterances. For a given utterance, all the n-gram probabilities of the token sequence are calculated and combined into an n-gram statistic vector. A high dimensional vector is obtained by concatenating multiple token sequences obtained from parallel phone recognizers or parallel GMM tokeniziers. Vector space modeling method is adopted as the backend classifier to model these high dimensional vectors.

The experimental results show that the acoustic features are more effective in speaker recognition. The phonotactic features also provide complementary information and can improve the system performance significantly on longer speech segments. The experiment results on the subsystem fusion proved that the combination of the discriminative features from multiple sources is an effective method to improve the speaker recognition accuracy.

References

1. Reynolds, D. A., Quatieri, T. F. and Dunn, R. B.: Speaker Verification Using Adapted Gaussian Mixture Modeling. Digital Signal Processing, 10 (2000), pp. 19-41.
2. Campbell, W. M., Campbell, J. P., Reynolds, D. A., Singer, E. and Torres-Carrasquillo, P. A.: Support Vector Machines for Speaker and Language Recognition. Computer Speech and Language, 20 (2006), pp. 210-229.
3. Doddington, G.: Speaker Recognition based on Idiolectal Differences between Speakers. Proc. Eurospeech, 2001.
4. Zissman, M. A.: Comparison of Four Approaches to Automatic Language Identification of Telephone Speech. IEEE Trans. on Speech and Audio Processing, vol. 4, no. 1, 1996.
5. Torres-Carrasquillo, P. A., Reynolds, D. A. and Deller, Jr., J. R.: Language Identification using Gaussian Mixture Model Tokenization. Proc. ICASSP, 2002.
6. Ma, B., Zhu, D., Tong, R. and Li, H.: Speaker cluster based GMM tokenization for speaker recognition. To appear in Interspeech 2006.
7. Auckenthaler, R., Carey, M. and Lloyd-Thomas, H.: Score normalization for text-independent speaker verification systems. Digital Signal Processing, vol. 10, no 1-3, pp. 42-54, Jan 2000.
8. T. H. Kinnunen, C. W. E. Koh, L. Wang, H. Li and E. S. Chng, "Shifted delta cepstrum amd temporal discrete cosine transform features in speaker verification," submitted to 5th International Symposium on Chinese Spoken Language Processing, 2006.
9. Li, H. and Ma, B.: A Phonotactic Language Model for Spoken Language Identification", 43rd Annual Meeting of the Association for Computational Linguistics (ACL05), June 2005, Ann Arbor, USA.
10. http://www.nist.gov/speech/tests/spk/2006/sre-06_evalplan-v9.pdf
11. W.M. Campbell, "Generalized linear discrininant sequence kernels for speaker recognition," in Proc. ICASSP, pp. 161-164, 2002
12. Collobert, R. and Bengio, S.: SVMTorch: support vector machines for large-scale regression problems. Journal of Machine Learning Research, vol. 1, pp. 143-160, 2001.
13. H. Hermansky, "Exploring temporal domain for robustness in speech recognition," invited paper. Proceedings of the 15th International Congress on Acoustics, 3:61-64, 1995.
14. Language Identification Corpus of the Institute for Infocomm Research
15. Wang, H.-C.: MAT-a project to collect Mandarin speech data through networks in Taiwan. Int. J. Comput. Linguistics Chinese Language Process. 1 (2) (February 1997) 73-89.
16. http://cslu.cse.ogi.edu/corpora/corpCurrent.html
17. Campbell, W. M., Campbell, J. P., Reynolds, D. A., Jones, D. A. and Leek, T. R.: Phonetic speaker recognition with support vector machines. Proc NIPS, 2003