Temporal Discrete Cosine Transform: Towards Longer Term Temporal Features for Speaker Verification

Tomi Kinnunen[1], Chin Wei Eugene Koh[2], Lei Wang[2],
Haizhou Li[1], Eng Siong Chng[2]

[1]Speech and Dialogue Processing Lab, Institute for Infocomm Research (I2R)
Singapore 119613
[2]School of Computer Engineering, Nanyang Technological University (NTU)
Singapore 639798
{ktomi, hli}@i2r.a-star.edu.sg
{kohc0026, wang0161, aseschng}@ntu.edu.sg

Abstract. In this paper, we propose the *temporal discrete cosine transform* (TDCT) feature for the speaker verification task. The TDCT feature captures temporal information from a longer time context beyond the conventional delta and double-delta coefficients. We evaluate the effectiveness of the TDCT feature on the NIST 2001, NIST 2004, and NIST 2005 speaker recognition benchmark corpora by using a standard GMM-UBM recognizer. We compare our results against the standard MFCC+Δ+ΔΔ front end, and with the *shifted delta cepstrum* (SDC) feature which is commonly used in the language identification task. The results indicate that the TDCT and SDC give similar accuracy, and that the TDCT feature outperforms MFCC+Δ+ΔΔ in most of the cases.

Keywords: Text-independent speaker verification, temporal features, temporal discrete cosine transform, shifted delta cepstrum, Gaussian mixture model

1 Introduction

Speaker verification is a task of determining whether a given voice sample was produced by a claimed person. A speaker verification system extracts features from the unknown person's utterance and compares them with a previously stored model of the claimed person. The match score is then compared with a decision threshold so as to make an "accept" or "reject" decision.

Many features have been proposed for speaker verification, including spectral features, prosodic patterns, phonetic features and lexical features [1]. In this study, we focus on spectral features as they provide the best individual accuracy. The spectral front-end of speaker recognition systems extracts short-term features from 20-30 millisecond frames. The feature vectors are appended with their first- and second order time derivative estimates, respectively known as *delta*- (Δ) and *double-delta* (ΔΔ) features. Each speaker's training vectors are then used to train a Gaussian mixture model (GMM).

The delta and double-delta coefficients capture short-term speech dynamics (temporal context about 50-100ms). This interval however doesn't capture longer term features like prosodic gestures and syllable usage. These "high-level" features are already used in speaker recognition systems [1, 2]. The basic idea in these approaches is to first convert the utterance into a sequence of discrete symbols by using a tokenizer (such as phone recognizer). The symbol sequences are then modeled and matched as if they were text documents. One problem with this approach is the rather complex implementation. It is also time-consuming to include multiple recognizers based on different features and models.

In this paper, we propose a *temporal discrete cosine transform* (TDCT) feature for speaker verification. The TDCT feature applies discrete cosine transform (DCT) to the cepstral feature "signals" over consecutive frames. Our work was inspired by a recent study [3] in which DCT was applied as a pre-processing step in a keyword spotting application. In [3], DCT was applied to the mel-filtered spectrogram. In this paper, we apply DCT directly to the MFCC coefficients so as to reduce computational overhead.

Another similar temporal feature is the so-called *shifted delta cepstrum* (SDC) which stacks several delta cepstra vectors over a long time context into one vector. The SDC feature was originally proposed in [4] for the language identification task. A more recent study of the SDC feature is [5] and this feature is used presently as a standard method in language identification. We however are unaware of SDC studies in speaker verification. We were curious to know whether it could be applied to speaker verification as well.

The rest of the paper is organized as follows. Section 2 describes the SDC and TDCT features. Section 3 describes the details of the data sets and classifier used in the experiments. Section 4 gives the experimental results. Finally, discussion and conclusions are given in Section 5.

2 SDC and TDCT Features

The usage of multiple frames around the current one increases the temporal context of each feature vector generated. This increase in the temporal context allows for the capturing of longer term speech dynamics. Previously, each MFCC vector only captures information in the immediate locality of the current frame. We hypothesize that the longer term dynamics would allow for better speaker discrimination.

2.1 Shifted Delta Cepstrum

The SDC feature [4, 5] as illustrated in Fig. 1, stacks several delta cepstrum frames together to form a longer feature vector. The method has four control parameters:

- Number of MFCC coefficients (M)
- Time delay for delta feature computation (d)
- Number of delta vectors for concatenation (K)
- Frame advance for consecutive SDC block computation (P)

The typical MFCC+Δ+ΔΔ front-end for speaker recognition uses 12 coefficients and a time difference of 1 frame for delta and double-delta computation. This results in a 36-dimension feature vector. The MFCC vectors that have been extracted traditionally are used for the computation of SDC. Therefore, we use M=12 and d=1 as a starting point and consider K and P as the major control parameters of the SDC method.

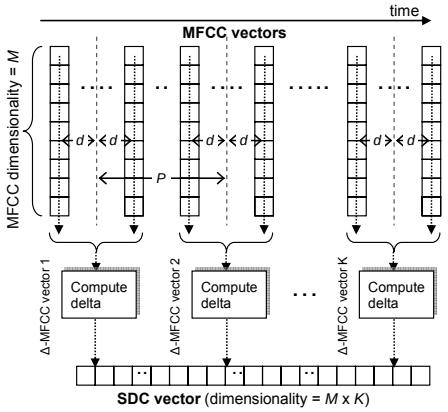

Figure 1: Illustration of the SDC feature computation.

2.2 Temporal Discrete Cosine Transform

TDCT features are also derived from MFCC features (see Fig. 2). In this method, each cepstral coefficient is considered as an independent "signal" which is windowed in blocks of length B. Each block is transformed into discrete cosine transform (DCT) coefficients. We retain the lowest K coefficients as they contain most of the energy. Finally, the DCT coefficients of all MFCC coefficients are stacked to form a long vector of dimensionality MK. The next TDCT vector is computed by advancing the block by one frame.

The difference between SDC and TDCT is in the way they reduce the dimensionality of the otherwise very high-dimensional temporal context. SDC sub-samples the feature vector sequence by dropping intermediate frames. TDCT, on the other hand, utilizes information from all the frames within the block and reduces the number of features by retaining the "low-frequency" part of the feature stream.

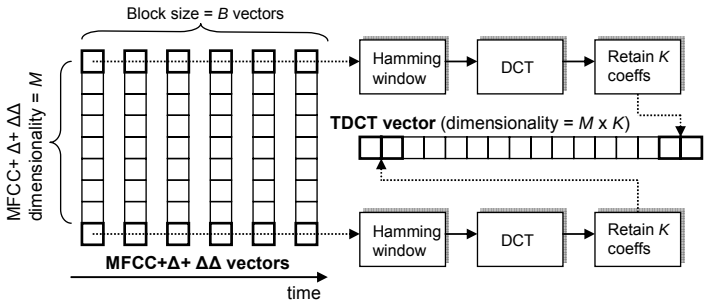

Figure 2: Illustration of the TDCT feature computation.

2.3 Setting the Parameters

In dynamic features such as SDC and TDCT, the duration of the time context in physical units is of special interest as it determines what kind of information is contained in the feature vectors. Assuming that the original MFCC frame length is L milliseconds and frame advance is S milliseconds, a single Δ vector spans over $S+L$ milliseconds. Similarly, a single $\Delta\Delta$ vector spans over $2S+L$ milliseconds. Thus, a typical MFCC+Δ+$\Delta\Delta$ vector (L=30ms, S=20ms) contains information over an interval of 70 milliseconds.

For the SDC feature, the parameters d, P and K control the duration of the time context. Each SDC vector is computed over $N = (K\text{-}1)P + 2d + 1$ MFCC frames. This corresponds to a physical length of $(N\text{-}1)S + L = [(K\text{-}1)P + 2d]S + L$ milliseconds. As d, P and K affect the time context duration, they are expected to be critical parameters. Table 1 lists some of the typical parameter values for this feature and the corresponding time context duration.

For the TDCT feature, the time context duration is B cepstral vectors. Since we use the MFCC+Δ+$\Delta\Delta$ vectors as the input to TDCT, this corresponds to a physical length of $(B\text{-}1)S + L + 2S$ milliseconds. Table 2 lists some typical parameter values for the TDCT feature and their corresponding time context durations.

Table 1: Time span of SDC feature in milliseconds (L=30ms,S=20ms).

	d =1			d =2			d =3		
	P =1	P =2	P =3	P =1	P =2	P =3	P =1	P =2	P =3
K=1	70	70	70	110	110	110	150	150	150
K=2	90	110	130	130	150	170	170	190	210
K=3	110	150	190	150	190	230	190	230	270
K=4	130	190	250	170	230	290	210	270	330
K=5	150	230	310	190	270	350	230	310	390

Table 2: Time span of TDCT feature in milliseconds (L=30ms, S=20ms, d=1)

B=1	B=2	B=3	B=4	B=5
70	90	110	130	150
B=6	B=8	B=12	B=16	B=20
170	210	290	370	450

3 Experimental Setup

We use the standard GMM-UBM classifier [6] with a global verification threshold to assess the performance of the features. We follow the NIST speaker recognition evaluation rules in our experiments [7].

3.1 Data Sets

We use the NIST 2001, NIST 2004, and NIST 2005 speaker recognition benchmark corpora for our experiments [7]. All of the corpora are conversational telephone speech recorded over various channel conditions and they are disjoint. The NIST 2001 corpus is used for tuning the control parameters of the SDC and TDCT methods. The NIST 2004 corpus is used for studying the effects of channel and language mismatch and also for tuning the combination weight for the fusion of TDCT and MFCC+Δ+ΔΔ features. The NIST 2005 corpus is then used for validating the effectiveness of the fusion. Table 3 summarizes the data sets used, including the training data for the universal background model.

Table 3: The data sets used in the experiments

Task	*Evaluation set*	*UBM training set*
1. Parameter setting of SDC and TDCT	NIST 2001 (41 speakers, 2000 trials)	NIST 2001 development set
2. Effects of VAD and feature normalization, and comparison of SDC, TDCT, and MFCC+Δ+ΔΔ	NIST 2001 (full set of 174 speakers and 22418 trials)	NIST 2001 development set
3. Study of mismatch, and tuning of fusion weight for MFCC+Δ+ΔΔ and TDCT	NIST 2004 core test (1conv4w-1conv4w)	NIST 2005 1conv training
4. Validation of fusion	NIST 2005 core test (1conv4w-1conv4w)	NIST 2004 1conv training

The 1-speaker detection task of the NIST 2001 corpus contains 174 target speakers and there are 22418 verification trials in total (2038 genuine + 20380 impostor). The amount of training data is two minutes per speaker and the length of the test segment varies from a few seconds up to one minute. In order to allow for faster experimentation, we used a small subset for preliminary parameter tuning. This subset has 41 target speakers (20M+21F) and 2000 verification trials. This list was generated by random sampling from the original NIST trial list while keeping the genuine/impostor ratio as per the original 1:10.

The core test of the NIST 2004 corpus includes data from 616 speakers (248M+368F) and there are 26224 verification trials in total (2386 genuine and 23838 impostors). On the other hand, the NIST 2005 corpus core test includes data from 646 speakers (274M+372F) and the number of trials is 31418 (1941 genuine and 29477 impostors). For both the NIST 2004 and NIST 2005 corpora, the training and testing files of the core test consist of five minutes of conversational speech. As each five minute conversation is shared between two parties, each conversation is estimated on average to contain about 2.5 minutes of speech data from the party of interest.

3.2 MFCC and GMM-UBM Setup

The MFCC coefficients 1-12 that we use are computed from a 27-channel mel-filterbank over L=30ms frames and window shift S=20ms. For SDC feature, the 24 MFCC+Δ features are used as input while for the TDCT feature, the 36-dimensional MFCC+Δ+ΔΔ features are used as input. We apply energy-based voice activity detection (VAD) followed by conversation-level mean and variance normalization for all of the three feature sets. Since the VAD was performed on a frame-by-frame basis, the dropping of low-energy feature vectors for the SDC and TDCT features is done using a voting scheme. In order for a feature vector to be dropped, 60% or more of the multiple frames that have been used to generate each SDC or TDCT vector must be deemed as non-speech.

We use diagonal-covariance Gaussian mixture model (GMM) as the speaker model [6]. An N-component universal background model (UBM) is developed by training independently two gender-dependent GMMs of size $N/2$ using the expectation-maximization (EM) algorithm. A gender-independent N-component background model is constructed by pooling the parameter vectors of the two models followed by mixture weight renormalization. The target speaker models are obtained by adapting the UBM parameters towards the speaker's training data using maximum a posteriori (MAP) adaptation principle. We adapt only the mean vectors and use a relevance factor of 15. The average log-likelihood ratio between the target model and the UBM is used as the match score. The fast GMM-UBM scoring algorithm [6] using the top-20 mixture components is employed.

4 Results

4.1 Setting the Parameters

We first study the effect of control parameters on recognition accuracy by using the 41 speaker tuning set, each speaker model having 64 Gaussian components. For the SDC feature, parameters d=1, P=3, and K=3 are suggested in [5] for the language recognition task. We first start with P=3, K=3, using MFCC features as the input and vary d= {1, 2, 3}. The detection error trade-off (DET) curves in Fig. 3 show that d=1 gives worse result when compared with d=2, and d=3. For the rest of the SDC experiments, we fix d=2.

Next, we study the effect of the K and P parameters of the SDC feature. Equal error rates (EER) for the tuning set are given in Table 4. We observe that setting P=2 and K=4 gives the best result (EER=13.6%), along with using MFCC+Δ features as the source. The dimensionality of the SDC feature for this setting is 4x24 = 96, and the time context duration 230 milliseconds. Interestingly, the other parameter setting having exactly same time span (P=3, K=3) gives a worse error rate of EER=14.3%. One reason for this could be smaller dimensionality (3x24=72).

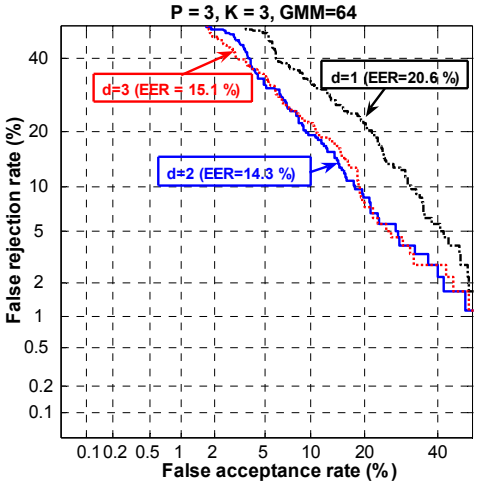

Figure 3: Effect of the delay parameter of the SDC method (using NIST 2001 subset).

Next, we study the parameters of the TDCT method. The results in Table 5 show that best parameter setting is B=8 and K=3, which corresponds to a time duration of 210 millisecond and dimensionality of 3x36=108. Comparing the parameter stability of the SDC and TDCT method, the latter has smaller variance in the error rates over the experimented combinations.

We next study the effects of voice activity detection (VAD) and feature normalization by using the full NIST 2001 detection list with GMM of 512 components. Based on the tuning results, we set (P, K) = (2, 4) for the SDC feature and (B, K) = (8, 3) for the TDCT feature. The results in Figures 4 and 5 show that the usage of the VAD and feature normalization improved accuracy.

Table 4: Effect of SDC control parameter on the NIST 2001 subset, EER (%)

d=2, K=3		d=2, P=2	
P=1	15.1	K=1	19.4
P=2	14.1	K=2	15.3
P=3	14.3	K=3	14.1
P=4	15.5	K=4	**13.6**
P=5	14.8	K=5	14.4
Avg.=15.1, *Std*=1.6			

Table 5: Effect of TDCT control parameters on the NIST 2001 subset, EER (%)

	K=2	K=3	K=4
B=16	14.2	13.6	14.4
B=12	13.8	14.1	13.9
B=8	14.1	**12.4**	14.2
Avg.=14.4, *Std.*=1.2			

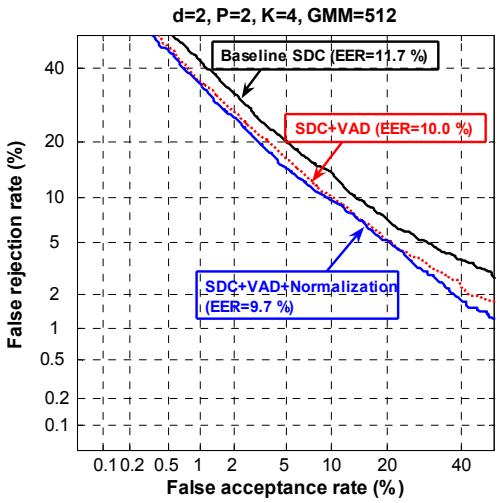

Figure 4: Effect of VAD & feature normalization on SDC (NIST 2001 core test).

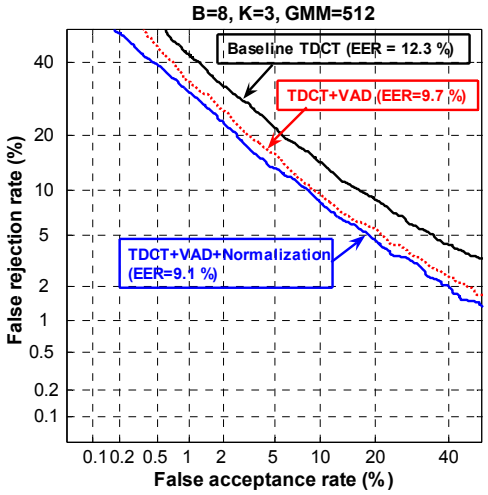

Figure 5: Effect of VAD & feature normalization on TDCT (NIST 2001 core test).

4.2 Comparison of MFCC+Δ+ΔΔ, SDC, and TDCT

We next compare the MFCC+Δ+ΔΔ, SDC and TDCT features (see Fig. 6). VAD and feature normalization is used for all the features as explained in Section 3.2. The three individual recognizers yield rather similar performance in terms of EERs. However, the individual curves are rotated differently, implying differences in the extreme ends of the DET curve. For instance, the MFCC+Δ+ΔΔ recognizer is worse compared with

the SDC and TDCT recognizers at low false rejection rates (user-convenient applications). The SDC feature in turn is worse compared with MFCC+Δ+ΔΔ and TDCT features at low false acceptance rates (secure applications). Since the SDC feature seems to be more sensitive to it's parameter settings, we drop it from further experiments and concentrate on comparing the properties of MFCC+Δ+ΔΔ versus TDCT.

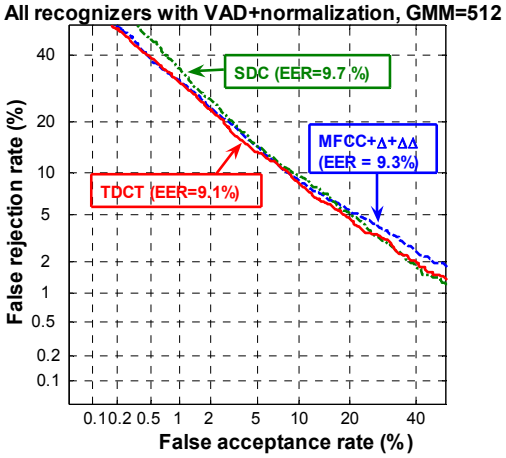

Figure 6: Comparison of the three features (NIST 2001 core test).

4.3 Channel and Language Mismatch

We next study the effects of channel and language mismatch on the core test of the NIST 2004 corpus. We further enhance the GMM-UBM recognizer by using gender-dependent scoring in these tests, i.e. using a background model that has the same gender as the target speaker, as permitted by the NIST test protocol [7]. This yields an absolute improvement of about 1% in EER over the combined background model. Both the MFCC+Δ+ΔΔ and TDCT classifiers use the same background speakers, though the number of Gaussian components is optimized separately for each feature. We use 512- and 256-component gender-dependent GMMs for MFCC+Δ+ΔΔ and TDCT respectively.

Using the information provided in the key file, we extract subsets of trials according to the similarity or difference of the testing and training utterance channel type (landline /cellular /cordless), and language (English / Spanish / Russian / Mandarin / Arabic). We decided to consider the "cordless" and "landline" channel types as being the same. From the results shown in Table 6, we observe that both features degrade when either channel or language mismatch is present. This is an expected result. Furthermore, the TDCT feature gives consistently better results. The improvement however is marginal. We therefore next study whether fusion of the two features improves accuracy.

Table 6: Effect of channel and language mismatch (NIST 2004 core test), EER (%)

Factor	*Condition*	*#Trials*	*MFCC+ Δ+ΔΔ*	*TDCT*
Channel	Same channel	15023	13.9	13.5
	Different channel	10649	15.9	15.6
	Unknown	552	11.5	10.3
Language	Same language	16981	13.8	13.4
	Different language	9243	16.2	16.1
	All trials	**26224**	**14.9**	**14.5**

4.4 Combining MFCC+Δ+ΔΔ and TDCT

We combine the scores of the MFCC+Δ+ΔΔ and TDCT classifiers as $w_{MFCC}s_{MFCC}$ + $(1-w_{MFCC})s_{TDCT}$, where $0 \le w_{MFCC} \le 1$ is the weight of the MFCC+Δ+ΔΔ classifier and $s_{(.)}$ are the average log likelihood ratio scores of the individual classifiers. The result on the NIST 2004 corpus is shown in Fig. 7. Any combination weight improves the accuracy over the TDCT classifier alone. The best result is obtained by selecting w_{MFCC}=0.8, which reduces the EER from 14.5% to 14.0%.

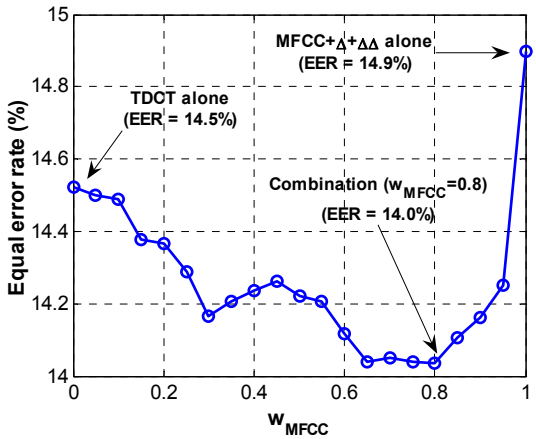

Figure 7: Accuracy of the score fusion of MFCC+Δ+ΔΔ and TDCT (NIST 2004 core set).

Table 7 displays the fusion result on an independent validation set, the NIST 2005 corpus. In addition to the core test (5min train + 5min test) we optimize the weights for six other test conditions. In all cases, the weight is set to an optimum value on the corresponding condition of the NIST 2004 corpus. We observe that the fusion is successful in most of the cases. The improvement itself however is minor, which would suggest perhaps using a different fusion strategy.

Table 7: Accuracy of the recognizers on the NIST 2005 core test. The fusion weights have been optimized separately for each of the conditions on the NIST 2004 core test.

Average segment length		*EER (%)*		
Training	*Testing*	*MFCC*	*TDCT*	*Fusion*
10 sec	10 sec	28.6	30.4	**27.8**
5 min	10 sec	19.8	18.8	**18.6**
5 min	5 min	13.6	13.8	**13.3**
3 x 5 min	10 sec	16.2	**15.6**	15.7
3 x 5 min	5 min	10.4	10.1	**10.0**
8 x 5 min	10 sec	14.5	14.5	**13.8**
8 x 5 min	5 min	9.6	**9.2**	9.5

5 Discussion and Conclusions

In this paper, we introduced the TDCT feature for the speaker verification task with the aim of incorporating longer-term temporal information into the features. We have studied the accuracy of the TDCT feature using a GMM-UBM classifier approach on three NIST corpora, and compared the results with the conventional MFCC+Δ+ΔΔ front-end. We also compared results to an SDC front-end which is a standard feature for the language identification task.

Comparing the SDC and TDCT methods, the differences as observed on the NIST 2001 corpus were small. However, the TDCT feature with fewer number of control parameters seems more attractive. The best parameter settings of the SDC (d=2, P=2, K=4) and the TDCT (B=8, K=3) methods correspond to time contexts of 230 and 210 milliseconds respectively. These durations are close to each other and they both are significantly longer compared with the 70 milliseconds of the standard MFCC+Δ+ΔΔ front-end. We hypothesize that these longer-term features contain added speaker information and believe this approach has room for further studies.

Further analysis of the MFCC+Δ+ΔΔ and TDCT front-ends on the NIST 2004 and NIST 2005 corpora indicated that the TDCT outperforms MFCC+Δ+ΔΔ in most of the cases and a weighted score combination of the features yields also yields improvement. The improvement however was minor. A possible approach to better harness the potential of the TDCT feature would be to use a different classifier, such as the *support vector machine* (SVM) [8]. We are currently working on this direction.

The simple weighted score fusion might also be inadequate to take full advantage of the individual benefits of the short- and long-term temporal features. A possible clue in that direction would be that we have successfully combined MFCC+Δ+ΔΔ, TDCT and LPCC+Δ features using a multilayer perceptron (MLP) score combiner in [9]. The relative improvements on the EER over the best individual classifiers when tested on the ISCSLP 2006 speaker evaluation special session's evaluation corpus were 18% and 52% respectively on mismatched and matched channel conditions [9].

References

[1] D.A. Reynolds, W. Andrews, J. Campbell, J. Navratil, B. Peskin, A. Adami, Q. Jin,D. Klusacek, J. Abramson,R. Mihaescu, J. Godfrey,D. Jones, and B. Xiang, "The SuperSID Project: exploiting High-Level Information for High-Accuracy Speaker Recognition", *Proc. ICASSP 2003*, pp. 784—787, Hong Kong, 2003.

[2] E. Shriberg, L. Ferrer, S. Kajarekar, A. Venkataraman, and A. Stolcke, "Modeling prosodic feature sequences for speaker recognition", *Speech Communication*, 46(3-4): 455-472, July 2005.

[3] I. Szöke, P. *Schwarz, P. Matejka,* L. Burget, M. Karafiat, M. Fapso, J Cernocky, "Comparison of Keyword Spotting Approaches for Informal Continuous Speech", *Proc. 9th European Conf. Speech Communication and Technology (Interspeech-Eurospeech 2005)*, pp. 633—636, Lisbon, Portugal, 2005.

[4] B. Bielefeld, "Language identification using shifted delta cepstrum," In Fourteenth Annual Speech Research Symposium, 1994.

[5] P.A. Torres-Carrasquillo, D.A. Reynolds, E. Singer, M.A. Kohler, R.J. Greene, and J.R. Deller, Jr., "Approaches to language identification using Gaussian mixture models and shifted delta cepstral features," *Proc. Int. Conf. Spoken Language Processing*, pp. 89-92, Denver, USA, 2002.

[6] D.A. Reynolds and T.F. Quatieri and R.B. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models", *Digital Signal Processing,* 10(1):19-41, 2000.

[7] *NIST Speaker Recognition Evaluations homepage*, http://www.nist.gov/speech/tests/spk/index.htm (valid June 2006).

[8] W.M. Campbell, J.P. Campbell, D.A. Reynolds, E. Singer, P.A. Torres-Carrasquillo, "Support Vector Machines for Speaker and Language Recognition", *Computer Speech & Language*, 20(2-3):210-229, 2006.

[9] K.A. Lee, H. Sun, R. Tong, B. Ma, M. Dong, C. You, D. Zhu, C.W.E. Koh, L. Wang, T. Kinnunen, E.S. Chng, H. Li, "The IIR Submission to CSLP 2006 Speaker Recognition Evaluation", submitted to *5th International Symposium on Chinese Spoken Language Processing,* 2006.