

Multi-resolution time frequency feature and complementary combination for short utterance speaker recognition

Zhi-Yi Li · Wei-Qiang Zhang · Jia Liu

Published online: 23 October 2013
© Springer Science+Business Media New York 2013

Abstract A human speaker recognition expert often observes the speech spectrogram in multiple different scales for speaker recognition, especially under the short utterance condition. Inspired by this action, this paper proposes a novel multi-resolution time frequency feature (MRTF) extraction method, which is obtained by performing a 2-Dimensional discrete cosine transform (DCT) in multi-scale on the time frequency spectrogram matrix and then selecting and combining to the final multi-scaled transformed elements. Compared to the traditional Mel-Frequency Cepstral Coefficient (MFCC) feature extraction, the proposed method can make better use of multi-resolution temporal-frequency information. Beyond this, we also proposed three complementary combination strategies of MFCC and MRTF: in feature level, in i-vector level and in score level. Comparing their performance. We found the best results are obtained by combination in i-vector level. In the three NIST 2008 Speaker Recognition Evaluation datasets, the proposed method is the most effective for improving the performance under short utterance than under long utterance. And after the combination, we can achieve an EER of 11.32 % and MinDCF of 0.054 in the 10sec-10sec trials on the male dataset, which is an absolute 3 % improvement of EER than the best reported result in this field.

Keywords Multi-resolution time frequency feature · I-vector · Complementary combination · Speaker recognition · Short utterance

Z.-Y. Li (✉) · W.-Q. Zhang · J. Liu
Tsinghua National Laboratory for Information Science and Technology,
Department of Electronic Engineering, Tsinghua University,
Beijing 100084, China
e-mail: lizhiyi.china@gmail.com

W.-Q. Zhang
e-mail: wqzhang@tsinghua.edu.cn

J. Liu
e-mail: liuj@tsinghua.edu.cn

1 Introduction

Speaker recognition (SRE) refers to recognizing persons from their voice. No two individuals sound identical because of their different physical parts of voice production organs, such as vocal tract shapes and larynx sizes, and also their different speaking manner such as the accent, rhythm, intonation style etc. [13]. It has many widely used applications, such as telephone banking, information security, and forensics. However, some studies have shown that all well known technologies, including the state-of-the-art *i*-vector, exhibit a sharp decline in performance and seriously limit the widespread use of speaker recognition technologies in many practical applications [3, 9, 17].

For improving the performance, especially in short utterance condition, one of the main solutions is to propose a new better feature extraction method. Traditional feature extraction method Mel frequency cepstral coefficient (MFCC) has some disadvantages, even though it is often simple and efficient. One of the main disadvantages of MFCC is appending delta and acceleration feature to basic feature, which is not a good way to extract the temporal information between frames, because it introduces correlation into the feature vector. This is not good for the backend classifier. To improve MFCC, time frequency cepstral (TFC) feature was proposed based on MFCC in [23, 24], which is obtained by performing a temporal discrete cosine transform (DCT) on the cepstrum matrix for decorrelation of feature vectors. TFC feature performs better than MFCC feature. However, it obviously is a fixed single resolution analysis without efficiently using the information in multi-resolution analysis. As a matter of fact, a voiceprint recognition human expert often observes the speech spectrogram in different scales when doing speaker recognition, especially on short utterance. In fact, the reason behind this action is that integrating the information from multi-resolution observation can improve the accuracy of recognition.

Based on the above consideration, in this paper we propose a novel multi-resolution time frequency feature (MRTF) extraction method under short utterance condition. First, the spectrogram of speech is divided to the multi-resolution blocks by the different scales. Then it is followed by 2-dimensional DCT and we select the major components under every single resolution. Finally, MRTF feature is obtained by stacking all the selected components of multi-resolution analysis and doing PCA to reduce the dimension and redundancy. This proposed method can take advantage of the proposed TFC extraction method and the multi-resolution analysis. It is obviously different from the same named multi-resolution feature extraction methods in [1, 7, 8], which either use multiple frame lengths and frame rates to extract the feature or use wavelet transformation in feature extraction.

From an information-theoretic viewpoint, complementary combination and fusion can integrate more information for classification and improve the recognition performance. It has been proved to be effective in our previous work [14, 15]. In this paper, we will propose three complementary combination strategies between the traditional MFCC and proposed MRTF, which are: in feature level, in *i*-vector level and in score level, and then compare their performance. We use the state-of-the-art *i*-vector based speaker recognition technologies with cosine distance scoring (CDS) and evaluate the performance on NIST 2008 SRE dataset.

The outline of the paper is as follows. At first, we briefly describe state-of-the-art i-vector baseline in Section 2. Section 3 presents the proposed multi-resolution time frequency feature extraction method. The three complementary strategies are proposed in Section 4. The experiments and results are given in Section 5. Finally, Section 6 concludes the paper.

2 The i-vector based speaker recognition

Inspired by classical Joint Factor Analysis (JFA) modeling based on factor analysis and some following work [2, 4, 11], Dehak et al. [3] have recently proposed the state-of-the-art i-vector based speaker recognition technology, which was derived from the GMM-UBM technology [20]. This modeling method is based on defining only a single space, instead of two separate spaces in JFA. This new space, which is named as total variability space, simultaneously contains the speaker and channel variabilities. Hence, during i-vector extraction, there is no distinction between the speaker effects and the channel effects in GMM supervector space, because experiments have demonstrated the channel factors of the JFA, which normally model only channel effects, also contain speaker information [2].

In i-vector model, the new speaker- and channel-dependent GMM supervector defined can be written as follows:

$$M = m + Tw \quad (1)$$

where m is a both speaker and channel independent component (usually using UBM supervector), T is a rectangular matrix of low rank representing the primary directions of variability and w is a random vector having a standard normal distribution $\mathcal{N}(0, I)$. The component w are named total factors, also called the i-vector for short. M is assumed to be normally distributed with mean vector m and covariance matrix TT^t . The process of training the matrix T is exactly the same as training the eigenvoice V matrix in JFA [10, 12], except that there is no need to consider the speaker labels. Because of this, the i-vector modeling can be seen as a simple factor analysis, which allows us to project a high-dimensional GMM supervector derived from speech utterance onto the low-dimensional i-vector in total variability space.

As the space T contains both the speaker's session and channel variability, extracted i-vector require further compensation technologies to attenuate the effects. Comparing a number of existing approaches such as Within-class Covariance Normalization (WCCN), Linear Discriminant Analysis (LDA) and Nuisance Attribute Projection (NAP), the experiments in [3] show that the best results are obtained with the LDA and WCCN combination followed by cosine distance scoring (CDS) classifier or other classifier like PLDA [22]. For fairly comparing to the result, we use the same configuration as in [3]. The method we used will be outlined in the remainder of this section.

2.1 LDA

In the case where all utterances of a given speaker are assumed to represent one class, the motivation for using LDA to i-vector is to define new special axes that

minimize the intra-class variance caused by channel effects, and to maximize the variance between speakers [16]. The advantage of the LDA approach is based on discriminative criteria designed to remove unwanted directions and to minimize the information removed about variance between speakers. Mathematically, the optimization problem of LDA can be defined as the Rayleigh coefficient as function of space direction v .

$$J = \frac{v^t S_b v}{v^t S_w v} \tag{2}$$

The solution of this problem is to maximize the Rayleigh coefficient. This maximization is used to obtain a project matrix P composed by the best eigenvectors (those with highest eigenvalues) of the general eigenvalue problem as following equation:

$$S_b v = \lambda S_w v \tag{3}$$

$$S_b = \sum_{s=1}^S (w_s - \bar{w})(w_s - \bar{w})^t \tag{4}$$

$$S_w = \sum_{s=1}^S \frac{1}{n_s} \sum_{i=1}^{n_s} (w_s^i - \bar{w}_s)(w_s^i - \bar{w}_s)^t \tag{5}$$

where the λ is the diagonal matrix of the eigenvalues, $\bar{w}_s = \frac{1}{n_s} \sum_{i=1}^{n_s} w_s^i$ is the mean of the i -vectors of each speaker, S is the number of speakers and n_s is the number of utterances of s th speaker. The raw i -vectors are then submitted to the projection matrix P obtained from LDA.

2.2 WCCN

WCCN is proposed to use the within-class covariance matrix to normalize the cosine kernel function in order to compensate for intersession variability [6]. We also assume that all utterances of a given speaker belong to one class. The within class covariance matrix is computed as follows:

$$W = \frac{1}{S} \sum_{s=1}^S \frac{1}{n_s} \sum_{i=1}^{n_s} (w_s^i - \bar{w}_s)(w_s^i - \bar{w}_s)^t \tag{6}$$

where $\bar{w}_s = \frac{1}{n_s} \sum_{i=1}^{n_s} w_s^i$ is the mean of the i -vectors of each speaker, S is the number of speakers and n_s is the number of utterances of s -th speaker. Then, the mapping matrix L can be obtained through a Cholesky decomposition of matrix $W^{-1} = LL^t$.

2.3 CDS classifier

The CDS is an efficient and well-performed classifier for the i -vector based system, which directly uses the value of the cosine kernel between the target speaker i -vector and the test i -vector as a final decision score. The value of this kernel is then compared to the threshold θ to take the final decision. The CDS scoring of the target and test i -vectors, following the LDA and WCCN transformation, is as follows:

$$score(w_{target}, w_{test}) = \frac{\langle (L^t P^t w_{target})^t (L^t P^t w_{test}) \rangle}{\sqrt{(L^t P^t w_{target})^t (L^t P^t w_{target})} \sqrt{(L^t P^t w_{test})^t (L^t P^t w_{test})}} \tag{7}$$

Another advantage of this modeling in speaker recognition is that no target speaker enrollment step is required, because the target and the test i-vectors are estimated exactly in the same manner without extra process between estimating the target and the test i-vectors. In addition, this method can make the modeling and scoring process faster with less complexities than the other classifiers and can also make the score normalization like ZT-Norm very fast [5].

3 Multi-resolution time frequency (MRTF) feature

In this section, we will detail the proposed novel multi-resolution time frequency feature extraction method. At first, we briefly introduce the time frequency cepstrum matrix analysis proposed in [24].

3.1 Time frequency cepstrum matrix analysis

Supposing c_i represents the i -th frame basic cepstrum vector, the cepstrum matrix X_i can be given as:

$$X_i = [c_i \ c_{i+1} \ \cdots \ c_{i+(M-1)}] \tag{8}$$

where M is the context width. TFC tries to extract more context related information from a block of the cepstrum matrix and further remove the correlation between elements.

Because of TFC based on the MFCC basic cepstral feature vector, only a 1-Dimensional DCT in the temporal (horizontal) direction is used. Letting C denote the DCT transform matrix, the cepstrum matrix X_i can be decorrelated by

$$Y_i = X_i C^t \tag{9}$$

After this operation, most of the variability in X_i will be concentrated in the coefficients in the upper left part of Y_i . By scanning the matrix Y_i in zigzag order, written as

$$y_i = \text{zigzag}(Y_i) \tag{10}$$

the upper left components corresponds to the lower index, and the lower frequency can then be truncated to D -dimensional vector y_i .

3.2 Multi-resolution time frequency (MRTF) feature extraction

The method is different from the method in [24], the MFTF method completes the operation directly in the spectrogram matrix and is more similar to compression tasks in image processing, in which the 2-Dimensional DCT is often used to decorrelate and reduce the dimensionality, after using the multi-resolution time frequency window function. Suppose the i -th is X

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1M} \\ x_{21} & x_{22} & \cdots & x_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{NM} \end{bmatrix} \tag{11}$$

where M is the context width and the subscript of i has been omitted for simplicity. We split the matrix to multi-resolution sub matrix in a different scale, shown as B'_1, B'_2, \dots, B'_R , and where R is the scale number. As a simple example, through a 2×2 block operation by a 2×2 window function, B' can be as

$$B = \text{TFBW}(X) = \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} \tag{12}$$

The TFBW mean the time frequency block window function operation on the spectrogram matrix.

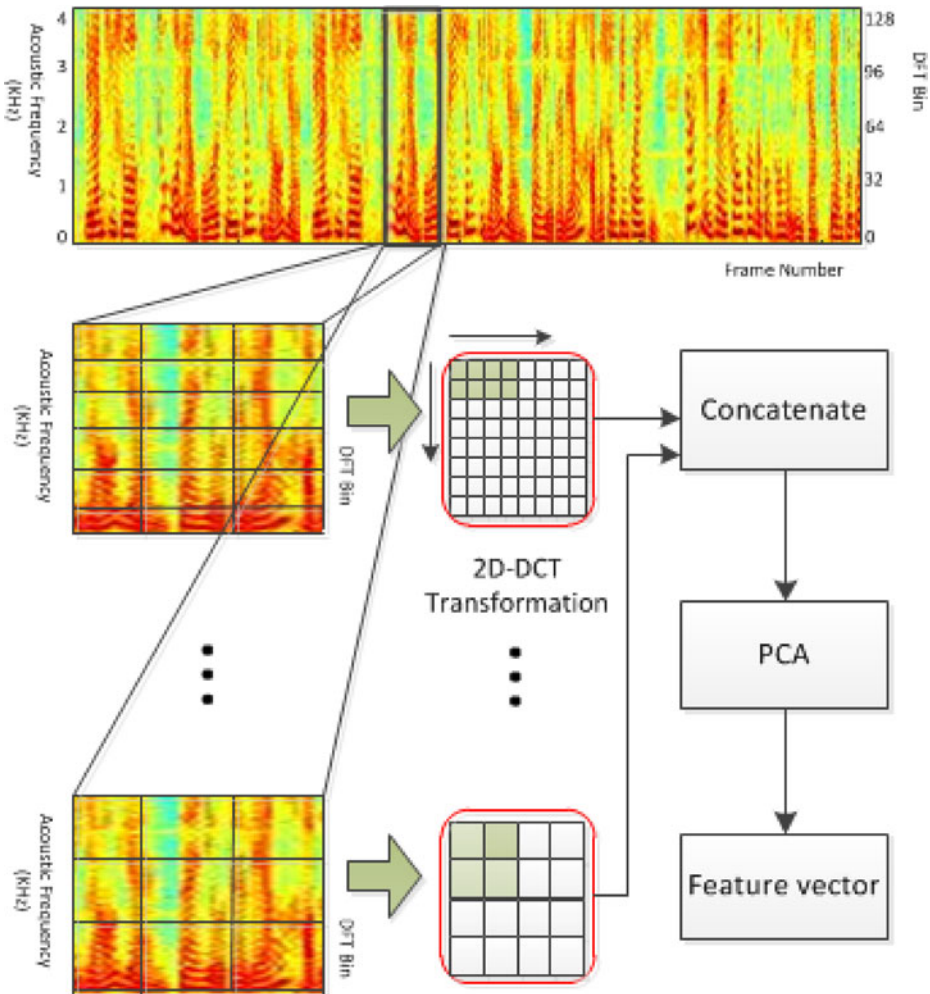


Fig. 1 Framework of MRTF feature extraction

With similar block operation, we can obtain different scaled multi-resolution matrix. Next for each B'_j , do the 2-dimensional DCT as

$$Y'_j = \text{zigzag}(C_j^v B'_j C_j^h) \tag{13}$$

where C_j^v and C_j^h are the vertical and horizontal DCT transform matrix of B'_j , respectively. After this, we obtain a vector by concatenating each Y'_j vector as

$$Y = \begin{bmatrix} Y'_1 \\ Y'_2 \\ \vdots \\ Y'_R \end{bmatrix} \tag{14}$$

Then, we make use of the principal component analysis (PCA) to reduce the dimension and decorrelate the elements between the dimensions. The framework of MRTF extraction method is shown in Fig. 1.

In MRTF, there are several control parameters that need to be considered, such as the context width and the window function. Next we will analyze in detail and select the optimal value for each parameter. For simplicity and speed, we select the value in term of the performance of the GMM-UBM system on male SRE 2008 short2-short3 tel-tel dataset. The dataset used to train UBM was the male SRE 2004 1-side dataset.

At first, we check and choose the context width M for a single resolution $M*32$ using the Mel-FBank time-frequency windowing and the performance is show in Fig. 2.

As show in Figure, we choose the context width $M = 8$ and fix this value to following analysis. After choosing the fixed context width $M (M = 8)$, we compare the different time-frequency window function: Linear-FBank windowing, Mel-FBank window function, Triangle window function, Hamming window function and the performance is shown in Fig. 3.

Fig. 2 Performance comparison of features under different context widths

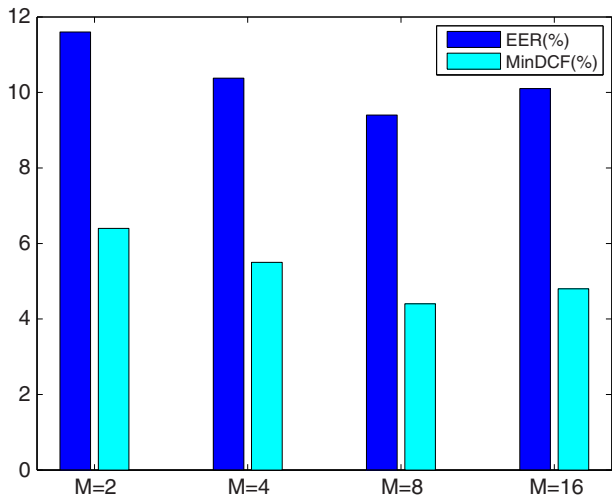
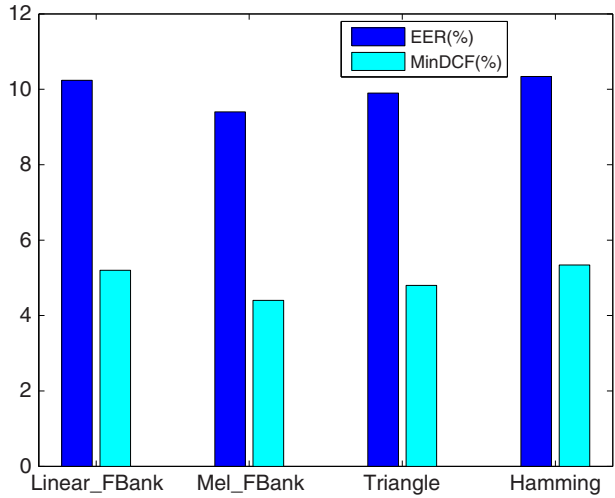


Fig. 3 Performance comparison of features using different window functions

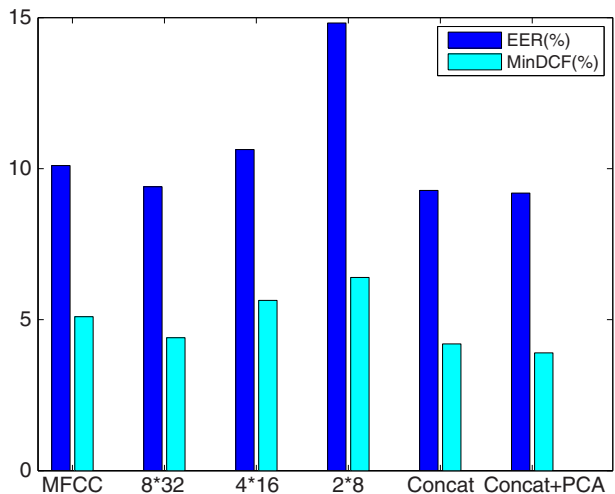


The above result demonstrates that the Mel-FBank window function provides a best performance than all the other functions. Next, we check the contribution of three single different resolutions, which are 8*32, 4*16, 2*8, and their concatenated feature as shown in Fig. 4.

We can see that all single resolution feature contains more or less speaker information and the concatenated feature can integrate the information and obtain the better performance.

In order to further reduce the dimensionality of concatenated MRTF feature, we evaluate the performance of concatenate MRTF feature followed by PCA. For comparable to MFCC, dimension of MRTF after PCA is set to 39 with accounting

Fig. 4 Performance comparison of MFCC, different single resolution features, MRTF without PCA, and MRTF following by PCA



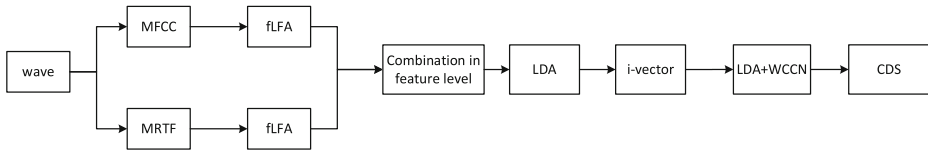


Fig. 5 Framework of complementary combination in feature level

for roughly 90 %. We can see that PCA can not only reduce the dimension but also improve the performance slightly and the final feature performs better than the traditional MFCC feature as also shown in Fig. 4.

4 Three complementary combination strategies

From an information-theoretic viewpoint, combination or fusion between different complementary features or systems can integrate more information for classification and improve the recognition performance. It has have been proved to be effective in our several previous works [14, 15]. In this paper, we will extend our research and propose three complementary combination strategies between the traditional MFCC and the proposed MRTF, which are in feature level, in i-vector level and in score level, and then evaluate their performance as follows, respectively. A combination in different level means there is a different tradeoff between the compute cost and the efficiency of information sharing.

4.1 Complementary combination in feature level

In this section, we propose a complementary combination in the acoustic feature level for the traditional MFCC and our proposed MRTF feature. To simplify the combination, we recommend using the same voice activity detection to promise that all varieties of features have the same frame number. As following, a feature-domain channel compensation method (fLFA) can be adopted. In a GMM-UBM based system, the fLFA has been proved to be a very useful feature-domain channel compensation technology for improving the performance [14, 21] and the effectiveness in i-vector based system [15]. Then correspondence features of each frame are concatenated. To delimit the redundant information for classification,

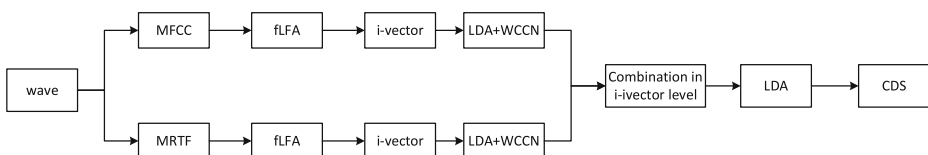


Fig. 6 Framework of complementary combination in i-vector level

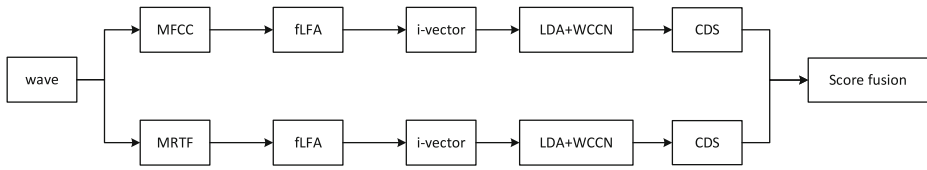


Fig. 7 Framework of complementary combination in score level

LDA is used in the following. The framework of the complementary combination strategy in the feature level can be shows as in Fig. 5.

4.2 Complementary combination in i-vector level

In this section, we also propose a complementary combination in i-vector level for the traditional MFCC and our proposed MRTF feature. Different from the above method in Section 4.1, the method of this section works in the i-vector level. Similarly, i-vectors are all extracted from the different acoustic features, which are also compensated by flFA. And then we concatenate the multiple i-vectors in i-vector level. In order to reduce the dimension and remove the useless information, we apply supervised LDA to the concatenated i-vector. The framework of the complementary combination strategy in i-vector level can be shows as in Fig. 6.

4.3 Complementary combination in score level

In addition, we also describe the combination strategy in score level. Without using the development data to train the fusion coefficient, we use only the equal weight to calculate the score fusion. The framework of the complementary combination strategy in score level can be shows as in Fig. 7.

5 Experiments

5.1 Databases

In this paper, our experiments are based only on telephone data for both training and testing. All experiments were carried out in short2-10sec and 10sec-10sec two short utterance conditions of the NIST 2008 SRE and also in core-core long utterance condition named short2-short3 and we chose male subsets first and female subsets lastly. In all three conditions, we all have only one telephone conversation to enroll the target model and one telephone speech to verify the identity of the speaker. The short2-10sec and 10sec-10sec male corpus 648eakers and 7799 test files and the short2-short3 corpus contain also 648 speakers but 12922 test files. For short2 and short3 data, a five-minute telephone conversation recording is available containing roughly two minutes of speech for enrolling the target speaker model or testing, while

for the 10sec segment, a 10-second telephone speech segment for enrolling the target speaker or for testing was used.

In the NIST evaluation protocol [18], we can use all previous NIST evaluation data and also other corpus to train our systems. For this purpose, we used all the following corpus to estimate our system hyperparameters:

- Switchboard: Switchboard II, Phase 2 and 3. Switchboard II Cellular, Part 1 and 2.
- NIST2004 : NIST 2004 telephone data of Speaker recognition evaluation.
- NIST2005 : NIST 2005 telephone data of Speaker recognition evaluation.
- NIST2006 : NIST 2006 telephone data of Speaker recognition evaluation.

The configure of data for training is as shown in Table 1.

5.2 Experimental setup

In our experiments, both the proposed MRTF and MFCC features are extracted using a 20 ms Hamming window and 10 ms frame shift. In MFCC, 12 Mel Frequency Cepstral Coefficients (MFCC) together with log energy were calculated and this 13-dimensional feature vector was subjected to feature warping [19] using a 3 s sliding window. Delta and delta-delta coefficients were then calculated to produce a final 39-dimensional feature vectors. And in the MRTF, 39 dimensional final feature vectors were extracted with the configuration as show in Section 3.2. We used gender-dependent UBMs containing 2048 Gaussians.

We used 400 total factors defined by the total variability matrix T and the decision scores obtained by i-vector scoring were normalized using ZT-norm. These data are random select from the same dataset as the lambda T training. Table 1 summarizes all corpora that are used to estimate the UBM, total variability matrix T , LDA, WCCN, ZT-norm. We also used different datasets to estimate both LDA and WCCN matrices. As is said in [3] that LDA tries to model between speaker variability, so adding more speaker will definitely help to improve the performance while WCCN models the channel so there is an advantage of using only NIST SRE datasets, because of containing several speakers that talk simultaneously in the different session channels.

The detection task is completed and evaluated with equal error rate (EER) signed by black circle in detection error tradeoff (DET) plot and minimum detected cost function (MinDCF) signed by red circle in DET plot [18].

Table 1 Data corpora used to estimate the UBM, total variability matrix (T), LDA, WCCN, zt-NORM

	Switchboard	NIST 2004	NIST 2005	NIST 2006
UBM	✓	✓	✓	✓
T	✓	✓	✓	✓
LDA	✓	✓	✓	✓
WCCN		✓	✓	✓
zt-norm		✓	✓	✓

Table 2 The results are given as EER and minDCF of MFCC, and MRTF on the trials of three male datasets of the NIST 2008 SRE

Dataset	Result in [3]		MFCC		MRTF	
	EER (%)	minDCF	EER (%)	minDCF	EER (%)	minDCF
short2-short3	4.48	0.024	3.64	0.020	3.53	0.018
short2-10sec	7.38	0.036	6.63	0.035	6.23	0.031
10sec-10sec	14.44	0.063	12.65	0.058	12.04	0.057

Bold entries shows the best result of the experiments

Table 3 Performance comparison of proposed three combination strategies on the trials of three male datasets of the NIST 2008 SRE

Dataset	Combined in feature		Combined in i-vector		Combined in score	
	EER (%)	minDCF	EER (%)	minDCF	EER (%)	minDCF
short2-short3	3.42	0.018	3.34	0.018	3.40	0.018
short2-10sec	6.03	0.032	5.69	0.031	5.90	0.032
10sec-10sec	11.90	0.057	11.32	0.054	11.90	0.056

Bold entries shows the best result of the experiments

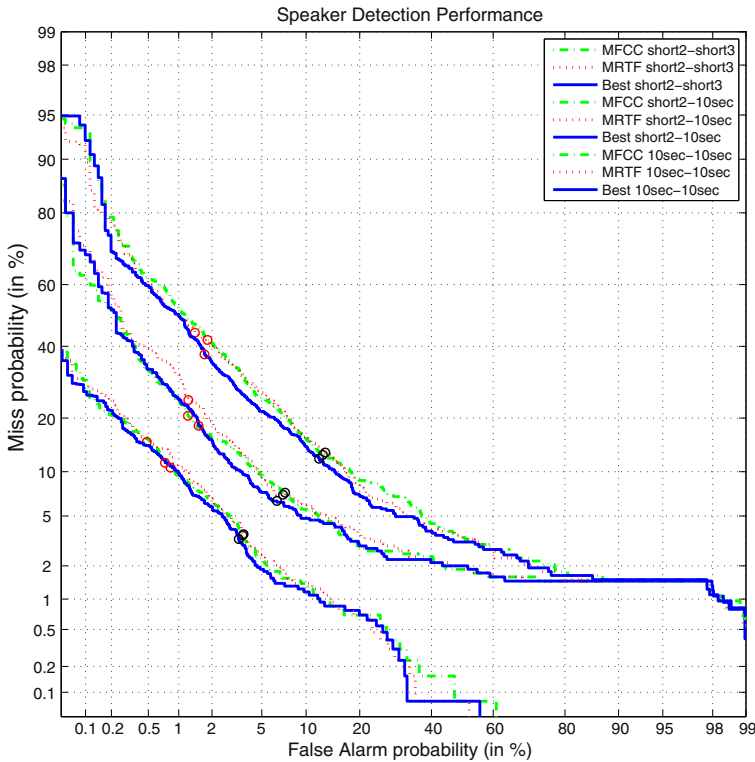


Fig. 8 DET performance curve of MFCC, MRTF and the best combination method under three male conditions

Table 4 Performance comparison of three combination strategies on the trials of three female datasets of the NIST 2008 SRE

Dataset	Result in [3]		MFCC		MRTF		Combined in i-vector	
	EER (%)	minDCF	EER (%)	minDCF	EER (%)	minDCF	EER (%)	minDCF
short2-short3	5.76	0.032	5.43	0.031	5.36	0.030	5.24	0.030
short2-10sec	9.59	0.050	8.77	0.048	8.75	0.047	8.06	0.047
10sec-10sec	16.59	0.072	14.94	0.068	14.59	0.064	13.71	0.062

Bold entries shows the best result of the experiments

5.3 Experimental results

5.3.1 Performance of MRTF

The experiments were carried out on the short2-short3 (core condition), short2-10sec and 10sec-10sec conditions of the NIST 2008 male SRE dataset. The experiments carried out in this section compare the results obtained with the traditional MFCC and the proposed MRTF in three condition. We obtain a better result than the result reported in [3]. We found that the MRTF feature obtains a better performance than the traditional MFCC as shown in Table 2.

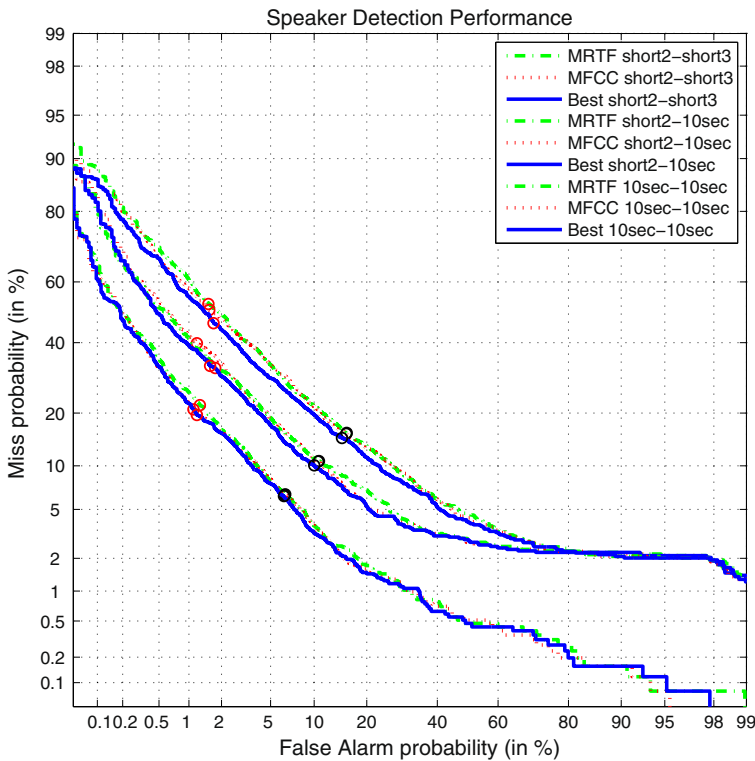


Fig. 9 DET performance curve of MFCC, MRTF and the best combination method under three female conditions

The results given in this table show that the proposed MRTF definitively gave the better results in all conditions of the NIST evaluation compared to traditional MFCC. In 10sec-10sec condition, the performance achieve a relative 5 % improvement in EER. In short2-10sec condition, the performance achieve about 6 % improvement in EER. The experiments also demonstrate the proposed method also effective in long utterance with 3 % relative improvement in EER in short2-short3 condition, even though the gains not better than in short utterance.

5.3.2 Performance comparison of three combination strategies

Table 3 presents the results obtained in three proposed combination strategies, which are in the feature level, in the i-vector level and in the score level. And the results were also in three conditions.

The result in Table 3 reveals that the complementary combination in the i-vector level achieves better results than the two other methods. Compared to the result show in Tabel 2, we obtained around a 10 % improvement of the EER in the 10sec-10sec and a 14 % improvement of the EER in short2-10sec improvement than MFCC. The result also gives an absolutely 3 % improvement in the 10sec-10sec condition, compared to the result in [3], which is the best reported results. The result gives the 8 % improvement of the EER in the short2-short3 condition, showing its an effective complement to present methods. The DET performance curve of MFCC, MRTF and the best i-vector combination method is shown in Fig. 8.

Table 4 also reveals the performance comparison of the Results in [3] and the performance of MFCC, proposed MRTF and the best combination method in the i-vector level on the female datasets. The results shows that we can also obtained the consistently better improvement of the EER than MFCC under three conditions. The DET performance curve of MFCC, MRTF and the best i-vector combination method is shown in Fig. 9.

6 Conclusion

Inspired by multiple different scales analysis, especially of short utterance, this paper proposes a novel multi-resolution time frequency feature (shortened as MRTF) extraction method, which is obtained by performing a multi-scaled 2-D DCT operation on the temporal spectrogram matrix block matrix and combining and selecting multi-scale transformed elements. Compared to the traditional MFCC feature, the proposed method can make better use of multi-resolution time-frequency information. Beyond this, we also proposed three complementary combination strategies, which are: in feature level, in i-vector level and in score level, and compared their performance. We found the best results are obtained by combination in the i-vector level. In the three NIST 2008 Speaker Recognition male Evaluation datasets, the proposed method is more effective for improving the performance under short utterance than in a long utterance. And after the combination, we can achieve an EER of 11.32 % and MinDCF of 0.054 in the 10sec-10sec trials, which is an absolute 3 % improvement of the EER than the best reported result in this field.

Acknowledgements This work is supported by National Natural Science Foundation of China (Project 61370034 , 61273268, 61005019, 90920302) and by Beijing Natural Science Foundation Program (Project KZ201110005005).

References

1. Ajmera P, Holambe R (2009) Multiresolution features based polynomial kernel discriminant analysis for speaker recognition. In: International conference on advances in computing, control, telecommunication technologies, 2009, ACT '09, pp 333–337
2. Dehak N (2009) Discriminative and generative approaches for long- and short-term speaker characteristics modeling: application to speaker verification. Ph.D. thesis, École de Technologie Supérieure, Montreal
3. Dehak N, Kenny P, Dehak R, Dumouchel P, Ouellet P (2011) Front-end factor analysis for speaker verification. *IEEE Trans Audio Speech Language Proc* 19(4):788–798. doi:10.1109/TASL.2010.2064307
4. Dehak N, Kenny P, da Dehak R, Glembek O, Dumouchel P, Burget L, Hubeika V, Castaldo F (2009) Support vector machines and joint factor analysis for speaker verification. In: ICASSP, pp 4237–4240
5. Dehak N, Dehak R, Glass J, Reynolds D, Kenny P (2010) Cosine similarity scoring without score normalization techniques. In: Odyssey 2010—the speaker and language recognition workshop
6. Hatch AO, Kajarekar S, Stolcke A (2006) Within-class covariance normalization for svm-based speaker recognition. In: Ninth international conference on spoken language processing
7. Impedovo D, Pirlo G, Petrone M (2012) A multi-resolution multi-classifier system for speaker verification. *Expert Syst* 29(5):442–455
8. Jayanna H, Mahadeva Prasanna S (2010) Limited data speaker identification. *Sadhana* 35(5):525–546
9. Kanagasundaram A, Vogt R, Dean DB, Sridharan S, Mason MW (2011) I-vector based speaker recognition on short utterances. In: Interspeech 2011, International Speech Communication Association (ISCA), Firenze Fiera, Florence, pp 2341–2344
10. Kenny P, Boulianne G, Dumouchel P (2005) Eigenvoice modeling with sparse training data. *IEEE Trans Audio Speech Language Proc* 13(3):345–354
11. Kenny P, Boulianne G, Ouellet P, Dumouchel P (2007) Joint factor analysis versus eigenchannels in speaker recognition. *IEEE Trans Audio Speech Language Proc* 15(4):1435–1447
12. Kenny P, Boulianne G, Ouellet P, Dumouchel P (2007) Speaker and session variability in gmm-based speaker verification. *IEEE Trans Audio Speech Language Proc* 15(4):1448–1460
13. Kinnunen T, Li H (2010) An overview of text-independent speaker recognition: from features to supervectors. *Speech Comm* 52(1):12–40. doi:10.1016/j.specom.2009.08.009. <http://www.sciencedirect.com/science/article/pii/S0167639309001289>
14. Li ZY, He L, Zhang WQ, Liu J (2010) Multi-feature combination for speaker recognition. In: 2010 7th International Symposium on Chinese Spoken Language Processing (ISCSLP). IEEE, pp 318–321
15. Li ZY, Zhang WQ, He L, Liu J (2012) Complementary combination in i-vector level for language recognition. In: Odyssey 2012—the speaker and language recognition workshop
16. Martinez AM, Kak AC (2001) Pca versus lda. *IEEE Trans Pattern Anal Mach Intell* 23(2):228–233
17. McLaren M, Vogt R, Baker B, Sridharan S, Sridharan S (2010) Experiments in svm-based speaker verification using short utterances. In: Odyssey 2012—the speaker and language recognition workshop
18. NIST (2008) The nist year 2008 speaker recognition evaluation plan. National Institute of Standards and Technology
19. Pelecanos J, Sridharan S (2001) Feature warping for robust speaker verification. In: Odyssey 2001—the speaker and language recognition workshop
20. Reynolds DA, Quatieri TF, Dunn RB (2000) Speaker verification using adapted gaussian mixture models. *Digital Sign Proc* 10(1–3):19–41
21. Shan Y, Liu J (2011) Robust speaker recognition in cross-channel condition based on gaussian mixture model. *Multimed Tools Appl* 52(1):159–173
22. Stafylakis T, Kenny P, Senoussaoui M, Dumouchel P (2012) Plda using gaussian restricted boltzmann machines with application to speaker verification. In: INTERSPEECH

23. Zhang WQ, Deng Y, He L, Liu J (2010) Variant time-frequency cepstral features for speaker recognition. In: INTERSPEECH, pp 2122–2125
24. Zhang WQ, He L, Deng Y, Liu J, Johnson M (2011) Time frequency cepstral features and heteroscedastic linear discriminant analysis for language recognition. *IEEE Trans Audio Speech Language Proc* 19(2):266–276



Zhi-Yi Li Ph.D. candidate in the Department of Electronic Engineering, Tsinghua University. His main research interest covers speaker recognition and language recognition. Corresponding author of this paper.



Wei-Qiang Zhang assistant professor in the Department of Electronic Engineering, Tsinghua University. His main research interest covers speaker recognition and language recognition.



Jia Liu Professor in the Department of Electronic Engineering, Tsinghua University. His main research interest covers speech recognition and signal processing.