Yan *et al. EURASIP Journal on Audio, Speech, and Music Processing* 2014, **2014**:40
http://asmp.eurasipjournals.com/content/2014/1/40

EURASIP Journal on Audio, Speech, and Music Processing
a SpringerOpen Journal

RESEARCH Open Access

A sub-band-based feature reconstruction approach for robust speaker recognition

Furong Yan*, Yanbin Zhang and Jiachang Yan

Abstract

Although the field of automatic speaker or speech recognition has been extensively studied over the past decades, the lack of robustness has remained a major challenge. The missing data technique (MDT) is a promising approach. However, its performance depends on the correlation across frequency bands. This paper presents a new reconstruction method for feature enhancement based on the trait. In this paper, the degree of concentration across frequency bands is measured with principal component analysis (PCA). Through theoretical analysis and experimental results, it is found that the correlation of the feature vector extracted from the sub-band (SB) is much stronger than the ones extracted from the full-band (FB). Thus, rather than dealing with the spectral features as a whole, this paper splits full-band into sub-bands and then individually reconstructs spectral features extracted from each SB based on MDT. At the end, those constructed features from all sub-bands will be recombined to yield the conventional mel-frequency cepstral coefficient (MFCC) for recognition experiments. The 2-sub-band reconstruction approach is evaluated in speaker recognition system. The results show that the proposed approach outperforms full-band reconstruction in terms of recognition performance in all noise conditions. Finally, we particularly discuss the optimal selection of frequency division ways for the recognition task. When FB is divided into much more sub-bands, some of the correlations across frequency channels are lost. Consequently, efficient division ways need to be investigated to perform further recognition performance.

Keywords: Robustness; Missing data technique (MDT); Reconstruction; Sub-band (SB); Full-band (FB); Principal component analysis (PCA)

1 Introduction

The performance of speaker or speech recognition systems degrades rapidly when they operate under conditions that differ from those used for training. Therefore, accomplishing noise robustness is a key issue to make these systems deployable in real world conditions. Solutions have been presented to solve this issue, such as feature-based [1-3], score-based [4,5], model-based [6-8], i-vectors [9], and the missing data technique (MDT) [10-12].

MDT can compensate for disturbances of the arbitrated type, so that this method which is based on the time-frequency representation is suitable to the problem of noise mismatch [12].

In MDT, two different methods have been considered to perform speech or speaker recognition with incomplete data: marginalization [13-15] and reconstruction [16,17]. In marginalization, the unreliable components are discarded or integrated up to the observed values. While the reconstruction method involves the estimation of the corrupted features using statistical methods, such as minimum mean square error (MMSE) [10], maximum *a posteriori* (MAP), and maximum likelihood (ML). Marginalization [11,14] and reconstruction [10] have been applied in speaker recognition system. However, marginalization suffers from two main drawbacks [17,18]. First, as known to us, utterance-level processing, such as mean and variance normalization, is capable of improving the recognition performance, but it cannot be performed with an incomplete spectrum [18]. Second, recognition has been carried out with spectral features. However, it is well known that cepstral features outperform spectral

*Correspondence: yfrhappy@126.com
Beijing University of Posts and Telecommunications, No.10 Xitucheng Road, Beijing 100876, China

ones. Moreover, of all the methods, marginalization is assumed to have the most overhead. Consequently, if the complete reconstructed spectrogram is available, the recognizer is no longer constrained to perform recognition using spectral features. A more optimal set of parameters from the reconstructed spectrum will be derived.

In this paper, MAP reconstruction method [10] is used. Its efficiency significantly depends on the correlation between the spectral features. Conventional MAP reconstruction method is conducted on full-band [18,19]. According to our analysis, the spectral vectors extracted from the sub-band have more relevance than the ones extracted from the full-band. The conclusion will be illustrated in Section 2. Based on the above theory and the sub-band idea [20-22], a multi-sub-band reconstruction approach is proposed to improve on the recognition performance. The principle is to divide the full-band into multiple sub-bands and then independently reconstruct missing features extracted from every sub-band. After that, those features from all sub-bands will be recombined to yield the typical mel-frequency cepstral coefficient (MFCC) vector.

As one of many feature enhancement methods, the proposed reconstruction approach can be used in speaker and speech recognition system. To evaluate its validity, this paper will combine the new reconstruction method with speaker recognition system.

This paper is organized as follows. In the next section, the theory of the proposed reconstruction approach is analyzed. Section 3 is devoted to describing the proposed reconstruction approach. Section 4 describes the baseline experiment system and the experimental framework which is adopted to evaluate the proposed technique. Finally, Section 5 concludes this paper and discusses some future directions.

2 The analysis of concentration

As we know, the more concentrated the feature vector is, the higher its redundancy is, that is, the greater its correlation is [23]. It is measured by the degree of concentration with principal component analysis (PCA) .

In this paper, the P-dimensional mel log-spectral vector is used for reconstruction. Mel filters are used to represent a frame spectrum as a log-spectral vector of P-dimensional (termed as full-band feature vector). The frequency region $(0, f_s/2)$ is divided into C sub-bands. Let P_i denote the number of mel filters corresponding to the ith sub-band. Apparently,

$$\sum_{i=1}^{C} P_i = P \tag{1}$$

Corresponding to the tth frame and ith sub-band, the output of mel filters (termed as the ith sub-band feature vector) is represented as follows:

$$\overrightarrow{Y}_i^t = (y(1,t), \cdots y(P_i,t))^T \tag{2}$$

In order to analyze the degree of concentration of the feature vector $\overrightarrow{Y}_i^t$, the eigenvalues of associated covariance matrix Θ_i need to be calculated and then need to be arranged in descending order. It is represented as $\left[\lambda_{i,1}, \lambda_{i,2}, \cdots \lambda_{i,P_i}\right]$.

To learn how closely the ith sub-band feature vector $\overrightarrow{Y}_i^t$ is in the space of the P_i-dimension, the so-called concentration level $M_R^i(r)$ is introduced and computed as follows:

$$R_i(m) = \frac{\sum_{i=1}^{m} \lambda_{i,l}}{\sum_{i=1}^{P_i} \lambda_{i,l}}, m = 1, 2, \cdots P_i \tag{3}$$

$$M_R^i(r) = \arg\min_m (R_i(m) > r) \tag{4}$$

That is, $R_i(m)$ is the accumulative contribution rate of the first m principle components. Concentration level $M_R^i(r)$ is the minimum m that makes $R_i(m) > r$, where r is a predefined concentration coefficient.

For certain r, a smaller $M_R^i(r)$ implies that the ith sub-band feature vector is confined along a smaller number of principle directions, and therefore, the feature vector is much more closely related to each other according to the above definition.

In the same manner, the degree of concentration of the full-band feature vector could be analyzed.

The accumulative contribution rate of the first m principle components corresponding to the 4-sub-band and full-band is shown in Figure 1. The conclusion should be clear. The concentration level corresponding to each sub-band in the 4-sub-band is smaller than the one corresponding to the full-band.

The correlation between the redundancy and accuracy of the prediction is best visualized using 2-dimensional examples as shown in Figure 2. The 2-dimensional examples involve the feature vector extracted from clean and noisy utterances, together with MAP reconstruction obtained for the noisy utterance. Babble noise at 0 dB signal-to-noise ratio (SNR) has been added to obtain the noisy utterance. Panels (a) and (b), respectively, reflect a range of 2-dimensional feature vectors with different redundancies. The redundancy of data in panel (b) is lower than that in panel (a). The reconstruction data corresponding to the data with high redundancy and low redundancy is defined along the first principle direction and scattered. In short, the fatter the cloud is, the lower the prediction accuracy is in a 2-dimensional case.

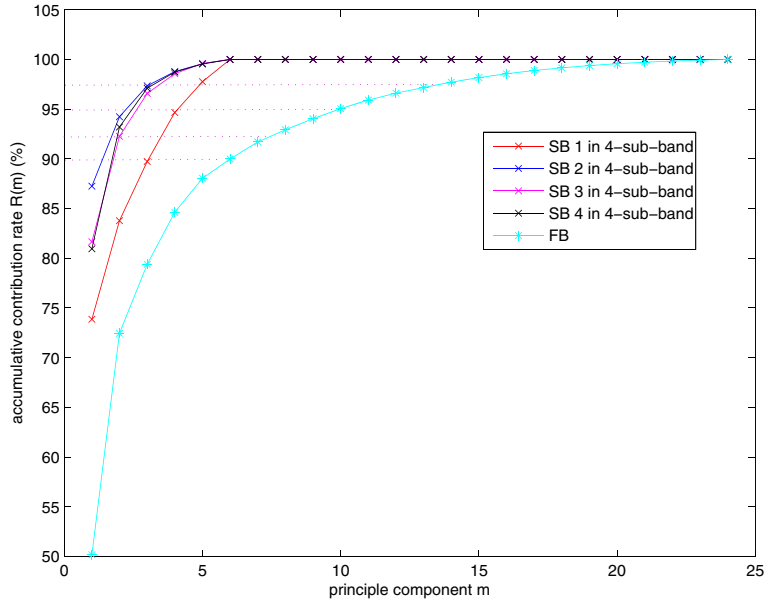

Figure 1 The accumulative contribution rate of the first *m* principle components corresponding to the 4-sub-band and full-band.

Figure 3 shows the contribution rate of two principle components which are obtained from the covariance matrix of the 2-dimensional feature vector. When the value of the predefined concentration coefficient r is 0.9, the concentration level which is corresponding to the data shown in Figure 2a,b is $M_R^{(high)}(r) = 1$ and $M_R^{(low)}(r) = 2$, respectively.

Considering the recorded positions of the 2-dimensional feature vector in Figure 2 and the corresponding contribution rate, together with our analysis, the following conclusion is obtained: the higher the redundancy of the data is, that is, the greater its correlation is, the smaller the corresponding concentration level is. As MAP reconstruction method is based on

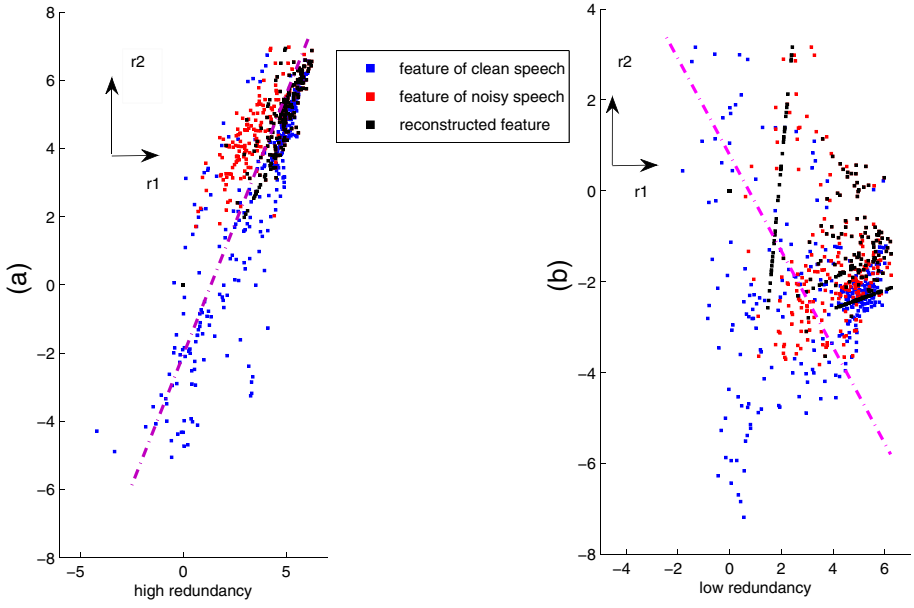

Figure 2 A spectrum of redundancies in data from the two separate recordings r1 and r2. The best-fit line is indicated by the dashed lines for **(a)** high redundancy and **(b)** low redundancy.

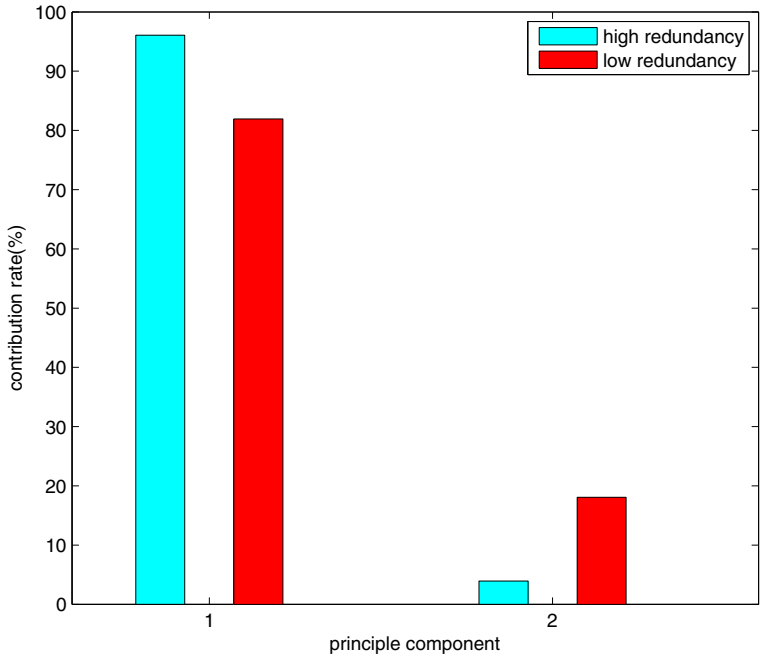

Figure 3 The contribution rate of two principle components in a 2-dimensional case.

the correlation between the feature vectors, the smaller the concentration level is, the higher the validity of the reconstruction is.

3 Multi-sub-band reconstruction for speaker recognition system

As one of many feature enhancement methods, the multi-sub-band reconstruction method in MDT can be applied in the Gaussian mixture model (GMM) [24], the SVM-GMM [25], and the universal background model (UBM)-GMM recognition system. Based on the validity of the UBM-GMM system shown in [11], the proposed reconstruction method is evaluated in a UBM-GMM speaker recognition system. In this section, the MDT-based speaker recognition system is described.

3.1 UBM-GMM model

In this paper, a speaker-independent UBM is used. A speaker-dependent model can be derived from UBM by adapting the UBM parameters to the speech material of the corresponding speaker using MAP estimation [11,26].

3.2 Feature vector

Mel log-spectral vector and MFCC are used in the reconstruction and recognition stage, respectively. The unreliable components are reconstructed based on the statistical relationship between the log-spectral vector.

3.3 Mask estimation

In order to perform MDT, a mask must be required which classifies the time-frequency (T-F) units into reliable and unreliable components. Various strategies have been proposed to estimate a mask, such as SNR-based estimation [27], auditory and perceptual estimation [14,28], classifier-based estimation [29], and DNN-based estimation [30]. It is, however, outside the scope of this paper to analyze and compare all existing approaches. Because the focus of this paper is to robustly identify speakers in the presence of noise, the mask $m(t,k)$ is determined by estimating the local SNR in individual T-F units. SNR-based mask estimation method is applied to decide whether a T-F unit is reliable.

$$m(t,k) = \begin{cases} 0, \text{ if } \left|\widehat{\overrightarrow{S}}(t,k)\right|^2 \leq \left|\widehat{\overrightarrow{N}}(t,k)\right|^2 \\ 1, \text{ otherwise} \end{cases} \quad (5)$$

where $\left|\widehat{\overrightarrow{S}}(t,k)\right|^2$ and $\left|\widehat{\overrightarrow{N}}(t,k)\right|^2$ represent the kth frequency bands of the power spectrum of speech and noise, respectively, in individual T-F units. What calls for special attention is that the estimation of speech and noise components is carried out in the spectral domain before applying mel filter.

The estimate of the noise spectrum is derived from the noisy signal spectrum. The estimation method is shown in [31]. The estimate of the speech spectrum $\left|\widehat{\overrightarrow{S}}(t,k)\right|^2$ can be derived by subtracting the estimated noise spectrum

$\left|\widehat{\overrightarrow{N}}(t,k)\right|^2$ from the corrupted signal spectrum. In this paper, the technique to accomplish this is to perform spectral subtraction by applying an SNR-dependent gain function MMSE log-STSA [32] in the frequency domain.

3.4 MAP estimation for unreliable components

In MAP estimation, the unreliable components are estimated by making their likelihood condition on the reliable components [18] be maximum.

$$\hat{\overrightarrow{x}}_u = \arg\max_{\overrightarrow{x}_u} p\left(\overrightarrow{x}_u|\overrightarrow{x}_r, \overrightarrow{\mu}, \Theta\right) \tag{6}$$

A feature vector $\overrightarrow{x} \in \Re^{P_j*1}$ is divided into reliable and unreliable components based on SNR-based mask estimation method.

$$\overrightarrow{x_r} \in \Re^{D_1*1} \quad \overrightarrow{x_u} \in \Re^{D_2*1}, D_1 + D_2 = P_j \tag{7}$$

$$\overrightarrow{x} = \left[\overrightarrow{x}_r, \overrightarrow{x}_u\right] \tag{8}$$

assuming that $p\left(\overrightarrow{x}; \overrightarrow{\mu}, \Theta\right)$ is the probability distribution function (pdf) of a Gaussian distribution with mean vector μ and covariance matrix Θ. According to the nature of Gaussian distribution, $p\left(\overrightarrow{x}_r; \overrightarrow{\mu}, \Theta\right)$ and $p\left(\overrightarrow{x}_u; \overrightarrow{\mu}, \Theta\right)$ would therefore also be Gaussian [33]. Consequently,

$$\overrightarrow{\mu} = \left[\overrightarrow{\mu}_r, \overrightarrow{\mu}_u\right] \tag{9}$$

$$\Theta = \begin{bmatrix} \Theta_{rr} & \Theta_{ru} \\ \Theta_{ur}, & \Theta_{uu} \end{bmatrix} \tag{10}$$

$$p\left(\overrightarrow{x}_r, \overrightarrow{\mu}_r, \Theta_{rr}\right) = C_1 \exp\left[-0.5\left(\overrightarrow{x}_r - \overrightarrow{\mu}_r\right)^T \Theta_{rr}^{-1}\left(\overrightarrow{x}_r - \overrightarrow{\mu}_r\right)\right] \tag{11}$$

$$p\left(\overrightarrow{x}_u, \overrightarrow{x}_r, \overrightarrow{\mu}, \Theta\right) = C_2 \exp\left[-0.5\left(\overrightarrow{x} - \overrightarrow{\mu}\right)^T \Theta^{-1}\left(\overrightarrow{x} - \overrightarrow{\mu}\right)\right] \tag{12}$$

where Θ_{ru} is the cross covariance between $\overrightarrow{x}_r$ and $\overrightarrow{x}_u$ and $\Theta_{ru} = \Theta_{ur}^T$.

It can now be shown that $p\left(\overrightarrow{x}_u|\overrightarrow{x}_r, \overrightarrow{\mu}, \Theta\right)$ is given by

$$p\left(\overrightarrow{x}_u|\overrightarrow{x}_r, \overrightarrow{\mu}, \Theta\right) = \frac{p(\overrightarrow{x}_u, \overrightarrow{x}_r, \overrightarrow{\mu}, \Theta)}{p(\overrightarrow{x}_r, \overrightarrow{\mu}_r, \Theta_{rr})}$$
$$= C \exp\left[-0.5\left(\overrightarrow{x}_u - \overrightarrow{\mu}_u\right) - \Theta_{ur}\Theta_{rr}^{-1}\left(\overrightarrow{x}_r - \overrightarrow{\mu}_r\right)\right] \tag{13}$$

where C is a normalizing constant. The following equation can be obtained from Equations 11, 12, and 13.

$$\widehat{\overrightarrow{x}}_u = \arg\max_{\overrightarrow{x}_u}\left[p\left(\overrightarrow{x}_u|\overrightarrow{x}_r, \overrightarrow{\mu}, \Theta\right)\right] = \overrightarrow{\mu}_u + \Theta_{ur}\Theta_{rr}^{-1}\left(\overrightarrow{x}_r - \overrightarrow{\mu}_r\right) \tag{14}$$

Figure 4 shows the process of reconstruction. The values of the statistical parameters such as $\overrightarrow{\mu}_r$, $\overrightarrow{\mu}_u$, Θ_{ur}, and Θ_{rr} must be learned from the training corpus. A vector is said to belong to the cluster that is most likely to have generated it. As the distribution of the vector is assumed to be Gaussian, the cluster membership $\widehat{m}_{\overrightarrow{x}(t)}$ of a vector $\overrightarrow{x}(t)$ is defined as

$$\widehat{m}_{\overrightarrow{x}(t)} = \arg\max_m\left[p\left(m|\overrightarrow{x}(t)\right)\right] = \arg\max_m\left[p\left(\overrightarrow{x}(t)|m\right)p(t)\right] \tag{15}$$

and then the unreliable components of the vector are reconstructed using MAP estimation method.

3.5 The proposed multi-sub-band reconstruction approach

Assuming that utilizing P mel filter to smooth the N FFT magnitude coefficients. The reconstruction is individually conducted on 2 sub-bands consisting of consecutive channels (P/2-dimensional channels) with no band overlap (sub-band 1: channel 1 to P/2, sub-band 2: channel P/2+1 to P). The reconstruction method falls neatly into two parts as shown in Figure 5. In the first part, the statistical parameters (SP) used in construction are individually trained for different sub-bands. The steps of the second part are as follows:

(a) The estimation of speech and noise components is carried out in the spectral domain.
(b) A mask will be obtained which classifies the T-F representation into reliable and unreliable components corresponding to the frequency range of P mel filters. The above two steps are carried out before applying the mel filter.
(c) P mel filters are used to smooth the power spectrum and then its logarithm is taken.
(d) The mel log-spectral vector is multiplied by the mask estimated in step (b).
(e) The feature vector corresponding to full-band is divided into ones corresponding to 2 sub-bands.
(f) Based on SP trained in the first part, the feature vectors corresponding to every sub-band are reconstructed, individually.
(g) The reconstructed vector of 2-sub-band is recombined to yield the typical MFCC vector.

3.6 Baseline system

The system described in [11] assumes that the unreliable components are bounded between zero and the observed mel log-spectrum and the mel log-spectrum is independent, and marginalization is applied to process the corrupted vector. The feature vector used in recognition is a P-dimensional mel log-spectrum. We compare the performance of the proposed system with the baseline system.

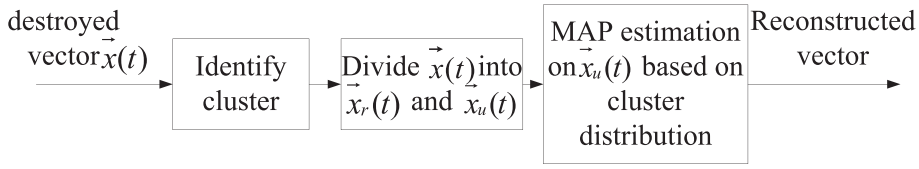

Figure 4 Block diagram presenting estimation of missing components in a vector using MAP estimation method.

4 Experiments

New reconstruction method is evaluated on a closed set of 30 speakers and 140 utterances per speaker. The sampling frequency is 16 KHz. For each speaker, 70% of the available speech material is randomly selected to train the corresponding speaker model, 7% is used for training SP for reconstruction stage, and the remaining 23% is used for test.

In the training stage, we use a voice activity detector (VAD) based on power to ensure that silence frames would not impact on the establishing model.

Speaker recognition performance is evaluated on a subset of ten randomly selected speakers involving a total of 30 sentences per speaker (20 sentences for training speaker-dependent GMM and 10 sentences for testing). In the test phase, utterances are mixed at various SNRs with noise signals drawn from the NOISEX database [34].

Figure 6 describes evaluation system in which 24 mel filters are used to smooth the spectrum and the full-band is divided into 2 sub-bands (SB1: channel 1 to 12, SB2: channel 13 to 24), and 34-demensional MFCC consisting of 16 static MFCC coefficients including the 0th order coefficient and first order temporal derivatives is used for recognition. At the end, cepstral mean normalization (CMN) is applied to improve robustness.

4.1 Experiment 1: performance comparison between marginalization and reconstruction including full-band and 2-sub-band reconstruction

In the first experiment, we compare the performances of two systems which use the marginalization and reconstruction methods to process the corrupted features and then evaluate the validity of the proposed reconstruction method. The point is that recognition has to be carried out with spectral features in the former system. While in the latter system, MFCC are extracted for recognition.

The DET curves visualize the trade-off between missed detections and false alarms [35]. Figure 7 gives the recognition performance of two systems in destroyer-engine noise at a SNR of 0 dB. The results in the figure show that cepstral features outperform spectral ones for speaker recognition.

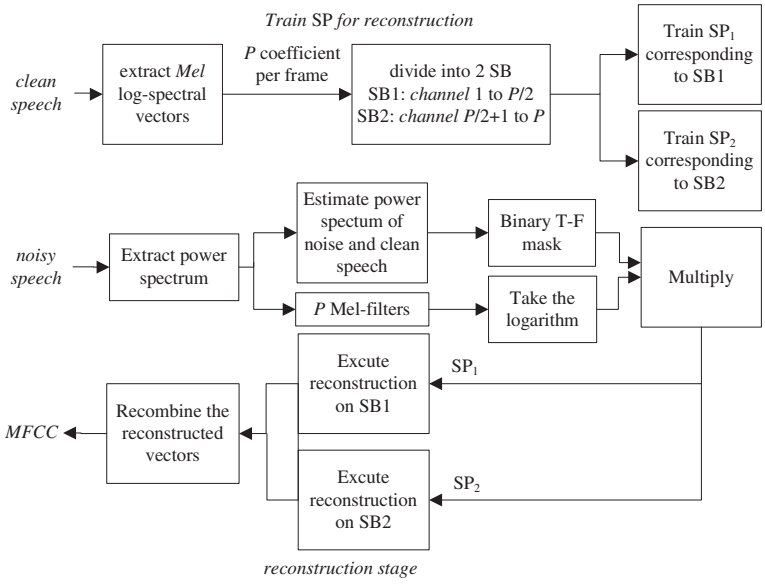

Figure 5 Diagram of the proposed reconstruction method.

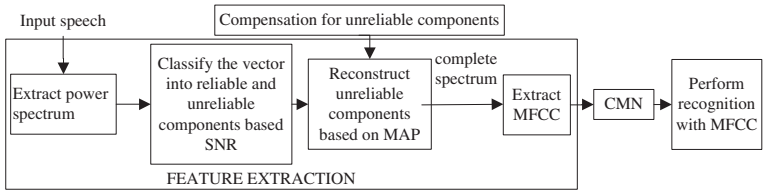

Figure 6 Schematic diagram of the evaluation system.

Figure 8 shows that the recognition performance of the latter system improves when reconstruction is applied to process the corrupted features, and the recognition accuracy of the latter system using 2-sub-band reconstruction method is improved 5.65% more than full-band reconstruction method.

In order to evaluate the validity of the proposed reconstruction method in various noise types, this paper conducts recognition experiments in babble, factory1, pink, white, and destroyer-engine noise. The SNR-dependent recognition accuracy for recognition system is presented in Table 1. The last table depicts the average performance over all noise conditions.

Based on the experimental results reported in Table 1, the corresponding SNR-dependent curves are shown for all noise types in Figures 9, 10, and 11.

The following observations can be made:

(a) It can be observed in Table 1 that the performance obtained from both reconstruction methods clearly outperforms the baseline system.
(b) The results show that 2-sub-band reconstruction method performs better than full-band for all noise types. The recognition performance is higher at a larger SNR.
(c) The recognition performance in babble noise is higher than the other four noise types in most cases for two kinds of reconstruction methods.
(d) The corresponding relative improvements regarding full-band reconstruction are 2.55%, 1.49%, 1.10%, 1.63%, and 1.03% at a SNR of 0, 5, 10, 15, and 20 dB, respectively. Recognition performance improves the most at a SNR of 0 dB.

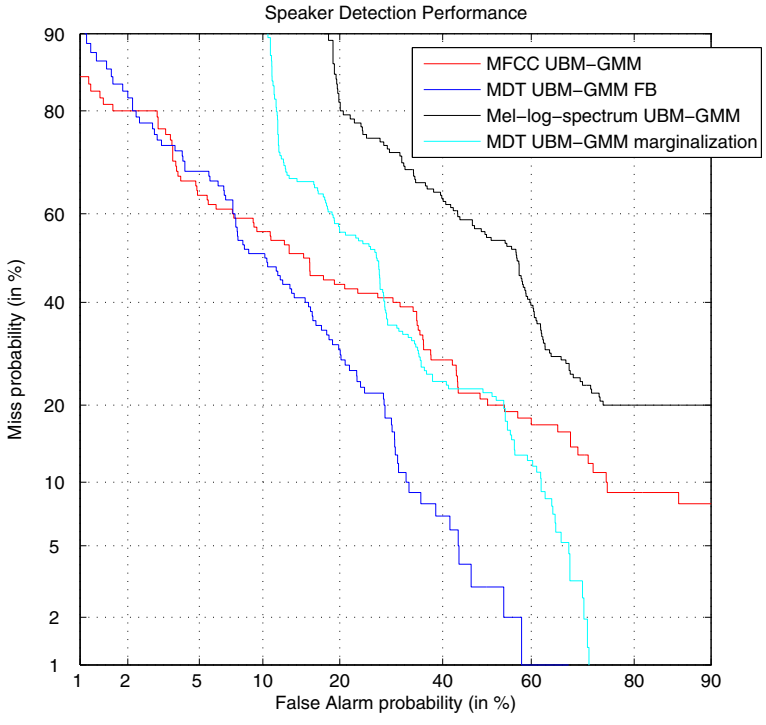

Figure 7 The DET curves for two recognition systems in destroyer-engine noise at a SNR of 0 dB.

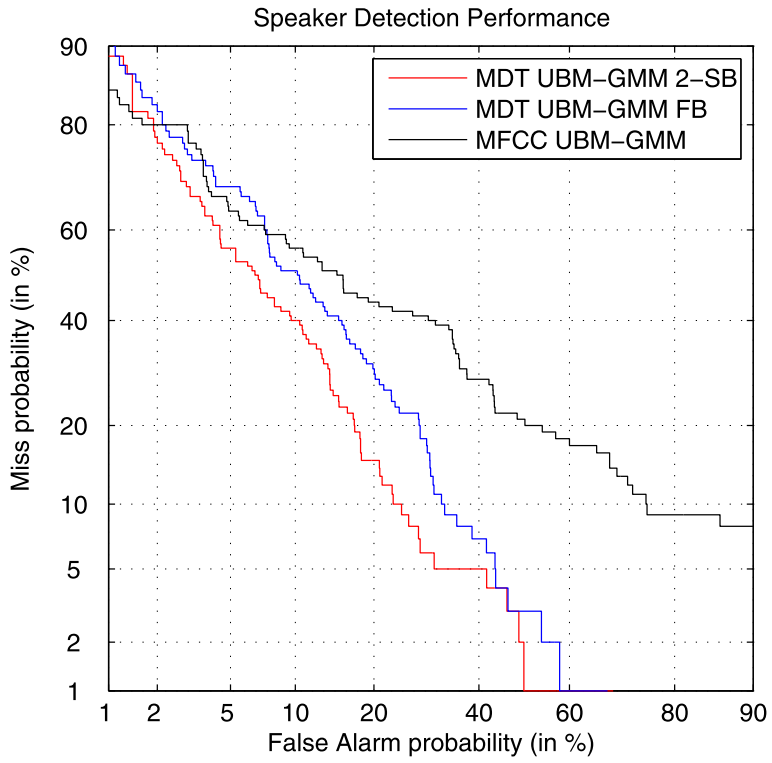

Figure 8 The DET curves for two kinds of reconstruction methods in destroyer-engine noise at a SNR of 0 dB.

Table 1 Recognition performance of FB, 2-SB reconstruction, and marginalization in the presence of different types of noise (unit: %)

		0 dB	5 dB	10 dB	15 dB	20 dB
Babble	FB	82.60	84.02	87.01	87.49	89.49
	2-SB	82.96	85.30	87.84	89.30	91.25
	Marginalization	63.71	64.88	66.44	67.69	69.75
Factory1	FB	76.52	82.10	87.40	88.00	88.33
	2-SB	80.01	82.25	87.70	89.13	90.23
	Marginalization	67.54	68.10	68.33	68.51	69.40
Pink	FB	75.12	80.78	84.83	87.55	89.11
	2-SB	76.66	82.43	87.40	89.62	90.27
	Marginalization	67.79	68.54	69.00	69.06	69.91
White	FB	77.10	83.02	84.40	87.16	89.71
	2-SB	78.81	83.60	86.09	88.70	89.82
	Marginalization	68.00	68.92	70.21	70.77	71.21
Destroyer-engine	FB	76.51	82.81	86.43	86.29	88.47
	2-SB	82.16	86.60	86.52	87.92	88.68
	Marginalization	66.64	67.26	68.30	68.48	69.97
Average	FB	77.57	82.55	86.01	87.30	89.02
	2-SB	80.12	84.04	87.11	88.93	90.05
	Marginalization	66.74	67.54	68.46	68.90	70.05

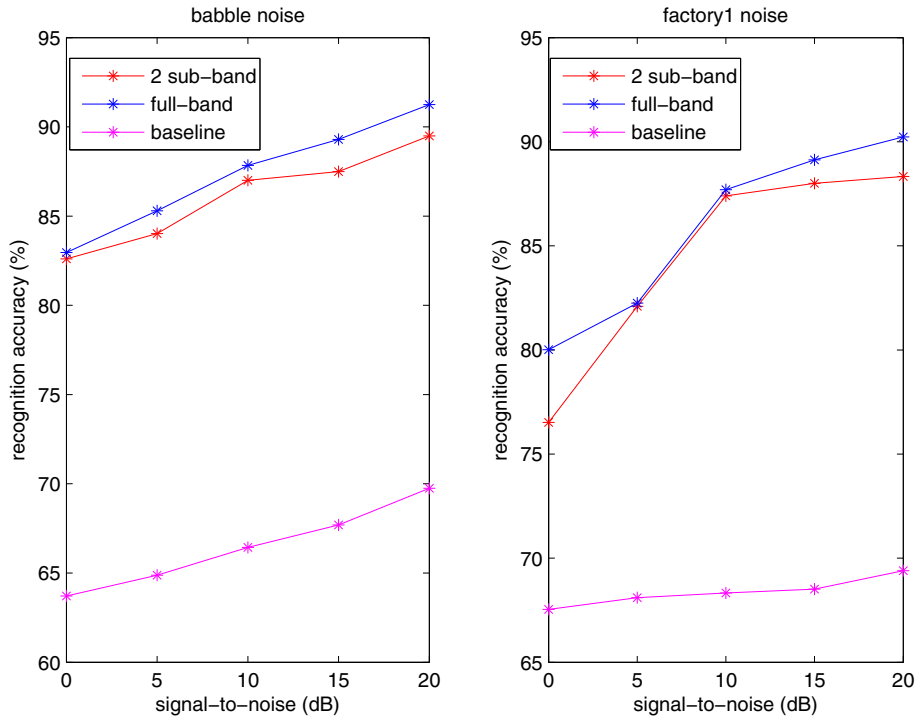

Figure 9 SNR-dependent speaker recognition performance in babble and factory1 noise.

(e) The improved recognition performance is 6.04%, 6.97%, 8.99%, 5.63%, and 11.37% in babble, factory1, pink, white, and destroyer-engine noise, respectively. Recognition performance improves the most in destroyer-engine noise.

We analyze the relationship between reconstruction performance and the correlation of the feature vector with PCA. Table 2 shows the contribution rate of every principle component. When the value of concentration coefficient r is 0.95, the corresponding concentration

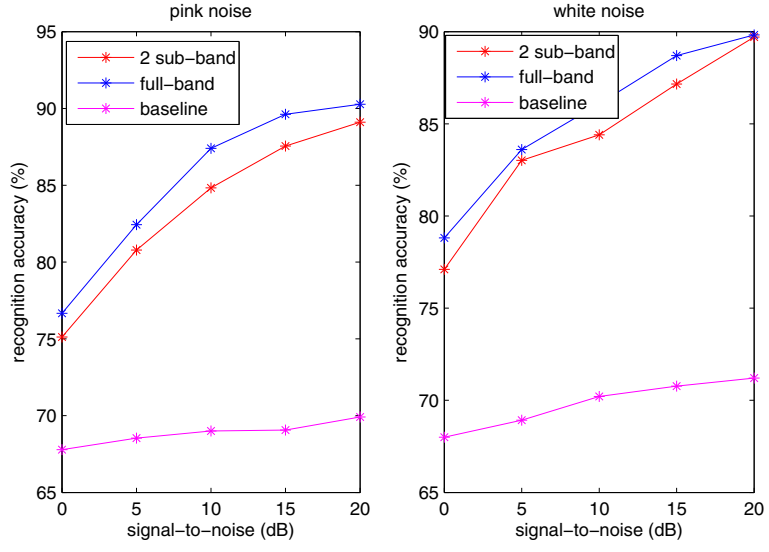

Figure 10 SNR-dependent speaker recognition performance in pink and white noise.

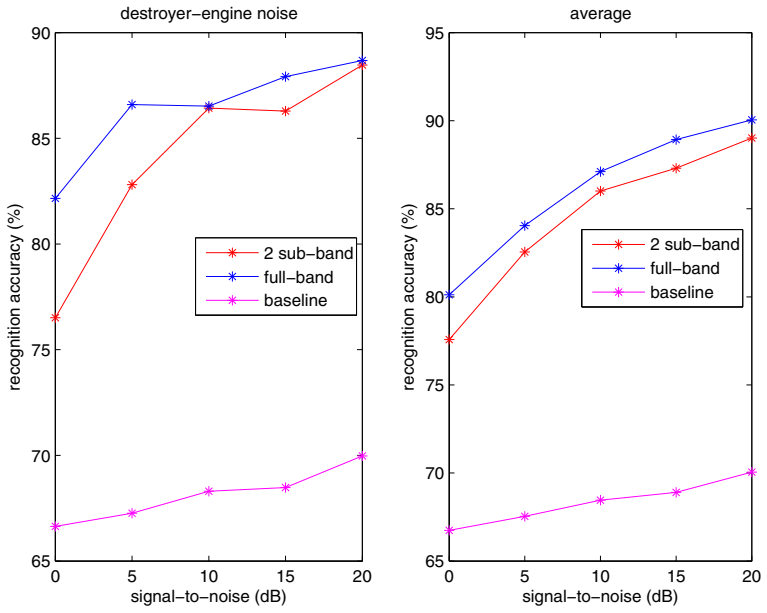

Figure 11 SNR-dependent speaker recognition performance in destroyer-engine noise and average condition.

levels are $M_R^{FB}(r) = 10$, $M_R^1(r) = 6$, and $M_R^2(r) = 5$. Based on the conclusion shown in Section 2, a smaller $M_R^i(r)$ implies a stronger concentration for the feature vector. Consequently, since the correlation of every sub-band is stronger than the full-band, the performance of the 2-sub-band reconstruction approach is better.

The result of PCA will be obtained by decomposing eigenvalues of the covariance matrix, which is relevant to the reconstruction. The accumulative contribution rate of the principle components is shown in Figure 12.

Table 2 The contribution rate (%) of every principle component

	FB	SB1	SB2
1	50.161	69.455	66.811
2	22.291	13.475	17.025
3	6.935	3.292	6.812
4	5.236	2.315	3.084
5	3.424	2.315	2.051
6	1.945	2.139	1.345
7	1.708	1.514	0.951
8	1.218	1.185	0.715
9	1.125	0.898	0.490
10	1.002	0.651	0.327
11	0.869	0.445	0.233
12	0.698	0.240	0.157

4.2 Experiment 2: influence of different division ways of full-band

The conclusion that the recognition performance obtained by the proposed reconstruction method is superior to full-band reconstruction has been obtained in Experiment 1. The choice of an optimal division of full-band seems to be crucial for sub-band reconstruction method. In order to find the optimal division, this paper conducts a series of recognition experiments. The division ways and the corresponding recognition performance are shown in Table 3.

These experiments are conducted in babble noise which is highly non-stationary and a SNR of 0 dB. The results are shown in Figure 13.

The recognition performance is ranked corresponding to different division ways starting with the highest performance: 4 sub-bands, 2 sub-bands, 3 sub-bands, 8 sub-bands, 6 sub-bands, and 12 sub-bands. The relationship between channel number and recognition performance is not obvious. In order to explain the relationship between the recognition performance and the division ways, we analyze the case of 4 sub-band and 2 sub-band. The results are shown in Figure 14. Assume that the amount of information presented by the original data is 100%. When full-band is divided into 4 sub-bands, the amount of information presented by the first four principle components

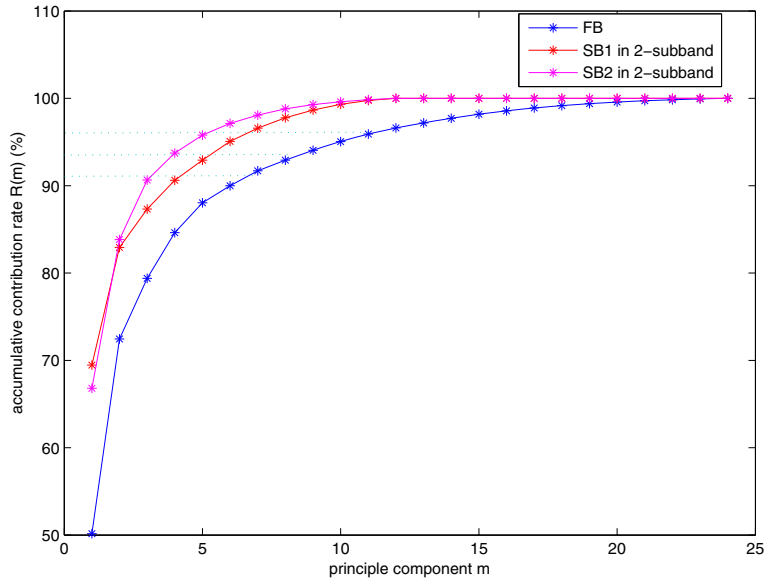

Figure 12 Accumulative contribution rate of principle components derived from feature vectors corresponding to 2 sub-bands and full-band.

derived from sub-band 1, sub-band 2, sub-band 3, and sub-band 4 is 94.67%, 98.79%, 98.58%, and 98.73%, respectively. However, if the full-band is divided into 2 sub-bands, the amount of information presented by the first four principle components derived from both sub-bands is 90.61% and 93.73%. That is, the redundancy of the feature vector extracted from each sub-band is higher on the condition that the full-band is divided into 4 sub-bands.

When the full-band is divided into 12 sub-bands, the recognition performance is inferior. The observation shows that the correlations between the feature vector are lost when the number of sub-bands is more numerous.

5 Conclusions

This paper presents a new feature enhancement method, which is evaluated in a UBM-GMM speaker recognition system. In the proposed method, the reconstruction is executed on a partial sub-band independently and then the reconstructed spectrum is recombined into a complete spectrum to yield the conventional MFCC for recognition. Compared to full-band reconstruction method, recognition performance obtained by the proposed reconstruction approach has been shown to be higher in five noise types. The experiment has also reflected that the recognition performance depends on the frequency division ways, thus the optimal division ways need to be developed.

Table 3 Different division ways of full-band and the corresponding recognition performance

Channel number	Sub-bands	Recognition (%)
12	1-2, 3-4, 5-6, 7-8, 9-10, 11-12,13-14,15-16, 17-18, 19-20, 21-22, 23-24	74.18
8	1-3, 4-6, 7-9, 10-12, 13-15, 16-18, 19-21, 22-24	78.15
6	1-4, 5-8, 9-12, 13-16, 17-20, 21-24	77.83
4	1-6, 7-12, 13-18, 19-24	84.86
3	1-8, 9-16, 17-24	81.78
2	1-12, 13-24	82.96

The first experiment has revealed the following results. First, MFCC features outperform spectral ones for speaker recognition. Second, the recognition performance obtained by reconstruction is higher than marginalization. Third, the recognition performance obtained by the 2-sub-band reconstruction method is superior to the full-band reconstruction in five noise types and at all SNRs. The second experiment has shown that different frequency division ways could influence on the recognition performance.

In order to achieve further recognition performance improvements, on the one hand, an optimal frequency division way will be very important. On the other hand, analyzing the distribution property of various noise types and then accurately identifying destroyed components are also research hot spots. In the end, research on mask estimation algorithms is required to precisely separate reliable from unreliable components.

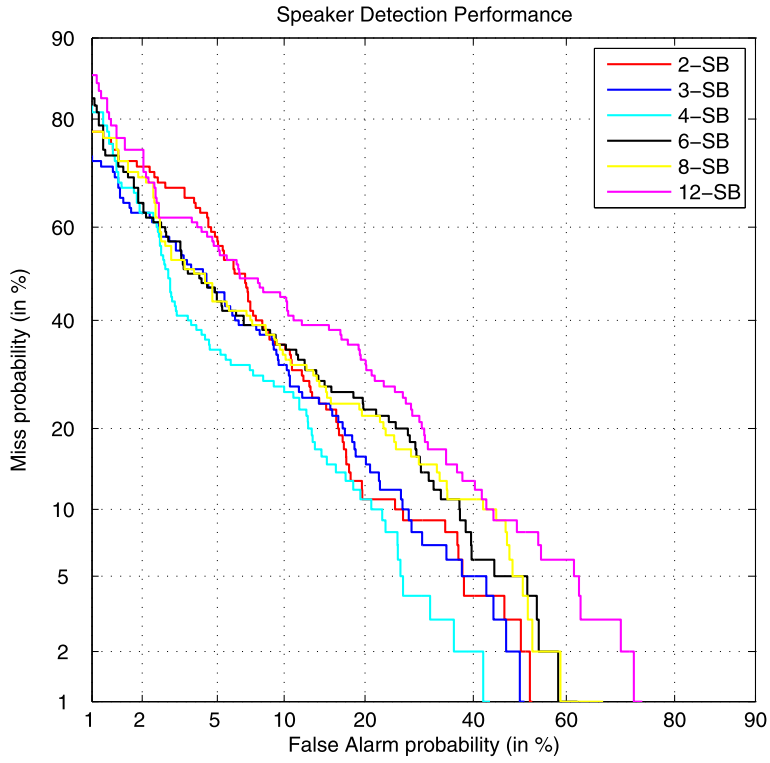

Figure 13 The recognition performance derived from different division ways.

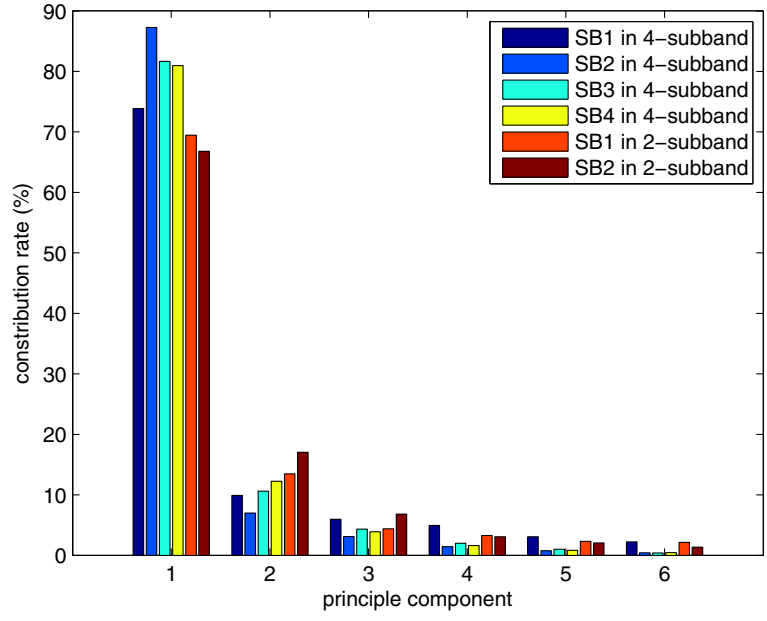

Figure 14 The comparison in contribution rate derived from 4 sub-bands and 2 sub-bands.

Competing interests
The authors declare that they have no competing interests.

Received: 26 April 2014 Accepted: 2 October 2014
Published online: 21 October 2014

References

1. J Pelecanos, S Sridharan, Feature warping for robust speaker verification. ISCA Workshop Speaker Recognition, June 213–218 (2001)
2. DA Reynolds, Channel robust speaker verification via feature mapping. ICASSP. **2**, 53–56 (2003)
3. V Chandran, D Ning, S Sridharan, in *Biometric Authentication*. Speaker identification using higher order spectral phase features and their effectiveness vis-avis mel-cepstral features, vol. 3072 (Springer Verlag Berlin, 2004), pp. 1–20
4. DA Reynolds, TF Quatieri, RB Dunn, Speaker verification using adapted Gaussian mixture models. Digital Signal Process. **10**(1–3), 19–41 (2000)
5. R Auckenthaler, M Carey, H Lloyd-Thomas, Score normalization for text-independent speaker verification systems. Digital Signal Process. **10**(1–3), 42–54 (2000)
6. P Kenny, P Dumouchel, in *Proc. ODYSSEY 2004-The Speaker and Language Recognition Workshop*. Experiments in speaker verification using factor analysis likelihood ratios (Toledo, Spain, May 31–June 3 2004), pp. 219–226
7. P Kenny, G Boulianne, P Ouellet, P Dumouchel, Factor analysis simplified. ICASSP. **1**, 637–640 (2005)
8. P Jančovič, M Köküer, Estimation of voicing-character of speech spectra based on spectral shape. IEEE Signal Process. Lett. **14**(1), 66–69 (2007)
9. M McLaren, D van Leeuwen, in *ICASSP*. Improved speaker recognition when using i-vectors from multiple speech sources (Prague, 2011), pp. 5460–5463
10. JA González, AM Peinado, N Ma, AM Gómez, J Barker, MMSE-based missing-feature reconstruction with temporal modeling for robust speech recognition. IEEE Trans. Audio, Speech, Lang. Process. **21**(3), 624–635 (2013)
11. T May, S van de Par, A Kohlrausch, Noise-robust speaker recognition combining missing data techniques and universal background modeling. IEEE Trans. Audio, Speech, Lang. Process. **20**(1), 108–121 (2012)
12. R Togneri, D Pullella, An overview of speaker identification: accuracy and robustness issues. IEEE Circuits Syst. Mag. **Second Quarter**, 23–61 (2011)
13. M Cooke, P Green, L Josifovski, A Vizinho, Robust automatic speech recognition with missing and unreliable acoustic data. Speech Commun. **34**, 267–285 (2001)
14. X Zhao, Y Shao, D Wang, CASA-Based Robust Speaker Identification. IEEE Trans. Audio, Speech Lang. Process. **20**, 1608–1616 (2012)
15. N Ma, J Barker, H Christensen, P Green, Combining speech fragment decoding and adaptive noise floor modelling. IEEE Trans. Audio Speech Lang. Process. **20**(3), 818–827 (2012)
16. JF Gemmeke, VanH Hamme, B Cranen, L Boves, Compressive sensing for missing data imputation in noise robust speech recognition. IEEE J. Sel. Topics Signal Process. **4**(2), 272–287 (2010)
17. JA González, AM Peinado, AM Gómez, N Ma, J Barker, in *IEEE Trans. Audio Speech Lang. Process*. Combining missing-data reconstruction and uncertainty decoding for robust speech recognition (Kyoto, 2012), pp. 4693–4696
18. B Raj, ML Seltzer, RM Stern, Reconstruction of missing features for robust speech recognition. Speech Commun. **43**, 195–202 (1997)
19. B Raj, *Reconstruction of incomplete spectrograms for robust speech recognition*. PhD dissertation, Pittsburgh, PA, Carnegie Mellon Univ, 2000
20. L Raj, JF Bonastre, in *Proc. Audio and Video based Biometric Person Authentication*. LNCS, ed. by J Bigün, G Chollet, and G Borgefors. Subband approach for automatic speaker recognition: optimal division of the frequency domain (Springer Heidelberg, 1997), pp. 195–202
21. L Besacier, JF Bonastre, C Fredouille, Localization and selection of speaker-specific information with statistical modeling. Speech Comm. **31**, 89–106 (2000)
22. H Bourlard, S Dupont, A new ASR approach based on independent processing and recombination of partial frequency bands. ICSLP. **1**, 426–429 (1996)
23. J Shlens, A tutorial on principal component analysis. Systems Neurobiology Laboratory, Salk Institute for Biological Studies, version 2, 1–13 (2005)
24. D Reynolds, R Rose, Robust text-independent speaker identification using Gaussian mixture speaker models. IEEE Trans. Speech Audio Process. **3**(1), 72–83 (1995)
25. W Campbell, D Sturim, D Reynolds, Support vector machines using GMM supervectors for speaker verification. IEEE Signal Process. Lett. **13**(5), 308–311 (2006)
26. DA Reynolds, TF Quatieri, RB Dunn, Speaker verification using adapted Gaussian mixture models. Digital Signal Process. **10**, 19–41 (2000)
27. B Raj, RM Stern, Missing-feature approaches in speech recognition. IEEE Signal Process. Mag. **22**(5), 101–116 (2005)
28. GJ Brown, D Wang, *Separation of Speech by Computational Auditory Scene Analysis*. (Springer Verlag, New York, 2005), pp. 371–402
29. ML Seltzer, B Raj, RM Stern, A Bayesian classifier for spectrographic mask estimation for missing feature speech recognition. Speech Commun. **43**(4), 379–393 (2004)
30. X Zhao, Y Wang, D Wang, Robust speaker identification in noisy and reverberant conditions. IEEE Trans. Audio, Speech Lang. Process. **22**, 836–845 (2014, in press)
31. Martin Rainer, Noise power spectral density estimation based on optimal smoothing and minimum statistics. IEEE Trans. Speech Audio Process. **9**, 504–512 (2001)
32. M Brookes, VOICEBOX: Speech Processing Toolbox for MATLAB (2009). [Online] Available: http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html
33. A Papoulis, *Probability, Random Variables, and Stochastic Processes*. (Academic Press, New York, 1991). Third Edition, McGraw Hill, Inc
34. A Varga, H Steeneken, M Tomlinson, D Jones, in *Tech. Rep., Speech Res. Unit, Defense Res. Agency*. The NOISEX-92 study on the effect of additive noise on automatic speech recognition (Malvern, U.K., 1992). (Available from NOISEX-92 CD-ROMS)
35. A Martin, G Doddington, T Kamm, M Ordowski, in *Proceedings of the European Conference on Speech communication and Technology*. The DET curve in assessment of detection task performance, (1997), pp. 1895–1898

doi:10.1186/s13636-014-0040-7
Cite this article as: Yan *et al.*: A sub-band-based feature reconstruction approach for robust speaker recognition. *EURASIP Journal on Audio, Speech, and Music Processing* 2014 **2014**:40.