The IIR Submission to CSLP 2006 Speaker Recognition Evaluation

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

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

Abstract. This paper describes the design and implementation of a practical automatic speaker recognition system for the CSLP speaker recognition evaluation (SRE). The speaker recognition system is built upon four subsystems using speaker information from acoustic spectral features. In addition to the conventional spectral features, a novel *temporal discrete cosine transform* (TDCT) feature is introduced in order to capture long-term speech dynamic. The speaker information is modeled using two complementary speaker modeling techniques, namely, Gaussian mixture model (GMM) and support vector machine (SVM). The resulting subsystems are then integrated at the score level through a multilayer perceptron (MLP) neural network. Evaluation results confirm that the feature selection, classifier design, and fusion strategy are successful, giving rise to an effective speaker recognition system.

1 Introduction

Speaker recognition is the process of automatically establishing personal identity information by analyzing speech utterances [1]. The goal of speaker recognition is to identify people by voice. This paper describes and evaluates an automatic speaker recognition system that addresses two different tasks, namely, speaker verification and speaker identification. Speaker verification is the task of validating a claimed identity, whereas speaker identification refers to the task of determining who is speaking [1, 2]. Speaker recognition technology has been found important in various applications, such as, public security, anti-terrorism, justice, and telephone banking.

As part of the 5th *International Symposium on Chinese Spoken Language Processing* (ISCSLP 2006), a special session on speaker recognition is organized by the *Chinese Corpus Consortium* (CCC) [3]. The CSLP speaker recognition evaluation (SRE) aims to provide a common platform for researchers to evaluate their speaker recognition systems. The focus of the CSLP SRE is on Chinese speech, as opposed to some other well-known SRE events, e.g., those carried out by *National Institute of Standards and Technology* (NIST) [4], which focus on English speech. The CSLP

Q. Huo et al. (Eds.): ISCSLP 2006, LNAI 4274, pp. 494 – 505, 2006.

2006 SRE includes text-dependent and text-independent speaker recognition tasks under single-channel and cross-channel training-testing conditions. In this paper we focus on the text-independent speaker verification and identification tasks.

The development and evaluation sets provided for the text-independent tasks of the CSLP 2006 SRE are derived from the CCC-VPR2C2005-1000 corpus (*CCC 2-channel corpus for voiceprint recognition 2005–1000 speakers*) [3]. The development set contains telephone speech utterances from 300 male speakers, while the evaluation set involves 700 male speakers. The speakers in the two datasets do not overlap. In both datasets, the duration of training samples is guaranteed to be approximately longer than 30 seconds, however, the test segments are much shorter.

Table 1. CSLP 2006 SRE evaluation categories

		Text independent	Text-dependent
Speaker verification	Single channel	×	
	Cross channel	×	
Speaker identification	Single channel	×	
	Cross channel	×	

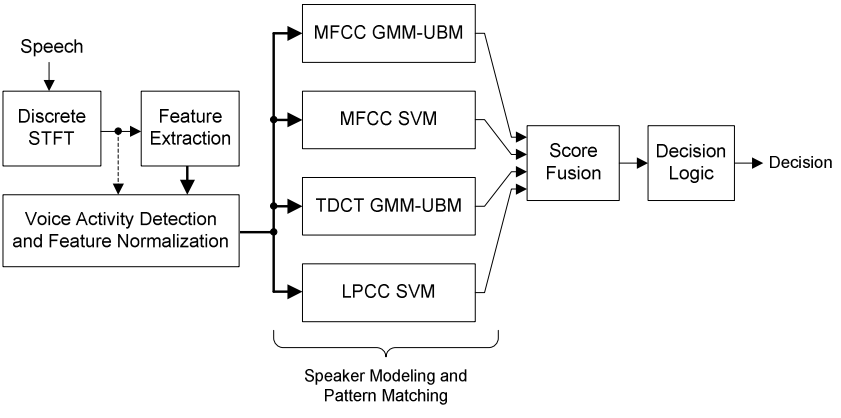

Fig. 1. An automatic speaker recognition system built upon four subsystems. Three different features (MFCC, LPCC, and TDCT) and two different speaker modeling techniques (SVM and GMM) are employed in the subsystems.

This paper describes the design and implementation of a practical automatic speaker recognition system for the CSLP 2006 SRE. The *Speech and Dialogue Processing Group* of *Institute for Infocomm Research* (IIR) participates in four (see checked boxes in Table 1) out of the six evaluation categories (see shaded boxes in Table 1) of this year SRE event. Our submission is built upon four subsystems using speaker information from acoustic spectral features [2, 5, 6, 7], as illustrated in Fig. 1. The speaker information represented in various forms is modeled using Gaussian

mixture model (GMM) [7, 8] and support vector machine (SVM) [9, 10]. Feature extraction and speaker modeling techniques employed in the subsystems are described in Section 2 and Section 3, respectively. The specifications of the subsystems, together with the system integration issue, are then detailed in Section 4. In Section 5, the evaluation results are presented. Finally, Section 6 concludes the paper.

2 Feature Extraction

As the front-end of the automatic speaker recognition system, the function of the feature extraction is to parameterize an input speech signal into a sequence of feature vectors [2]. The purpose of such transformation is to obtain a new representation of the speech signal, which is more compact and allows a tractable statistical modeling. Our speaker recognition system uses two basic sets of acoustic spectral features, namely, the *mel-frequency cepstral coefficients* (MFCC) and the *linear prediction cepstral coefficients* (LPCC) [2, 5, 7]. A third set of features is derived from the MFCC features by taking the discrete cosine transform (DCT) along the time axis, hence the name *temporal DCT* (TDCT) features [6].

2.1 Mel-Frequency Cepstral Coefficients

Prior to feature extraction, the input speech signal is pre-emphasized using a first order finite impulse response filter (FIR) with its zero located at $z = 0.97$. The pre-emphasis filter enhances the high frequencies of the speech signal, which are generally reduced by the speech production process [7].

MFCC feature extraction begins by applying a discrete short-time Fourier transform (STFT) on the pre-emphasized speech signal, using a 30 ms Hamming window with 10 ms overlap between frames. The magnitude spectrum of each speech frame, in the frequency range of 0 to 4000 Hz, is then weighted by a set of 27 mel-scale filters [5]. The mel-scale filter bank emulates the critical band filters of human hearing mechanism. Finally, a 27-point DCT is applied on the log energy of the mel-scale filter bank outputs giving rise to 27 cepstral coefficients. The first coefficient is discarded, and the subsequent 12 coefficients are taken to form a cepstral vector. Delta and delta-delta features are computed over a ± 1 frame span and appended to the cepstral vector, forming a 36-dimensional MFCC feature vector. The delta and delta-delta features contain the dynamic information about the way the cepstral features vary in time.

2.2 Linear Prediction Cepstral Coefficients

In addition to the MFCC feature, the input speech signal is also parameterized in terms of LPCC, which we believe is able to provide complementary information to the MFCC features. Similar to that of the MFCC feature, the LPCC feature is extracted from the pre-emphasized speech signal using a 30 ms Hamming window with 10 ms overlap between frames. For each of the speech frame, an 18th order linear prediction analysis is performed using the autocorrelation method. Finally, 18 cepstral coefficients are derived from the LP coefficients. Dynamic information of the

features is added by appending delta features, resulting in a 36-dimensional LPCC feature vector. Note that we do not include delta-delta features. Preliminary experiment on the NIST 2001 SRE dataset shows that a better performance can be achieved with the current setting.

2.3 Temporal Discrete Cosine Transform

In MFCC features, the delta and delta-delta features capture short-term dynamic information in the interval ranging from 50 to 100 ms. However, this interval is insufficient for longer term "high-level" features like prosodic gestures, and syllable usage. TDCT encodes the long-term dynamic of the cepstral features by taking the DCT over several frames [6]. Fig. 2 illustrates the TDCT features computation procedure. Each cepstral coefficient is considered as an independent signal which is windowed in blocks of length B. 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 of dimensionality $M \times L$. The next TDCT vector is computed by advancing the block by one frame. Experimental results show that a block size of B = 8 frames, and L = 3 for the DCT, give the best performance on the NIST 2001 SRE dataset [6]. The resulting TDCT feature vector has a dimension of 36×3 = 108, and corresponds to a total time span of 250ms.

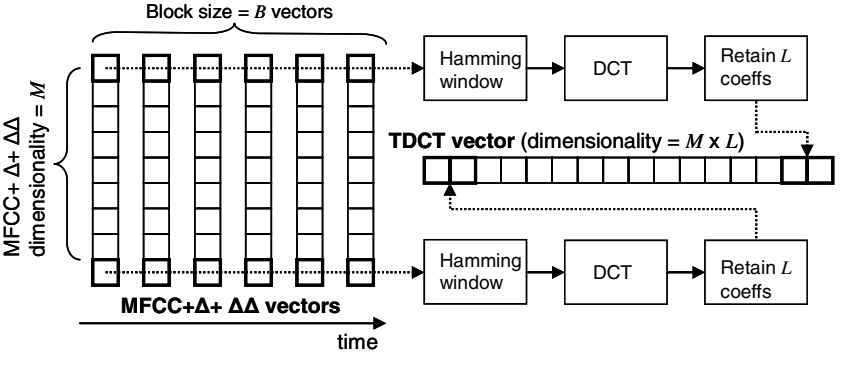

Fig. 2. Illustration of the TDCT features computation [6]

2.4 Voice Activity Detection

An energy-based voice activity detector (VAD) is applied after feature extraction. The VAD decides which feature vectors correspond to speech portions of the signal and which correspond to non-speech portions (i.e., silence and background noise). In particular, we use a GMM with 64 components to model the energy distribution of the speech frames pertaining to each of the two classes. The GMMs are trained beforehand using the development set of the NIST 2001 SRE corpus. The decision is then made through a likelihood ratio test, whereby speech frames with their energy having a higher likelihood with the speech GMM are retained, while those having a

higher likelihood with the non-speech GMM are discarded. Recall that a TDCT feature vector is derived from a block of *B* MFCC feature vectors. If most of the MFCC facture vectors in a certain block belong to speech portion, then the TDCT feature vector derived from that specific block can be determined to be corresponding to speech portion. In our implementation, a TDCT feature vector is retained if more than 40% of the MFCC feature vectors in the block belong to speech portion. Finally, mean subtraction and variance normalization are applied to the outputs of the VAD to produce zero mean, unit variance MFCC, LPCC, and TDCT features.

3 Speaker Modeling and Pattern Matching

Given a speech utterance represented in terms of spectral feature vectors, as described in the previous section, the next step is to model the speaker specific information embedded in the given set of feature vectors. Two different approaches to speaker modeling and verification, as listed below, are employed in our system.

3.1 GMM-UBM

The GMM-UBM subsystems in Fig. 1 uses the standard set-up described in [7, 8]. A GMM is a weighted combination of a finite number of Gaussian distributions in the following form

$$p(\mathbf{x} \mid \lambda) = \sum_{k=1}^{K} w_k \, p_k(\mathbf{x}), \tag{1}$$

where w_k is the mixture weight associated with the *k*th Gaussian component given by

$$p_k(\mathbf{x}) = \frac{1}{(2\pi)^{D/2} |\boldsymbol{\Sigma}_k|^{1/2}} \exp\left\{ -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\mathbf{x} - \boldsymbol{\mu}_k) \right\}. \tag{2}$$

In the above equations, each of the Gaussian densities is parameterized by a $D \times 1$ mean vector $\boldsymbol{\mu}_k$ and a $D \times D$ covariance matrix $\boldsymbol{\Sigma}_k$, where D is the dimension of the feature vector $\mathbf{x}$. The mixture weights of all the K mixture components are by definition ≥ 0 and have to satisfy the constraint $\sum_{k=1}^{K} w_k = 1$. Collectively, the parameters of the mixture density, i.e., $\lambda = \{w_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}$ for $k = 1, 2, \ldots, K$, represent a speaker model in the feature space of $\mathbf{x}$.

For a given test segment $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$, the average log likelihood of the speaker model λ for the test segment, assuming that the feature vectors $\mathbf{x}_n$ are independent, is given by

$$\log p(X \mid \lambda) = \frac{1}{N} \sum_{n=1}^{N} \log p(\mathbf{x}_n \mid \lambda). \tag{3}$$

Notice that log-likelihood value is divided by *N,* which essentially normalizes out the duration effects of test segments with different length. The final score is then taken as a log likelihood ratio, as follows

$$\log \text{ LR} = \log p(X \mid \lambda) - \log p(X \mid \lambda_{\text{UBM}}), \tag{4}$$

where λ_{UBM} is the *universal background model* (UBM) that represents a background set of speaker models. For computational simplicity, we use fast GMM-UBM scoring algorithm [8] using only the top 20 mixture components. It should be emphasized that the fast scoring algorithm makes sense only if the target model is adapted from the background model, as explained below.

In the training phase, speech segments from several background speakers are combined to train a UBM, thereby allowing the UBM to represent the speaker-independent distribution of features. The parameters of the UBM λ_{UBM} are estimated by maximum likelihood estimation, using the *expectation-maximization* (EM) algorithm. A speaker model λ is then derived by adapting the parameters of the UBM λ_{UBM} using the speech segment from the speaker by means of maximum *a posteriori* (MAP) training [8]. For numerical reasons, the covariance matrices pertaining to the Gaussian components are assumed to be diagonal.

3.2 Spectral SVM

SVM is a two-class classifier. For a given set of training samples with positive and negative labels, the SVM models the hyperplane that separates the two classes of samples. In the context of speaker verification, SVM models the boundary between a speaker and a set of background speakers that represent the population of impostors expected during recognition. The idea is different from the GMM-UBM, which models the distribution of the two classes. Furthermore, SVMs are non-probabilistic and use a different training philosophy compared to GMM. With a proper fusion strategy, both classifiers would complement each other in speaker recognition task [10].

The spectral SVM classifier in Fig. 1 closely follows the work reported in [9, 10], which greatly relies on polynomial expansion and the *generalized linear discriminant sequence* (GLDS) kernel. The central element of the GLDS kernel is a kernel inner product matrix defined as follows

$$\mathbf{R} \equiv E\left\{\mathbf{b}(\mathbf{x})\mathbf{b}^T(\mathbf{x})\right\}, \tag{5}$$

where $\mathbf{b}(\mathbf{x})$ denotes the polynomial expansion of the feature vector $\mathbf{x}$. For example, the second-order polynomial expansion of a two-dimensional vector $\mathbf{x} \equiv [x_1,\ x_2]^T$ is given by $\mathbf{b}(\mathbf{x}) = [1,\ x_1,\ x_2,\ x_1^2,\ x_1x_2,\ x_2^2]^T$. For computational simplicity, it is customary to assumed that the matrix $\mathbf{R}$ is diagonal, i.e., $\mathbf{R} \approx diag[\mathbf{r}] = \mathbf{\Lambda}$, where the vector $\mathbf{r}$ is given by

$$\mathbf{r} = \frac{1}{M}\sum_{m=1}^{M} diag[\mathbf{b}(\mathbf{x}_m)\mathbf{b}^T(\mathbf{x}_m)]. \tag{6}$$

In the above equation, $\{\mathbf{x}_m\}_{m=1}^{M}$ denotes a pool of M feature vectors from all the non-target background speakers, and $diag[.]$ denotes the operation forming a diagonal matrix from a column vector and vice versa.

During enrollment, all the utterances in the background and the utterance for the current speaker under training are represented in terms of average expanded feature vectors in the following form

$$\mathbf{b}_{av} = \left[\frac{1}{N} \sum_{n=1}^{N} \mathbf{b}(\mathbf{x}_n) \right], \tag{7}$$

where N denotes the length of any specific utterance. These average expanded feature vectors are then normalized in the form $\mathbf{\Lambda}^{-1/2}\mathbf{b}_{av}$, assigned with appropriate label (i.e., +1 for target speaker, -1 for other competing speakers in the background), and finally used for SVM training. The output of the training is a set of support vectors $\mathbf{b}_i$, weights α_i, and a bias d. A speaker model $\mathbf{w}$ is then obtained by collapsing all the support vectors, as follows

$$\mathbf{w} = \left(\mathbf{\Lambda}^{-1/2} \sum_{i=1}^{l} \alpha_i t_i \mathbf{b}_i \right) + \mathbf{d}, \tag{8}$$

where $\mathbf{d} = [d,\ 0,\ \ldots,\ 0]^T$ and l denotes the number of support vectors resulted from the discriminative training. In the verification phase, for a given test segment $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$, and a hypothesized speaker $\mathbf{w}$, the classifier score is obtained as the inner product between the speaker model $\mathbf{w}$ and the average expanded feature vector $\mathbf{b}_{av}$ pertaining to the test segment X, as follows

$$\text{score} = \mathbf{w}^T \mathbf{b}_{av}. \tag{9}$$

4 System Specifications

Given the approaches described in Section 2 and Section 3, four separate subsystems are constructed forming an ensemble of classifiers, as illustrated in Fig. 1. The four classifiers are (i) MFCC GMM-UBM, (ii) TDCT GMM-UBM, (iii) MFCC SVM, and (iv) LPCC SVM. For a given speech utterance, pattern matching is performed in the individual classifier, and a final score is obtained by combining the scores from all the subsystems. The specifications of the subsystems and fusion strategy are described below. The specifications presented below are obtained through numerous experiments carried using the development set of the CSLP SRE corpus, and some other corpora like CCC-VPR2C2005-6000 (*CCC 2-channel corpus for voiceprint recognition 2005 – 6000 speakers*) and NIST SRE corpus [4].

4.1 GMM-UBM

We have two separate GMM-UBM subsystems. The first one is based on MFCC, whereas the second one uses the new TDCT features described in Section 2.3. The UBMs are trained from the development set of the CSLP SRE corpus, which is guaranteed to be disjoint with the evaluation set [3].

Separate UBMs are used for the single-channel and cross-channel tasks. For single-channel task, we derive a 768-component UBM by training independently two

channel-dependent UBMs of size 512 and 256 components, respectively, for landline and cellular channel types. The final UBM model is obtained by aggregating the Gaussian components of the two UBMs, and normalizing the mixture weights so that they sum to one. It should be noted that, channel-dependent UBM is not applicable here because channel-type information is not available for the evaluation data. On the other hand, a different composition is used for cross-channel task. In particular, the UBM has 768 components (1024 components for TDCT GMM-UBM) with 512 components trained from the landline data, and the remaining 256 components (512 components for TDCT GMM-UBM) trained from cellular data. The speaker models are then obtained by adapting the UBM parameters towards the speaker's training data using MAP adaptation principle. Therefore, the speaker models have the same number of Gaussian components with the UBM.

4.2 Spectral SVM

Two different sets of acoustic spectral features, namely MFCC and LPCC, are used thereby forming two separate SVM subsystems. The background or anti-speaker data consist of 4000 utterances extracted from CCC-VPR2C2005-6000. The evaluation set (for text-independent verification and identification tasks) of the CSLP SRE is derived from the CCC-VPR2C2005-1000, which is a subset of the CCC-VPR2C2005-6000 corpus. The CCC-VPR2C2005-1000 subset is discarded from the CCC-VPR2C2005-6000 beforehand so that the 4000 utterances used as the background would not overlap with the evaluation data.

Similar background data is used for the single-channel and cross-channel tasks. For each utterance in the background and for the target speaker, an average expanded feature vector is created. All monomials up to order 3 are used, resulting in a feature space expansion from 36 to 9139 in dimension. These average expanded feature vectors are used in the SVM training. The commonly available SVMTorch [11] is used for this purpose. The result of the training is a vector $\mathbf{w}$ of dimension 9139 which represents the desired target speaker model.

Test normalization (T-norm) method [12] is used to normalize the score. A collection of 500 cohort models are derived from development set of the CSLP SRE corpus. Scores from the cohort models are used to normalize a hypothesized speaker score for a given test segment. Score normalization is accomplished by subtracting the mean and dividing by the standard deviation of the scores produced by the cohort models in response to a given test segment. In order to obtain an accurate estimation of the mean and standard deviation parameters, the population of the cohort models has to be large enough. Furthermore, cohort models have to closely resemble the target speaker models. We believe that it is the best to establish the cohort models from the development set of the CSLP 2006 SRE.

4.3 Subsystems Integration

For a given speech utterance and a hypothesized speaker, pattern matching is performed separately in the four classifiers, giving rise to a 4-dimensional score vector. A final score is then derived from the score vector through a multilayer perceptron (MLP) neural network. The scores from all the subsystems are normalized

to zero mean and unit variance before passing to the neural network. The MLP has 100 hidden neurons and one output neuron with sigmoid activation function. Conjugate gradient algorithm is used for the neural network training.

The development set of the CSLP SRE corpus is used to train two neural networks for score fusion, one for the single-channel verification and identification tasks, and the other one for cross-channel verification and identification tasks.

For speaker verification, the threshold (for the true/false decision) is set at a point whereby the following detection cost function (DCF) is minimized:

$$C_{\mathrm{DET}} = C_{\mathrm{Miss}} \times P_{\mathrm{Miss}} \times P_{\mathrm{Target}} + C_{\mathrm{FalseAlarm}} \times P_{\mathrm{FalseAlarm}} \times \left(1 - P_{\mathrm{Target}}\right), \tag{10}$$

where P_{Miss} and $P_{\mathrm{FalseAlarm}}$ are miss and false-alarm probabilities, respectively, and the parameters $C_{\mathrm{Miss}} = 10$, $C_{\mathrm{FalseAlarm}} = 1$, and $P_{\mathrm{Target}} = 0.05$ are as indicated in the evaluation plan [3].

The speaker identification task is handled through a ranking and pruning procedure. First, a MLP score is derived for each pair of test sample and model. For each test sample, we rank the corresponding trial models with their MLP scores in descending order. Second, we extract all the pairs of test sample and its top-best matching model, rank them in descending order. The top 50% of the pairs are selected as the genuine test trials.

5 Evaluation Results

Fig. 3 and Fig. 4 depict the *detection error tradeoff* (DET) curves of the individual subsystems for the single-channel and cross-channel verification tasks, respectively. As mentioned earlier, these subsystems are fused at the score level using a neural network classifier. The neural networks are trained using the provided development set. The results of fusion are shown in Fig. 3 and Fig. 4 as well. The characteristics of the development set matches well with that of the evaluation set thereby giving a satisfactory fusion result when the trained neural networks are used for the evaluation dataset. The final decision thresholds for the verification tasks are also determined using the development set. On the other hand, the thresholds for the identification tasks is set according to 1:1 in-set and out-of-set ratio stated in the evaluation plan [3]. That is, the speaker identification tasks are performed in an open set manner.

Table 2 summarizes the performance of our submission to the CSLP 2006 SRE based on the actual DCF value and the *identification correctness rate* [3] for verification and identification tasks, respectively. As expected, channel mismatch makes the recognition tasks more difficult. The degradation in performance can be observed from both the DCF value and the identification correctness rate.

Table 3 summarizes the equal-error rates (EERs) and the minimum DCF values for the individual and fused scores. Clearly, the subsystems fuse in a complementary way reducing error rates substantially. Taking the LPCC SVM as baseline, the fused systems give relative EER improvements of 52% and 22% for single-channel and cross-channel conditions, respectively. On the other hand, the relative improvements in minimum DCF for single-channel and cross-channel verification tasks are 57% and

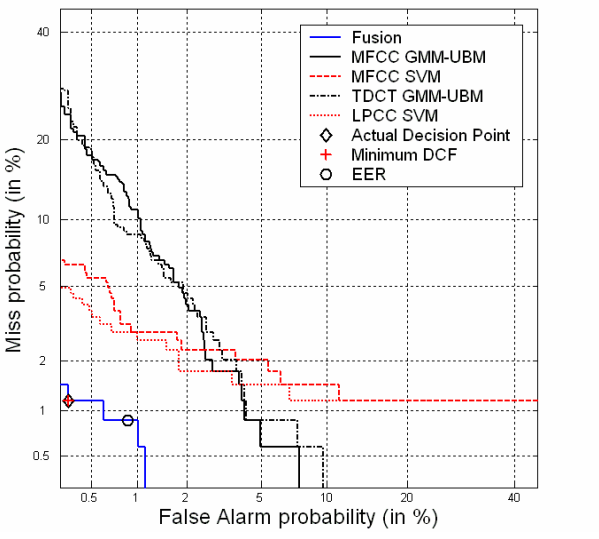

Fig. 3. DET curves for single-channel verification task

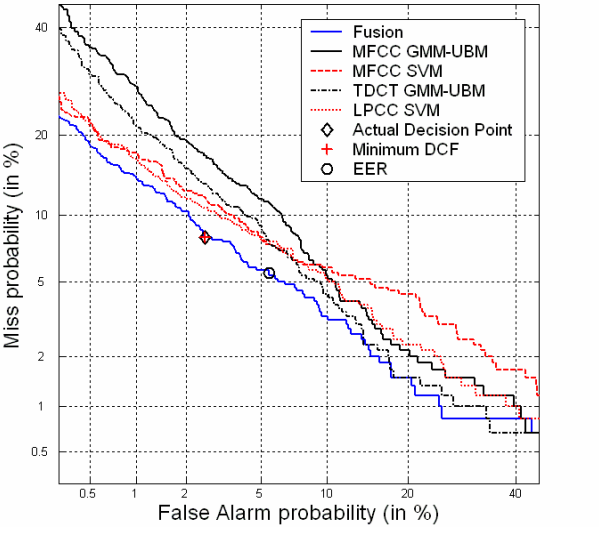

Fig. 4. DET curves for cross-channel verification task

18%, respectively. The gains in performance are due both to the different features (MFCC, LPCC, and TDCT) and the different speaker modeling techniques (SVM and GMM). From the DET curves, it can be noted that SVM and GMM complement each other at different threshold values. In particular, SVM performs best at high threshold values (i.e., upper left corner), while GMM dominates at low threshold values (i.e.,

Table 2. Performance of IIR submission to the 2006 CSLP SRE based on the DCF value and the identification correctness rate.

	Actual DCF value (×100)	Identification Correctness Rate
Single-Channel Verification Task	0.90	
Cross-Channel Verification Task	6.42	
Single-Channel Identification Task		97.16%
Cross-Channel Identification Task		86.45%

Table 3. Comparison of EER and minimum DCF for IIR individual subsystems/final system in speaker verification tasks

System	Single-channel verification task		Cross-channel verification task	
	EER (%)	Min DCF (×100)	EER (%)	Min DCF (×100)
MFCC GMM-UBM	2.54	3.44	7.70	10.22
MFCC SVM	2.31	2.31	6.71	8.10
TDCT GMM-UBM	2.85	3.89	6.69	8.68
LPCC SVM	1.81	2.09	7.03	7.79
Fusion	0.86	0.90	5.50	6.42

lower left corner). It can also be observed that SVM performs best with LPCC features. On the other hand, GMM performs best with MFCC and TDCT features for single and cross-channel tasks, respectively, mainly due to the difference in the UBMs. Further research into optimizing features for each of the modeling techniques should be carried out.

6 Conclusions

A description of a speaker recognition system has been presented as it was developed for the CSLP 2006 SRE. Our submission was built upon three different acoustic spectral features and two different speaker modeling techniques giving rise to four subsystems, namely, MFCC GMM-UBM, TDCT GMM-UBM, MFCC SVM, and LPCC SVM. These subsystems were combined at the score level through a MLP neural network in a complementary way. The fused system achieved an EER of 0.86% and 5.50% for single-channel and cross-channel verification tasks, respectively. Promising results were also obtained for identification tasks, where identification rates of 97.16% and 86.45% were obtained under single-channel and cross-channel conditions, respectively. The SRE results confirm a successful design and implementation of speaker recognition system. Nevertheless, continuous effort that makes use of the common platform provided by the CSLP SRE event should be carried out.

References

1. S. Furui, "Speaker verification," in *Digital Signal Processing Handbook*, V. K. Madisetti and D. B. Williams, Eds. Boca Raton: CRC Press LLC, 1999.
2. T. F. Quatieri, *Discrete-time speech signal processing: principles and practice*. Upper-Sadder River, NJ: Prentice-Hall, 2002.
3. *Evaluation Plan for ISCSLP'2006 Special Session on Speaker Recognition*, Chinese Corpus Consortium, Apr. 2006.
4. D. A. van Leeuwen, A. F. Martin, M. A. Przybocki, and J. S. Bouten, "NIST and NFI-TNO evaluations of automatic speaker recognition," *Computer Speech and Language*, vol. 20, pp. 128-158, 2006.
5. S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-28, no. 4, Aug. 1980.
6. 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," accepted for presentation in *International Symposium on Chinese Spoken Language Processing,* 2006.
7. F. Bimbot, J. F. Bonastre, C. Fredouille, G. Gravier, I. Margin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-Garcia, D. Petrovska, and D. A. Reynolds, "A tutorial on text-indepent speaker verification," *Eurasip Journal on Applied Signal Processing*, vol. 4, pp. 430-451, 2004.
8. 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.
9. W. M. Campbell, "Generalized linear discriminant sequence kernels for speaker recognition," in *Proc. ICASSP*, pp. 161-164, 2002.
10. 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.
11. R. Collobert and S. Bengio, "SVMTorch: support vector machines for large-scale regression problems," Journal of Machine Learning Research, vol. 1, pp. 143-160, 2001.
12. 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.