Research Article

Multitaper MFCC and normalized multitaper phase-based features for speaker verification

Arash Mansouri[1] · **Eduardo Castillo-Guerra**[1]

Abstract
One trend of current research activity in speaker verification systems focuses on features that capture complementary information to the Mel-frequency cepstral coefficients (MFCCs). In conventional speaker verification methods, MFCCs are usually computed from a single-tapered DFT spectrum. These methods use the power of the Fourier transform of the time-domain speech frames and ignore the phase component due to its large amount of uncertainty. The multitaper phase information has been also ignored when the multitaper MFCC method applies. We propose a phase information extraction method that normalizes the change variation in the multitaper phase. The goal of this work is to incorporate phase information using low-variance multitaper spectrum estimation method instead of conventional single-taper window and normalizes the change variation in multitaper phase according to the frame position of the input speech to reduce the uncertainty of multitaper phase information in both the state-of-the-art Gaussian mixture model-universal background model (GMM-UBM) baseline and the i-vector speaker verification system. The experiments are conducted on the TIMIT database for both the clean and noisy conditions (SNR = 10 dB). The multitaper inverted-MFCC is also added to the overall system to produce diversity with respect to the fused subsystem. The results show relative improvement of 34% and 25% in terms of equal error rate (EER) over the MFCC baseline in clean and noisy conditions for GMM-UBM system, respectively. The relative improvement in terms of EER with i-vector back-end are about 36% and 16% in clean and noisy conditions as well.

Keywords Gaussian Mixture Model (GMM) · Multitaper Phase Information · Probabilistic Linear Discriminant Analysis (PLDA) · I-Vector · Speaker Verification (SV) System

1 Introduction

Speaker verification (SV) is the task of authenticating a claimed identity based on a speech sample. Speaker verification systems fall into two categories: text-dependent and text-independent. Text-dependent SV system requires the same speech for both enrolment and verification phase. In this paper, however, we focus on text-independent speaker verification where the speaker must be recognized from any utterance. The feature extraction phase in conventional speaker verification system is based on Mel-frequency cepstral coefficients (MFCCs) which uses the power spectrum of the time-domain speech frames. However, the MFCCs are highly associated with phonetic information of the speech signals and are not fundamentally optimized to capture speaker-specific information [1]. Therefore, not all of MFCCs are equally relevant for speaker recognition and some of them are redundant or dependent on other coefficients [2]. Complementary features have been combined with MFCCs to address this shortcoming. The fusion of different features shows a trend in achieving improved performance specially under limited data conditions [3]. Fusion of MFCCs and these complementary features can be achieved in either score level or feature

✉ Arash Mansouri, a.mansouri@unb.ca; Eduardo Castillo-Guerra, ecastill@unb.ca | [1]Department of Electrical and Computer Engineering, University of New Brunswick, Fredericton, NB, Canada.

SN Applied Sciences (2019) 1:290 | https://doi.org/10.1007/s42452-019-0305-y

Received: 11 January 2019 / Accepted: 26 February 2019 / Published online: 2 March 2019

level [4–6]. In feature level, different types of feature vectors are concatenated to construct a new feature vector, whereas in score level fusion, as in this work, a linear function of different matching likelihoods is used to obtain the overall score.

One of these complement features is phase component of speech signal that is usually ignored in conventional MFCC-based SV systems. The importance of phase information is investigated by [7–11]. Paliwal and Alsteris [7] explored the usefulness of phase information in human speech and reported that phase spectrum can contribute to speech intelligibility as much as the magnitude spectrum. Murty and Yegnanarayana [8] investigated the role of the residual phase as a MFCC complementary feature. They extracted the residual phase information from speech signals by linear prediction analysis and reported improvement in overall speaker recognition performance. There are also other studies which concentrated on modeling the phase and group delay phase information [9–11], to name a few.

However, the effect of large uncertainty in phase information has been controversial in the literature. Paliwal et al. [7] concluded that the phase information for shorter windows could be informative if the shape of the window function is properly selected. Nakagawa et al. [12] normalized the phase information according to the frame position of the input speech signal in order to compensate the phase variations. Sahidullah and Saha [13] reported that the type of window function, its duration and overlap between windowed frames are important parameters that affect phase information uncertainty in the preprocessing step of SV system. This is due to the fact that by changing these parameters in phase-based systems not only the sound quality but also the phonetic value of a stop consonant could be changed [14]. Moreover, the windowing method in the frequency domain is a convolution of the speech spectrum and frequency response of the window function that has a large main lobe and several side lobes. This general shape of window functions is the source of two problems; (1) the frequency resolution problem, which is caused by the main lobe of the window function. The wider is the main lobe, the larger frequency interval of the speech spectrum gets smoothed. (2) the spectral leakage problem, which is caused by the small side lobes of the window function. The amount of spectral leakage increases with the magnitude of the side lobes. As a result, selection of window function for Discrete Fourier Transform (DFT) depends on the underlying application. It is necessary in magnitude spectrum analysis and cepstral feature extraction to choose a window function that proposes a better trade-off between frequency resolution and spectral leakage. One of the popular window function used in state-of-the-art SV systems is the Hamming window which has reasonable side lobe characteristics with a length of 10 ms $\leq T_{\omega} \leq$ 40 ms. Although the side lobes do not cause a major problem in phase information extraction, the smoothing effect caused by the main lobe is more serious problem in phase spectrum estimation using Hamming window [12].

Selecting the optimal window function and improving the front-end process is still an open challenge problem in speech and speaker recognition applications [15–19]. Recently, there has been work that focused on the design of new window function [13, 19–22]. For instance, Mottaghi-Kashtiban and Shayesteh [19] proposes a method to find the optimal amplitudes of DC term and cosine function by adding the third harmonic of the cosine function to the Hamming window. Nonetheless, such a single-taper window function, that is Hamming and the likes, still produces high variance for the direct spectral estimation. This issue can be recovered by averaging spectral estimates using a set of different tapers, leading to a so-called multitaper window functions. The multitaper technique reduces the variance of the spectral estimation by using multiple time domain window functions (tapers) rather than a single window function.

The use of multitaper MFCC features for speaker verification tasks was motivated in [23]. The idea of using multitaper in speaker verification addresses the problem of windowed periodogram that suffers from high variance as well as large bias [24]. The periodogram is a biased estimate due to spectral leakage, which is a tendency for power from strong peaks to spread into neighboring frequency intervals of lower power. While Hamming window reduces the large bias of the Fourier transformation squared magnitude, it is not able to reduce the high variance of the periodogram.

In this work, we combine phase spectral information extracted by multitaper window functions with magnitude spectral information at the score level in order to achieve more efficient verification system. We also propose a phase information normalization technique to address the phase changes according to the frame position in multitapering method. The uncertainty in features is also modeled by the variance of weighted mixture of GMM models; smaller feature variance results in less random variance of the GMM. We expect to reduce phase uncertainty by using normalization method on top of multitapering phase information extraction. The conclusion reached is that combining normalized multitaper phase information with MFCCs is a step forward to the state of the art SV systems such as [25] where single-taper phase information were employed. Our experiments on TIMIT dataset show 34% relative improvement in terms of EER when multitaper phase information is used compared to single-taper MFCC-based baseline. Moreover, our

experiments show that the contribution is also robust to noise with SNR = 10 dB, yielding to 25% EER improvement.

Furthermore, the importance of intra-speaker and inter-speaker variability have made state-of-the-art SV systems gradually migrate from GMM-UBM to i-vector with probabilistic discriminant analysis (PLDA) scoring [26]. Therefore, for the best performing system based on GMM-UBM we also provide the verification results with the state-of-the-art PLDA i-vector back-end.

The remainder of this paper is as follows: the next section explains the multitapering method. Section 3 investigates the effect of phase on speaker verification, and then formulate the normalized multitaper phase information. In Sect. 4, the results of our proposed method are compared with those of related works. Finally, Sect. 5 concludes the paper and introduces future directions for research.

2 Multitapering method

The multitapering method is obtained by multiplying the data sequence by a set of orthogonal windows to form several single taper periodogram that are averaged to estimate of the power spectral density. Let us assume that $r_c(n)$ is a discrete-time symmetrical cepstrum, where $r_c(-n) = r_c(n)$, and $S_x(f)$ is the real-valued spectral density function for a real-valued stationary process. $x = [x(0), x(1), \ldots, x(N-1)]^T$ denotes one frame of speech of N samples. The $r_c(n)$ can be obtained as:

$$r_c(n) = \int_{-\frac{1}{2}}^{+\frac{1}{2}} logS_x(f)e^{i2\pi fn}df; n \in [-N+1, N-1] \quad (1)$$

The mean square error (MSE) of the cepstrum is defined as:

$$MSE = \int_{-\frac{1}{2}}^{+\frac{1}{2}} E[(\hat{S}_c(f) - S_c(f))^2]df \quad (2)$$

where $\hat{S}_c(f)$ is the spectral estimation of $r_c(n)$. The goal of the multitapering method is to find the optimal estimator in terms of MSE based on multiple window functions. It is assumed that the statistical information discarded by one window is partially recovered by the other tapers. The multitaper spectrum estimator $\hat{S}_c(f)$ is obtained as a weighted average of subspectra as,

$$\hat{S}_c(f) = \sum_{j=1}^{k} \lambda(j) \left| \sum_{t=0}^{N-1} w_j(t)x(t)e^{-\frac{i2\pi tf}{N}} \right|^2 \quad (3)$$

where k is the number of tapers $w_j = [w(0), x(1), \ldots, w(N-1)]^T$ with corresponding weights $\lambda(j)$ [21]. The weighting function is defined to generate a smooth estimate with less variance than single taper methods. This spectrum estimation for the baseline system and for a single window is obtained by

$$\hat{S}_c(f) = \left| \sum_{t=0}^{N-1} w(t)x(t)e^{-\frac{i2\pi tf}{N}} \right|^2 \quad (4)$$

Figure 1 shows multitaper window shapes for Thomson, Multipeak, and Sine-Weighted Cepstrum Estimator (SWCE) window function where $k = 4$. A Hamming window can be considered as a single-taper in Eq. 3 where $k = 1$ and $\lambda = 1$. For a multitapering method, the tapers are typically chosen to be orthonormal. The choice of taper has a significant effect on the accuracy of the spectrum estimation. The details of finding the optimal number of tapers for a given process and different applications can be found in [27–30].

Once the tapers are selected as the window functions, we extract the phase information using multitapering method as explained in the next section.

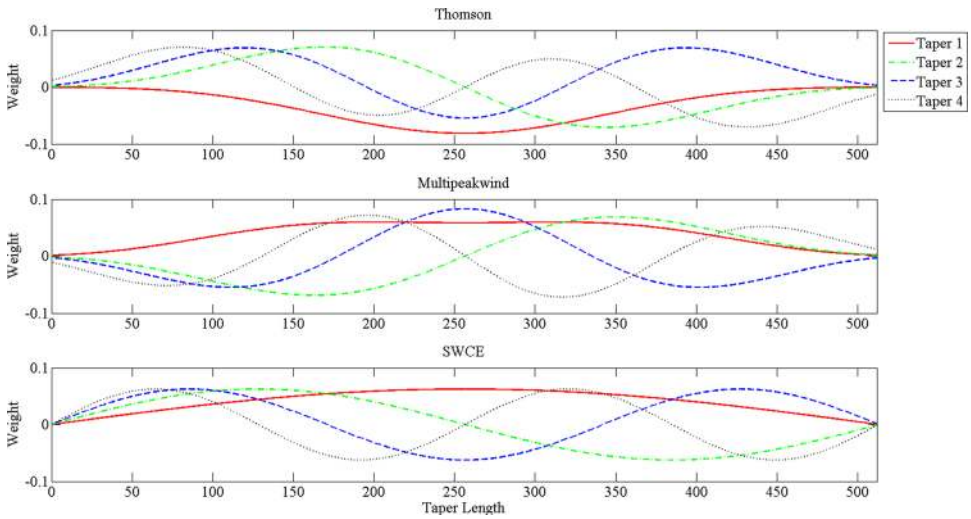

Fig. 1 Different multitaper window functions with k = 4

3 Multitaper phase information extraction

The dependency of the phase-based features on the starting sample window and difficulties of phase unwrapping are two challenges in combining phase information with MFCCs. Different phase unwrapping methods have been studied and group delay-based phase information has been proposed to address these challenges and to make the phase information less sensitive to the phase warping issue [31–35]. The fast changes of phase spectrum according to the position of a window could be more critical for applying multiple parallel window functions in multitapering methods. While the position of a window shifts, the phase $\theta(\omega, t)$ changes during the windowing process even for a same frequency. The difference of phase information between shifted windows is considerable and depends on the clipping position of the input speech. To overcome this problem, Nakagawa et al. [12] normalized the phase response with respect to the frame position for Hamming window. In this work, we adapt the method proposed by [12] to normalize the position of shifted windows in the phase information extraction step for multitapering method as follows.

The spectrum of a speech signal is obtained from input speech as,

$$S(\omega, t) = X(\omega, t) + jY(\omega, t) = \sqrt{X^2(\omega, t) + Y^2(\omega, t)} \times e^{j\theta(\omega, t)} \tag{5}$$

where $S(\omega, t)$ is the DFT of input speech, $X(\omega, t)$ and $Y(\omega, t)$ are the real and imaginary parts of $S(\omega, t)$, $X^2(\omega, t) + Y^2(\omega, t)$ is the power spectrum of $S(\omega, t)$ which is used to calculate MFCCs. $\theta(\omega, t)$ and $2\pi + \theta(\omega, t)$ are the phase information of speech signals extracted from different frames for the same voice data. The $\theta(\omega, t)$ and $2\pi + \theta(\omega, t)$ are mapped to the same value by constraining the phase values to $[-\pi, \pi]$ before sending the phase information to the classifier.

To normalize the changes of $\theta(\omega, t)$ from shifted frames, the phase of a certain basis frequency ω_b is kept constant. Then the phase values of other frequencies can be normalized according to the frequency of basis ω_b. Figure 2 shows the effect of shifted windows on phase information for a short subband before normalization. Although the windows shifted for just 10 ms, the phase information from these two windows is not the same. If we pass this unnormalized wrapped phase information directly to the classifier to create speaker models, the models could not accurately represent the corresponding speakers. By setting the phases of the basis frequency ω_b to a constant value, for example $\theta_{\omega_b} = \frac{\pi}{4}$, the spectrum of specific frequency is obtained as,

$$\begin{aligned} S'(\omega', t) &= \tilde{X}(\omega', t) + j\tilde{Y}(\omega', t) \\ &= \sqrt{X^2(\omega', t) + Y^2(\omega', t)} \times e^{j\theta(\omega', t)} \times e^{j\frac{\omega'}{\omega_b}\left(\frac{\pi}{4} - \theta(\omega', t)\right)} \end{aligned} \tag{6}$$

Without loss of generality, we can set the basis frequency to $2\pi \times 1000$, and calculate the real and imaginary parts of the spectrum at the desired frequency $\omega' = 2\pi f'$ as,

$$\begin{aligned} \tilde{X}(\omega', t) = &\sqrt{X^2(\omega', t) + Y^2(\omega', t)} \\ &\times cos\left[\theta(\omega', t) + \frac{\omega'}{2\pi \times 1000}\left(\frac{\pi}{4} - \theta(\omega', t)\right)\right] \end{aligned} \tag{7}$$

$$\begin{aligned} \tilde{Y}(\omega', t) = &\sqrt{X^2(\omega', t) + Y^2(\omega', t)} \\ &\times sin\left[\theta(\omega', t) + \frac{\omega'}{2\pi \times 1000}\left(\frac{\pi}{4} - \theta(\omega', t)\right)\right] \end{aligned} \tag{8}$$

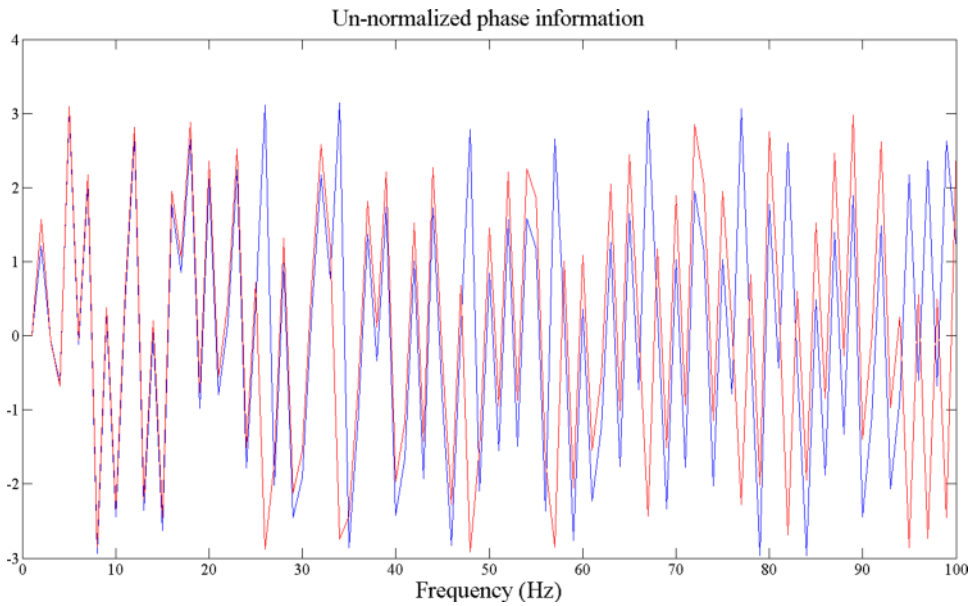

Fig. 2 Unnormalized wrapped phases of two different windows

Therefore, the final normalized phase is obtained as,

$$\tilde{\theta}(\omega', t) = \theta(\omega', t) + \frac{\omega'}{2\pi \times 1000}\left(\frac{\pi}{4} - \theta(\omega', t)\right) \quad (9)$$

Although this normalization method significantly reduces the effect of frame position for two shifted windows, it is not still a perfect match as illustrated in Fig. 3. As the final step of phase normalization and to address the illustrated differences in Fig. 3, let us compare two values of different phases $\pi - \tilde{\theta}_1$ and $\tilde{\theta}_2 = -\pi - \tilde{\theta}_1$. The difference of these two phases is $2\pi - 2\tilde{\theta}_1$, and if we set $\tilde{\theta}_1 \approx 0$, then the difference would be 2π, while we know that the two phases are very similar to each other. This problem can be solved if we modify the phase into coordinates on a unit circle as,

$$\tilde{\theta} \rightarrow \{cos\tilde{\theta}, sin\tilde{\theta}\} \quad (10)$$

Then the $cos\tilde{\theta}$ and $sin\tilde{\theta}$ values are passed through the GMM classifier for two different phases with very similar values. Figure 4 shows the effect of shifted frames after applying the proposed normalizing method to phase information for two different windows.

We need to replace the estimation of phase spectrum with amplitude estimation in the multitapers weighting equation in order to extract the multitaper phase-based features. Let us assume k multitapers, where $j = 1, \ldots, k$, are used with corresponding weights $\lambda(j)$ for one frame of speech of N samples. The multitaper phase information is therefore obtained as,

$$\tilde{\theta} = \sum_{j=1}^{k} \lambda(j)\left\{ cos \sum_{t=0}^{N-1} \theta(\omega', t) + \frac{\omega'}{2\pi \times 1000}\left(\frac{\pi}{4} - \theta(\omega', t)\right), \right.$$
$$\left. sin \sum_{t=0}^{N-1} \theta(\omega', t) + \frac{\omega'}{2\pi \times 1000}\left(\frac{\pi}{4} - \theta(\omega', t)\right)\right\} \quad (11)$$

Figure 5 illustrates the proposed multitapering method designed and adopted for magnitude and phase information. As seen, the final amplitude spectrum and phase information are formed as a weighted average of the individual tapered window. The tapers are designed to provide approximately uncorrelated spectrum estimates so that the averaging them reduces the variance of the spectrum estimation and makes the spectrum less sensitive to noise, compared with the conventional single-taper method.

The steps of extracting multitaper MFCC and inverted-MFCC (IMFCC) in a SV system is explained in [36]. The MFCC and IMFCC feature extraction begins with pre-processing and followed by short-time Fourier transform (STFT) analysis using multitaper method. The final static features are obtained using discrete cosine transform (DCT). The feature vectors of MFCC, IMFCC, and normalized phase information are separately sent to classifiers and the likelihoods of each GMM are fused to produce the total score S_{Total} as,

$$S_{Total} = (1-\gamma)S_{Phase} + \gamma(\alpha S_{MFCC} + (1-\alpha)S_{IMFCC}) \quad (12)$$

where S_{MFCC}, S_{IMFCC}, and S_{Phase} are the likelihoods produced by MFCC-based, IMFCC-based, and phase-based speaker models, respectively. Parameters α and γ denote weighting coefficients to adjust the weight of each speaker model in final score calculation. For example, if $\alpha = 1$ and $\gamma = 1$, the system is the conventional MFCC-GMM SV [25]. The effect of phase information to the overall system performance is evaluated by choosing $\gamma \in \mathcal{R}^{[0,1)}$. The phase-only GMM model is when $\gamma = 0$.

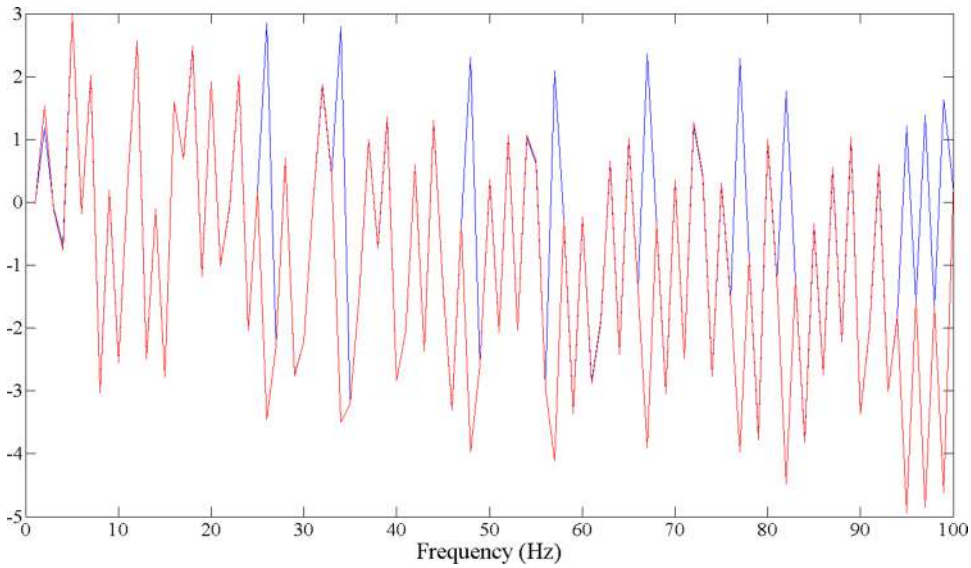

Fig. 3 Normalized wrapped phases of two different windows

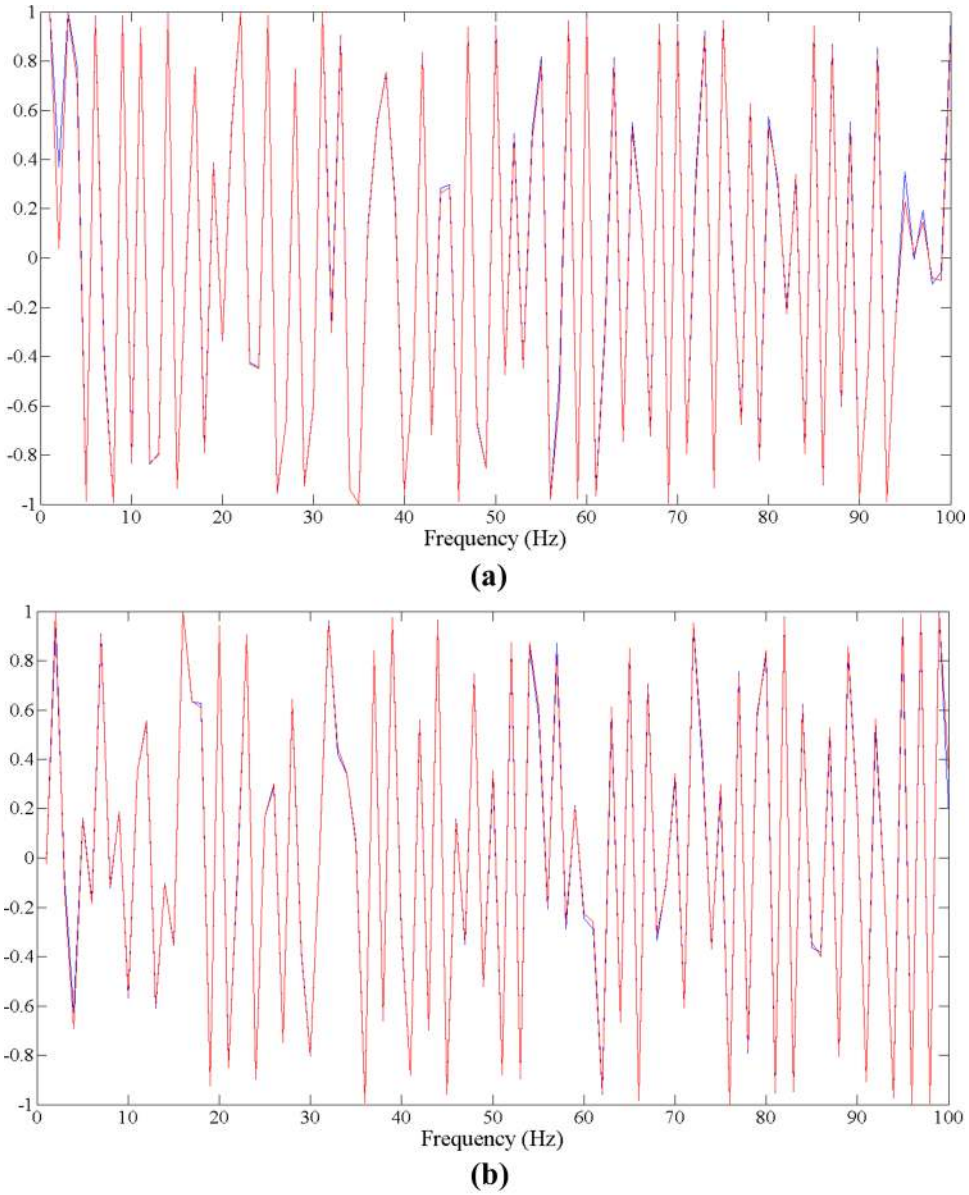

Fig. 4 Example of the effect of frame position on phase. **a** *Cosθ* parameters of normalized wrapped phases of two different windows. **b** *Sinθ* parameters of normalized wrapped phases of two different windows

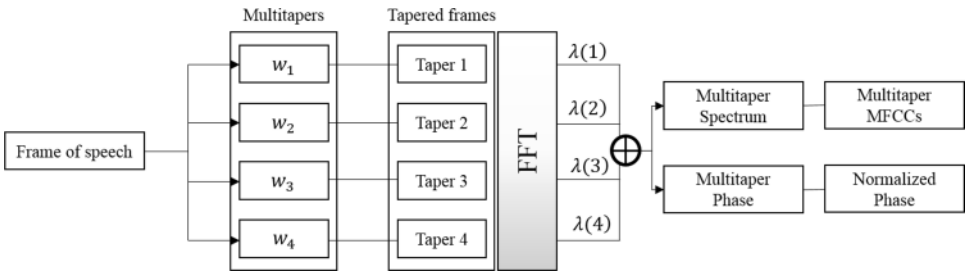

Fig. 5 The block diagram of proposed multitapering MFCC and phase information system

4 Experiments

We use TIMIT corpus [37] with 16 KHz sampling rate as the dataset in enrollment and verification phases. It contains 630 speakers (192 female and 438 male). It provides 10 utterances for each speaker from which we selected 7 utterances for training and 3 utterances for testing. We used 189 speakers (58 female, 131 male) for training and testing phases (gender-independent)

and all 10 utterances of other 441 speakers to generate universal background model. We train two gender-independent, full covariance UBMs; one with 256 component GMMs and the other with 8 number of mixture. All the experiments use a mixture size of 256 except the one that investigate the length of the analysis window for the phased-based system since it is a side experiment to investigate the trend of using short versus long duration window for phase information extraction. Verification trials consist of all possible model-test combinations resulting in a total of 107,163 trials ($3 * 189 = 567$ targets versus $3 * 188 * 189 = 166,596$ impostor trials). Note that all the speaker models are derived from the UBM via adaptation in this paper.

Two different validation tests are employed: (1) match and clean training where the train and test data are same and without additive noise, (2) mismatch and noisy training in which white noise is added to test data at SNR = 10 dB and a classifier that is trained when clean signals is applied. Different experiments are carried out in order to investigate the effects of several factors in the performance of proposed SV system including:

1. number of tapers using different multitapering methods,
2. short and long frame length in phase information extraction,
3. normalizing multitaper phase information and its fusion with multitaper MFCC,
4. fusion of all multitaper MFCC, multitaper IMFCC and normalized multitaper phase information.

EER is calculated to evaluate the accuracy of the proposed speaker verification system against the baselines.

4.1 Baseline design

MFCC for the baseline system are computed using Hamming window with a frame duration of 30 ms and 75% overlaps between frames, and 40-channel Mel-frequency filterbank. The lowest 16 MFCC are retained, excluding delta, double-delta, and energy coefficients. In this paper, we call this baseline the Hamming-based baseline. These configurations are the same for multitapering methods, Thomson [27], Sine-Weighted Cepstrum Estimator (SWCE) [30], and Multipeak [29], except that the spectrum is estimated from Eq. 3 instead of Eq. 4. The size of frames and the overlap factor for multitapers are the same as for the Hamming window.

4.2 Choice of tapers

First, the choice of the spectrum estimator is studied to investigate the effect of using different types of taper and the number of windowing functions (tapers) on the overall system accuracy. This experiment does not include the phase information. The number of tapers for each of the multitapering methods is varied from 2 to 12 tapers, and the result of each method is compared to the Hamming-based baseline. The accuracy of the speaker verification systems is studied under both the clean-match and noisy-mismatch conditions.

Figure 6 compares the results of EERs for multitaper-based system with different multitapering methods. The maximum relative improvement (RI) obtained by each multitaper is compared to the Hamming-based baseline in the last column. As shown, the maximum relative improvement is 18.9% for the Thomson method with 4 tapers. The three multitaper methods in most of the settings outperform the Hamming-based baseline in terms of EER. We attribute this improvement to the fact that employing the MFCC filterbank on top of the multitapering method brings additional averaging to the system which result in low EER and more accurate verification. Figure 7 illustrates the relative improvements obtained by each system. Positive RI indicates a reduction in EER whereas negative RI indicates an increase in EER. The average improvements obtained by using Thomson, Multipeak, and SWCE multitaper methods are 11.2%, 4.2%, and 8.9%, respectively. Although the optimum number of tapers depends on the method and objective of the system, it seems that 2–6 tapers is the best choice in the clean-match condition according to the settings of this experiment.

Next, we study the accuracy of multitapering methods under additive noise. The number of tapers for each method is varied from 2 to 12 tapers while the amount of SNR is set to 10 dB. Experiments are carried out with different multitapering methods keeping other parameters identical as the previous experiment. The pre-processing, feature extraction and classification are the same for all three multitaper-based baselines. Figure 8 illustrates the results of using multitapering methods in noisy condition. Although the accuracy of all multitaper-based baselines significantly drops under low SNR in compare with their identical setting in clean condition, they have better average performance compared to the Hamming-based baseline under additive noise. As shown in Fig. 9, the Thomson method performs the best when the number of tapers is between 2 and 6. Specifically, the average EER of Thomson, Multipeak, and SWCE systems are 28.11%, 30.35%, and 27.72%, while the maximum EER obtained by each of them are 25.57%, 26.73%, and 25.77%, respectively. Although the average EER of the Multipeak method is

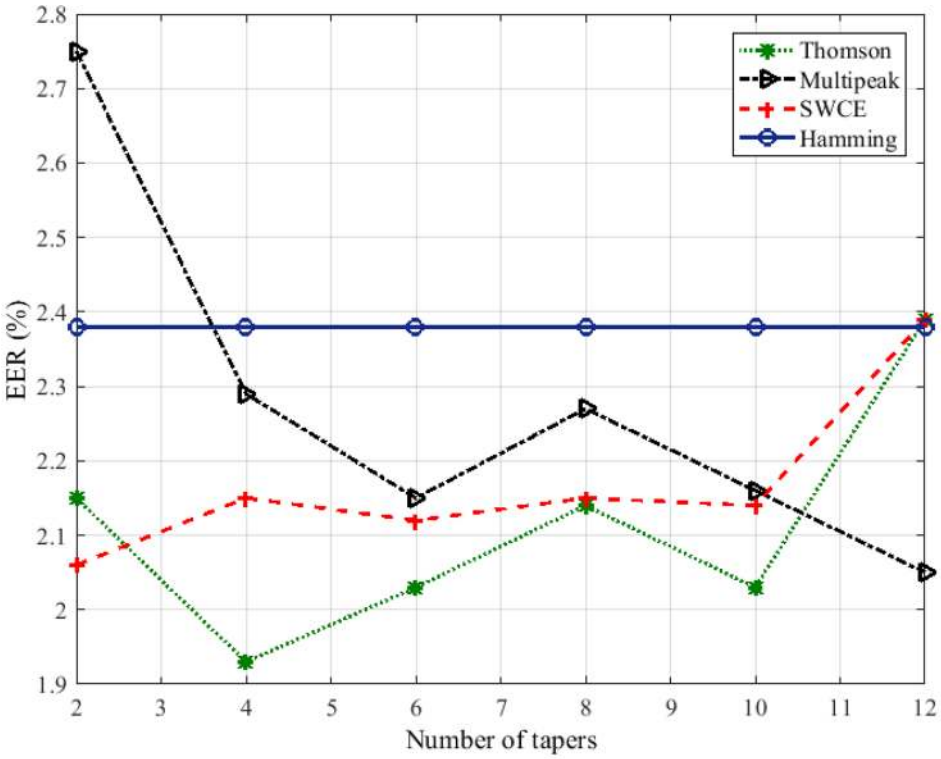

Fig. 6 Performance evaluation of multitaper MFCC in terms of EER (%) for different number of tapers in clean-match condition

Tapers No.	k=2	k=4	k=6	k=8	k=10	k=12	Max RI
Thomson	2.15	**1.93**	2.03	2.14	2.03	2.39	18.9%
Multipeak	2.75	2.29	2.15	2.27	2.16	**2.05**	13.8%
SWCE	**2.06**	2.15	2.12	2.15	2.14	2.39	13.4%
Hamming	2.38	2.38	2.38	2.38	2.38	2.38	-

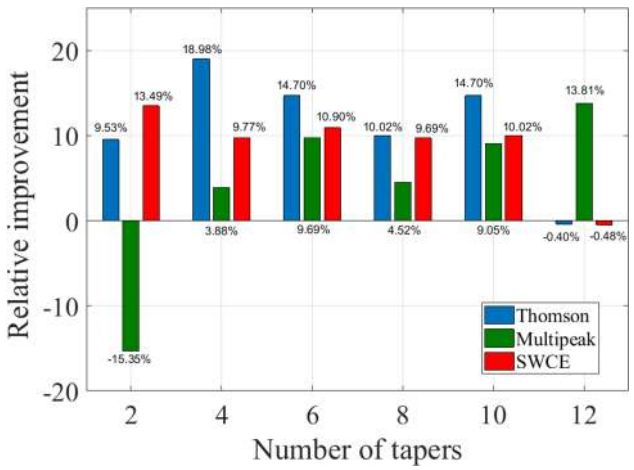

Fig. 7 Relative improvement (RI) of the multitaper systems in comparison with Hamming-baseline in clean-match condition

more than the Hamming-based baseline under additive noise, all three multitaper variants outperform the Hamming-based baseline for different number of tapers. In the presence of high noise with SNR = 10 dB, Thomson performs best on average. Figure 9 shows the details of relative improvements obtained by each multitaper-based system compared to the Hamming-based baseline. Based on the results of the above experiments, the intuition of improving Hamming-based baseline using multitapering methods seems feasible both in clean and noisy conditions. However, we could not reach a unique best setting for number of tapers and its type. As a result, our next experiments include varied number of tapers for all types of multitapering methods.

Figures 10 and 11 provide the SWCE multitaper spectrum for a 30 ms duration speech signal, as an example. As seen, the tapers are designed so that the estimation errors in the subspectra are approximately uncorrelated. The outcome of 4 tapers in SWCE method produces smoother spectrum compared to the Hamming-based baseline due to its variance reduction.

4.3 Short versus long duration window for phase-based systems

The length of analysis window in speech and speaker recognition applications is important in both the amplitude

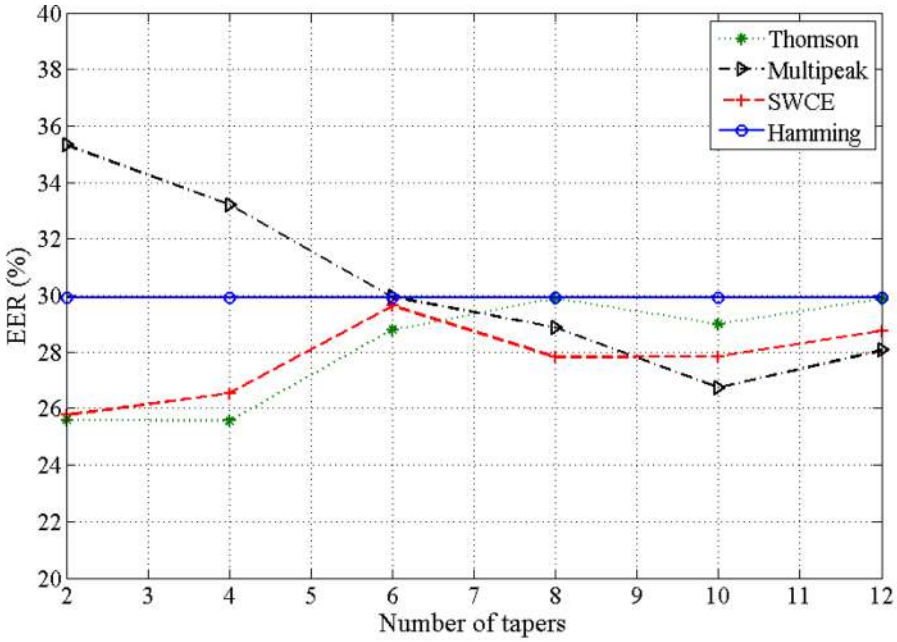

Fig. 8 Comparison of the baseline and multitaper systems under additive noise and mismatched condition (SNR = 10 dB)

Tapers No.	k=2	k=4	k=6	k=8	k=10	k=12	Max RI
Thomson	25.59	**25.57**	28.78	29.89	28.98	29.90	14.5%
Multipeak	35.33	33.20	29.95	28.86	26.73	**28.05**	10.7%
SWCE	**25.77**	26.53	29.63	27.82	27.85	28.75	13.9%
Hamming	29.93	29.93	29.93	29.93	29.93	29.93	-

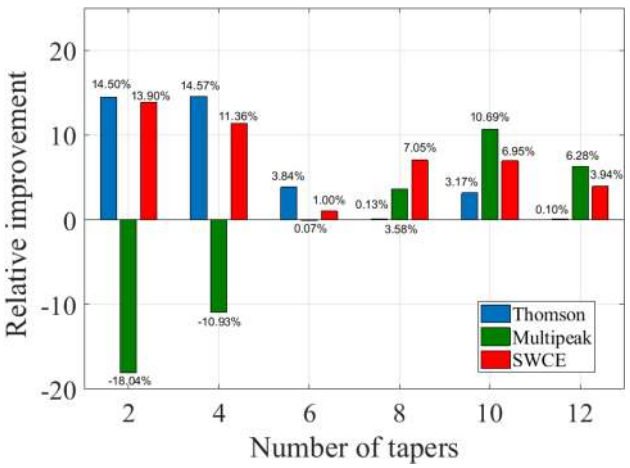

Fig. 9 Relative improvement (RI) of the multitaper systems in compared with Hamming-baseline under additive noise and SNR = 10 dB

and phase information extraction. The state-of-the-art SV systems use three major ranges for duration of analysis window which are sub-segmental (3–5 ms), segmental (10–40 ms) and suprasegmental (100–300 ms) [38]. It is traditionally believed in speech recognition applications that the magnitude spectrum contributes more when small window duration is employed while phase-based features are more informative for large window durations [14, 39]. For instance, use of large window with phase information, rather than small window, is more effective in order to recognize phoneme [40]. However, to the best of our knowledge, there is no study in speaker verification to investigate the effect of frame duration for phase information extraction. This section studies the effect of frame size, segmental (20 ms) versus suprasegmental (120 ms), on the performance of a MFCC + Phase-based system according to Eq. 12. Setting the weight $\gamma = 1$ shows an MFCC-only system without any phase information, while $\gamma = 0$ indicates phase-only system without MFCCs. The weight of IMFCC is set to zero during the experiment by setting $\alpha = 1$. Therefore, no high frequency information is involved in the likelihood score calculation.

The settings of MFCC system are identical to the Hamming-based baseline while the length of window for phase information extraction is either 20 ms (small duration window) or 120 ms (long duration window) with 90% overlap between frames. A DFT of 512 points is used in this experiment, which results in 256 individual frequency components. Since the TIMIT corpus has a 16 kHz sampling frequency, the interval between two adjacent frequency components is $\frac{8000}{256}$ Hz. The phase information is obtained from the lowest 12 components of the subband spectrum, from 30 Hz to 350 Hz. This experiment also investigates the effect of phase information normalization according to Eq. 11.

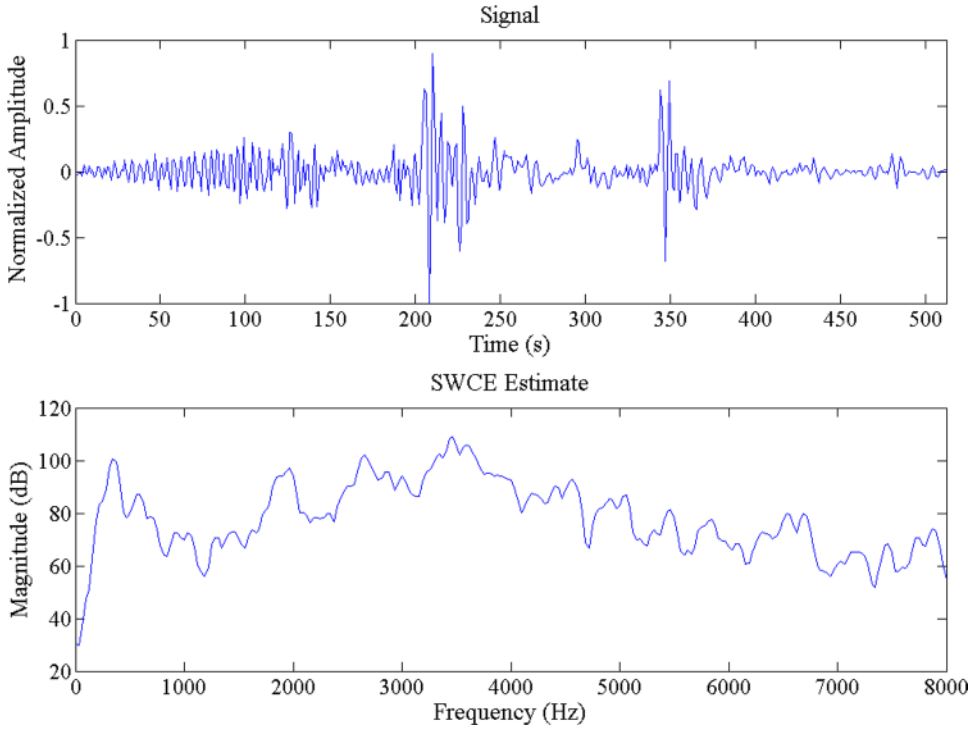

Fig. 10 30 ms framed speech signal (512 samples) and its SWCE final estimation

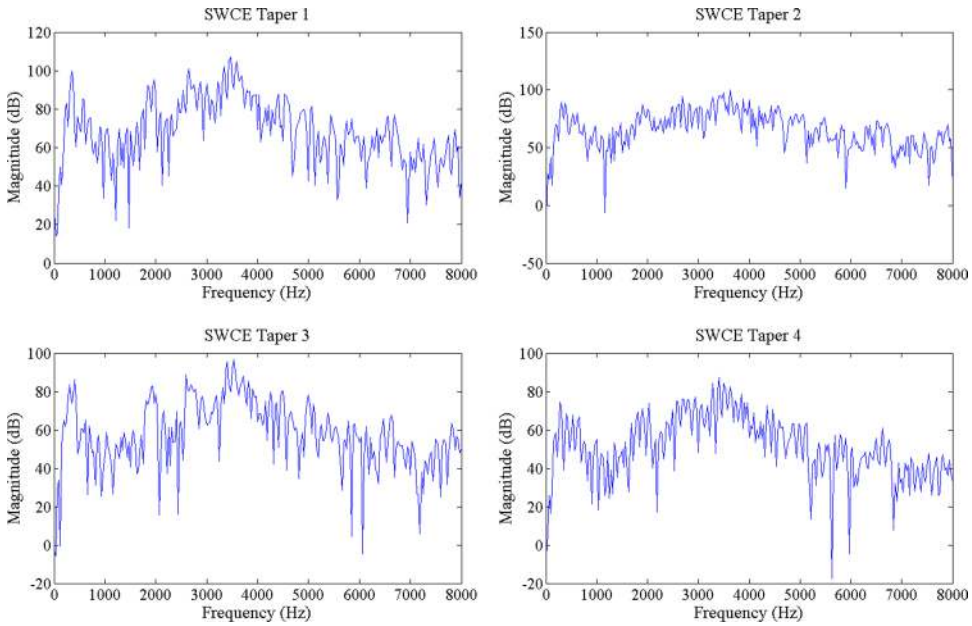

Fig. 11 SWCE spectral estimates with four tapers and uniform weights for a 30 ms duration framed speech signal

First, we compare the performance of short duration of 20 ms and large duration of 120 ms window in the clean-match condition for MFCC-only, phase-only, and fused systems. The comparison is also done based on the fact that whether our proposed normalization technique for phase-based features (Sect. 3) is in effect or not. As shown in Fig. 12, when there is no normalization, combining phase information through short window duration (20 ms-No Norm) is able to improve the MFCC-only system. Specifically, when $\gamma = 0.7$, the best RI is 1.13% compared to MFCC-only system. However, adding phase information for long window duration never outperforms the MFCC-only system. The fused system using long window duration reach to EER $= 2.64\%$ at best while the EER of MFCC-only system is still better and stays at 2.38%. Indeed, adding unnormalized long window duration phase information to the baseline system has a destructive effect in overall performance.

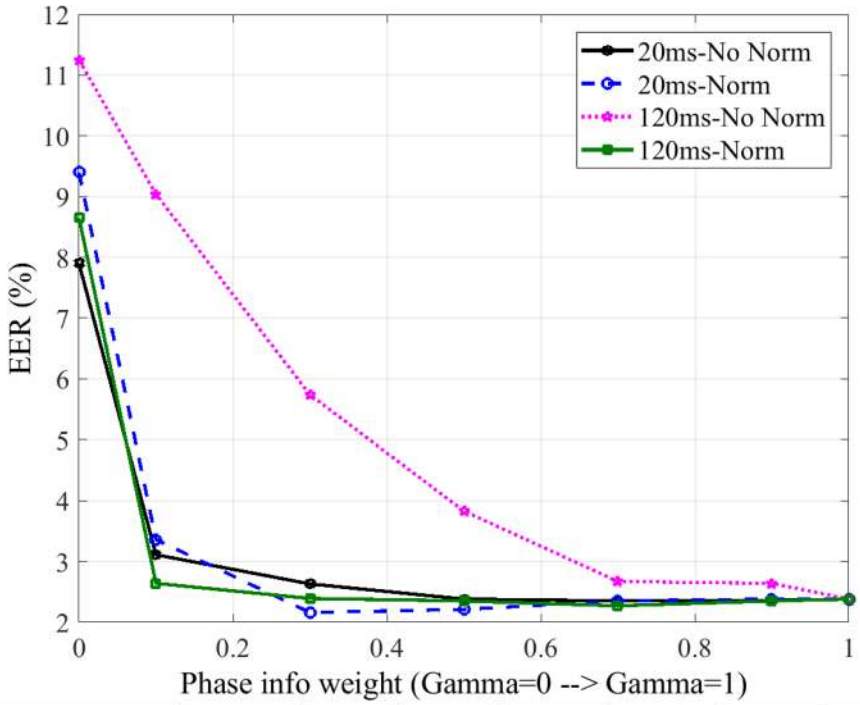

Fig. 12 Speaker verification results of the Hamming-baseline system in terms of EER (%) for combination of phase information with MFCC in the "Clean" condition and for the 20 ms and 120 ms window length sizes. Systems with phase normalization are represented with "Norm"

	γ=0	γ=0.1	γ=0.3	γ=0.5	γ=0.7	γ=0.9	γ=1
20ms-No Norm	7.90%	3.11%	2.63%	2.38%	**2.35%**	2.36%	2.38%
20ms-Norm	9.40%	3.36%	2.16%	**2.21%**	2.35%	2.38%	2.38%
120ms-No Norm	11.25%	9.04%	5.74%	3.38%	2.67%	2.64%	**2.38%**
120ms-Norm	8.65%	2.64%	2.39%	2.35%	**2.27%**	2.35%	2.38%

In this experiment and with respect to the normalization, we see that the proposed phase normalization improves the verification performance for both long and short window duration in fused systems (Fig. 12). However, the verification relative improvement for the short window duration is 3.38% better than the best result obtained for long window duration. The minimum EER for short window duration obtained when $\gamma = 0.5$ while this happens at $\gamma = 0.7$ for long duration window. Although the overall performances of short and long duration windows for proposed normalized phase information are near the same for both, the normalized phase information extracted with short duration window adds more discriminative information to the fused system. As a result, in clean-match condition short duration window can be a better choice rather than long duration window. It is worth noting that adding phase information outperforms the MFCC-only systems unanimously except for the system with the 120 ms framing size and without phase normalization (120 ms-No Norm). Now, we repeat experiments under same settings as above but under additive noise with SNR = 10 dB and in a mismatch condition to identify an appropriate frame size for the phase information extraction in noisy environment.

It is evident from Fig. 13 that the overall performance of systems follows the trend in previous experiment, that was clean environment. The long duration window for unnormalized system (120 ms-No Norm) never outperforms the MFCC-only baseline. However, the unnormalized phase information for short duration window (20 ms-No Norm) improves the system performance in term of EER by 5.54% compared to MFCC-only. The normalized phase information for short duration window (20 ms-Norm) could even achieve better accuracy and increases the RI of the verification system to 9.15%.

Other minor findings are as follows. As for the experiment in the clean condition, the MFCC-only systems always outperform the phase-only systems. Adding external noise to SNR = 10 dB causes the performance loss for the Hamming-based baseline from 2.38 to 29.92%, while this performance deterioration is smoother for phase-only systems. Using unnormalized phase information, the system with a 20 ms window size (20 ms-No Norm) obtains an EER = 2.35% ($\gamma = 0.7$), which is worse than both the other systems with phase normalization methods. Systems with a 20 ms window size and with phase normalization (20 ms-Norm), 20 ms window size without phase normalization (20 ms-No Norm), and 120 ms window size and with phase normalization (120 ms-Norm) relatively improve their MFCC-only systems in terms of EER by 7.26%, 1.13%, and 4.6%, respectively. For both the small and large window

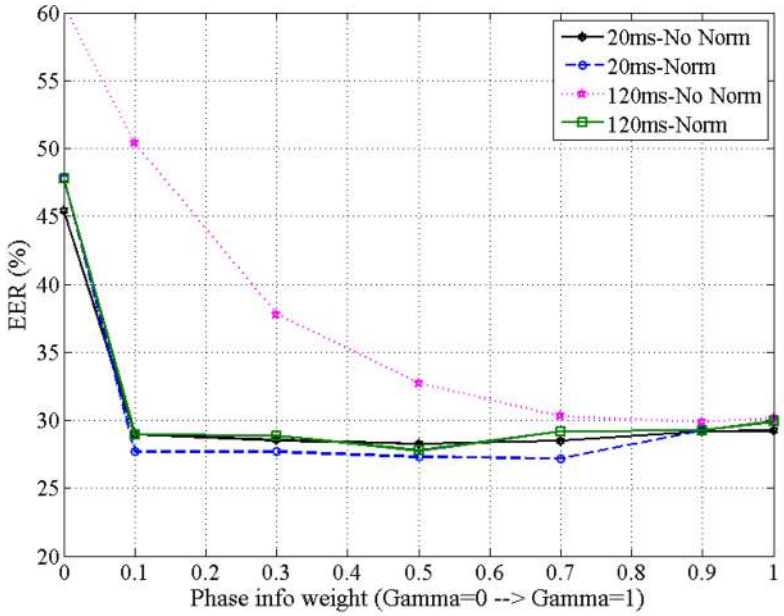

Fig. 13 Speaker verification results of the Hamming-baseline system in terms of EER (%) for combination of phase information with MFCC under additive noise and mismatched condition (SNR = 10 dB) and for the 20 ms and 120 ms window length sizes

	γ=0	γ=0.1	γ=0.3	γ=0.5	γ=0.7	γ=0.9	γ=1
20ms-No Norm	45.37%	28.93%	28.52%	**28.26%**	28.46%	29.18%	29.92%
20ms-Norm	47.80%	27.69%	27.67%	27.28%	**27.18%**	29.25%	29.92%
120ms-No Norm	60.49%	50.38%	37.76%	32.70%	30.27%	**29.90%**	29.92%
120ms-Norm	46.76%	28.93%	28.87%	**27.74%**	29.16%	29.25%	29.92%

duration, the best results from MFCC-only systems ($\gamma = 1$) outperform the best results from phase-only ($\gamma = 0$) systems.

In summary, according to the result of this section, choice of short duration window for phase information improves speaker verification performance more than long duration window. MFCC-only systems outperforms the phase-only systems in all conditions, which is consistent with the results reported earlier in the literature [12]. The results show that using the combination of MFCC and phase information outperforms MFCC-only system. The reason could be due to the fact that many of the important features of a signal are preserved if phase is properly retained. The results also support the use of phase information to complement MFCC for a better characterization of the vocal system since the fused systems improved the speaker verification performance. Moreover, normalized phase information always provides better results in compare with unnormalized phase information.

4.4 Multitaper phase-based system performance

The main goal in this paper is to improve the performance of SV by combining normalized multitaper phase information with multitaper MFCCs. We evaluated the proposed method against the state of the art, i.e. Hamming-based baseline, and report on our outcomes in the following. Here, we use short duration window for all experiments as its effectiveness for phase-based systems in both clean and noisy conditions was reported in previous section. For each of the multitaper methods-Thomson, Multipeak and SWCE—we vary the number of tapers and extract the MFCCs and normalized phase information. We retain the lowest 16 MFCCs from 40-channel Mel-frequency filterbank every 7.5 ms using windows of 30 ms (75% overlap). The number of mixtures is set to 8 for creating speaker models and the GMM-UBM model. The order of phase information components set to 12 and is calculated every 2.7 ms using windows of 30 ms, which leads to 90% overlap between frames. The phase-based features are normalized according to Eq. 11. The combination weight of MFCCs and phase information is varied from $\gamma = 0$ to $\gamma = 1$ according to Eq. 12, where $\gamma = 0$ is phase-only system and $\gamma = 1$ is MFCC-only system. The wide range of 2–12 taper counts is considered in this experiment.

For consistency with the previous baselines and settings of this experiment, we perform experiments for both the match-clean and mismatch-noisy conditions on the 30 ms Hamming-based system to have a fair comparison. The experiment is repeated for different combination of MFCC, normalized, and unnormalized phase information. As shown in Fig. 14, the combination of MFCC and normalized phase information outperforms the unnormalized phase information for both clean and noisy conditions. Adding the normalized phase information to the MFCC in the clean condition relatively improves the EER

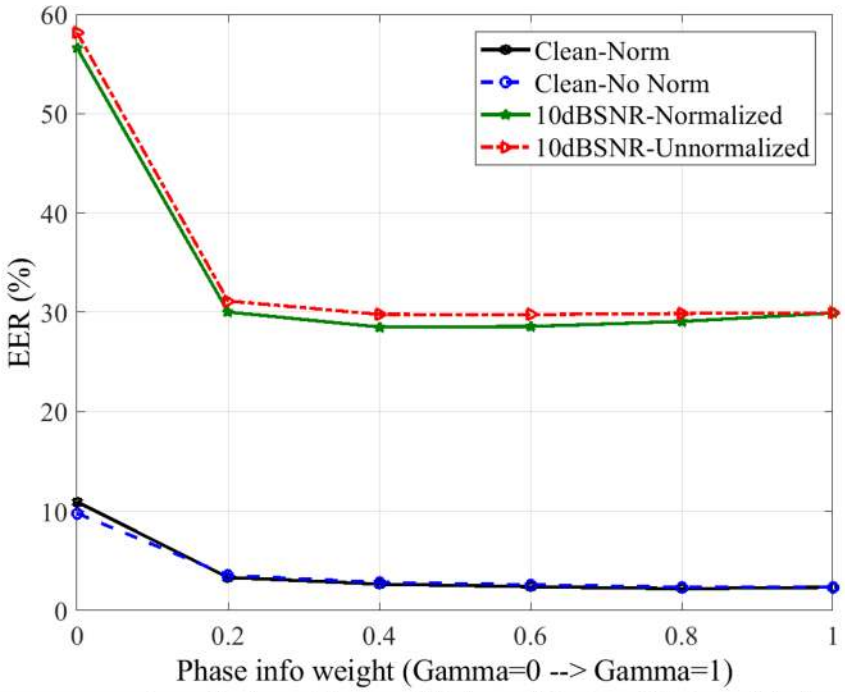

Fig. 14 Speaker verification results for the baseline system in clean and noisy conditions using a 30 ms Hamming window and a combination of MFCC and phase information

	γ=0	γ=0.1	γ=0.3	γ=0.5	γ=0.7	γ=0.9	γ=1
Clean-Norm	10.95%	3.34%	2.67%	2.40%	**2.21%**	2.35%	2.38%
Clean-No Norm	9.80%	3.55%	2.84%	2.61%	**2.36%**	2.36%	2.38%
Noisy-Norm	56.63%	32.75%	28.97%	**28.34%**	28.92%	29.27%	29.92%
Noisy-No Norm	58.17%	34.40%	29.98%	**29.73%**	29.84%	29.91%	29.92%

by 6.94% with $\gamma = 0.7$, while the RI under additive noise and for SNR = 10 dB is 5.31%. These results are in line with previous section for the 20 ms short-term framing. We use the results of this experiment as the baseline to study and compare the effects of extracting normalized phase information in multitaper-based systems.

The EERs for multitaper-based systems are compared with the Hamming-based baseline in Fig. 15 for the clean condition. The results of this experiment show that although the phase-only systems perform worse than MFCC-only systems for multitaper-based systems, a relatively high speaker verification performance is obtained for the multitaper-based system where normalized multitaper phase information is combined with multitaper MFCC. We see similar trend in Hamming-based baseline as well. It is already observed that adding normalized phase information to the Hamming-based baseline improves the EER by RI $\approx$ 7%. In another word, adding normalized phase information generally improves the performance both in Hamming-based baseline and multitaper-based system. However, in comparison with the Hamming-based baseline with normalized phase information, our proposed multitaper-based systems achieve superb performance. Specifically, multitaper-based system obtained a 33.94% relative improvement in EER where Thomson is used with normalized phase information for 6 tapers (k = 6). The maximum relative improvements obtained for Multipeak and SWCE multitapering methods with normalized phase information are 24.06% and 29.86%, respectively.

We next study the accuracy of combining MFCC and normalized phase information extracted from multitaper-based system under additive noise corruption and with SNR = 10 dB. The experimental settings of all systems under additive noise are identical to the clean condition. As shown in Fig. 16, the accuracy of all methods drops as SNR decreases to 10dB, as expected. The combination of MFCC and normalized phase information for multitaper-based systems outperforms the Hamming-based baseline in all cases. Some exceptions occur in the multitaper-based system with Multipeak and SWCE multitapering methods for different numbers of tapers, but the minimum EER obtained by each multitaper-based system is consistently lower than the Hamming-based baseline.

Moreover, MFCC-only systems outperform phase-only systems under additive noise in multitaper-based systems which is in line with the experiment in the match-clean condition. Specifically, the maximum relative improvements obtained for multitaper-based systems with Thomson, Multipeak and SWCE are 24.87%, 16.51%, and 20.32%, respectively, in comparison with the Hamming-based baseline with normalized phase information. Although the performance of the three multitapering methods at

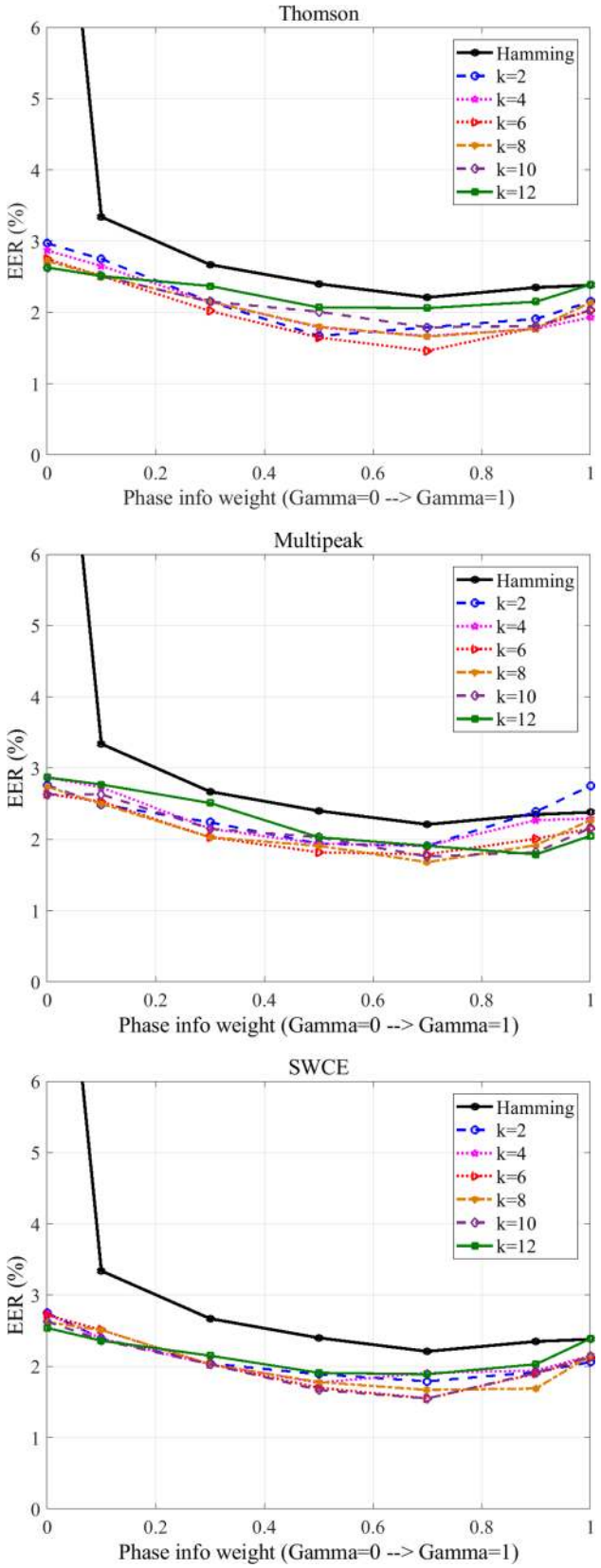

Fig. 15 Comparison of EERs (%) for combination of normalized phase information with MFCC extracted from multitaper-based systems in the clean and matched condition

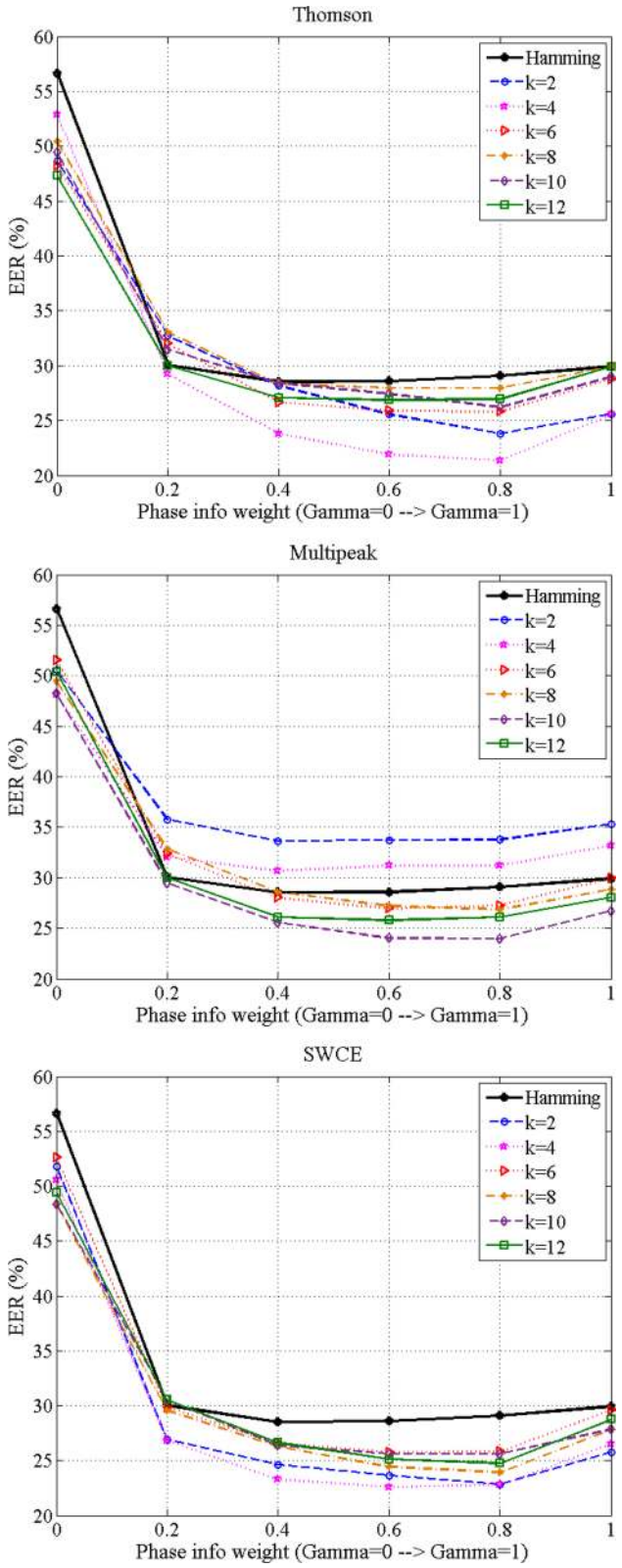

Fig. 16 EERs (%) for combination of normalized phase information and MFCC extracted by multitapering methods in multitaper-based systems compared to Hamming-based baseline under additive noise and mismatched condition (SNR = 10 dB)

their optima is near the same, the combination of phase information and MFCC for the Thomson performs best on average in the noisy condition (SNR = 10 dB). The minimum EER obtained by the Hamming-based baseline is 28.34%, while the multitaper-based system with Thomson obtains an EER of 21.29% with 4 tapers. Thomson shows sharper local minima than Multipeak and SWCE methods and yields lower error rates for different number of tapers in our experiments. One reason is that Thomson tapers are designed for flat spectra (added white noise). We also observe that the optimum value for the number of tapers depends on the method and there is no specific range for the entire multitaper-based systems.

It is worth noting that Thomson multitapering method achieves better performance both in clean and noisy conditions in comparison with other multitapering methods.

4.5 Combination of multitaper MFCC, IMFCC, and normalized phase-based features

In the previous section, we show that extracting normalized phase information using mutltipering methods could able to improve the performance of SV system. In this section, however, we are also curious to see the effect of adding the inverted-MFCC filterbank. This section evaluates the performance of the SV systems based on MFCC, IMFCC, and phase-based feature at score level fusion using multitapering method for (1) MFCC + IMFCC fused system, and (2) MFCC+ IMFCC + normalized phase information fused system. A 30 ms framing size is used for the experimental setup. The overlap between frames is set to 75% for MFCC and IMFCC, and is set to 90% for phase information extraction. The order of MFCC and IMFCC is set to 16 coefficients, while 12 components of phase information are normalized and used during the experiment. The fusion scores are obtained according to Eq. 12.

Table 1 shows the speaker verification accuracy in terms of EER for different combinations of MFCCs' complementary features. The results show that both IMFCC and phase information improve verification rates when separately combined with the MFCCs. However, when both feature sets combined with MFCC (i.e. MFCCs + IMFCCs + normalized phase-based system) the verification accuracy does not improve compared to neither MFCC+IMFCC system nor MFCC+normalized phase-based system. In fact, the worst performance among the four alternative systems for clean speech is the fusion of MFCC+IMFCC+normalized phase-based system (Fig. 17). The results also show that the combination of the normalized phase information, IMFCCs, and MFCCs is more robust in noisy condition (SNR = 10 dB). The MFCC+IMFCC+phase-based system outperforms the Hamming-based baseline, but still does not yield better EER in comparison with MFCC+IMFCC and MFCC+normalized phase-based systems. One of the explanation for this performance loss for the fused MFCC+IMFCC+phase-based system could be that GMM, like other classifiers, experiences performance degradation in the high dimensional feature settings. Therefore, the EM algorithm converges to an incorrect set of Gaussian parameters since there are many free parameters in the GMM covariance matrices of the cluster distributions to provide a good fit for many different assignments.

4.6 Experimental results using i-vector for the best performing system

For the sake of investigating the robustness of the proposed method on i-vector based system, it would be interesting to see the performance of the combination of multitaper MFCC and multitaper normalized phase information with the i-vector back-end. Since TIMIT database is not challenging for SV systems [41], repeating all the presented experiments may not necessary due to the fact that i-vectors perform better than the GMM-UBM approach in general [26]. However, for the best performing system with the related taper type, we also provide the verification results with the PLDA i-vector method. A gender-independent i-vector extractor of dimension 400 is

Table 1 Speaker verification results in term of EER (%) for different fused systems. Bolded values indicate lowest EER (%)

	Hamming	SWCE	Multipeak	Thomson
	Clean-match			
MFCC	2.4	2.1	2.1	**1.9**
MFCC + IMFCC	2.3	1.7	1.8	**1.6**
MFCC + normalized phase	2.2	1.5	1.7	**1.5**
MFCC + IMFCC + normalized phase	**2.4**	2.9	2.9	2.8
	SNR = 10 dB-mismatch			
MFCC	29.9	25.7	26.7	**25.5**
MFCC + IMFCC	28.9	23.3	24.6	**21.4**
MFCC + normalized phase	28.3	22.5	23.6	**21.2**
MFCC + IMFCC + normalized phase	28.9	24.7	24.7	**23.6**

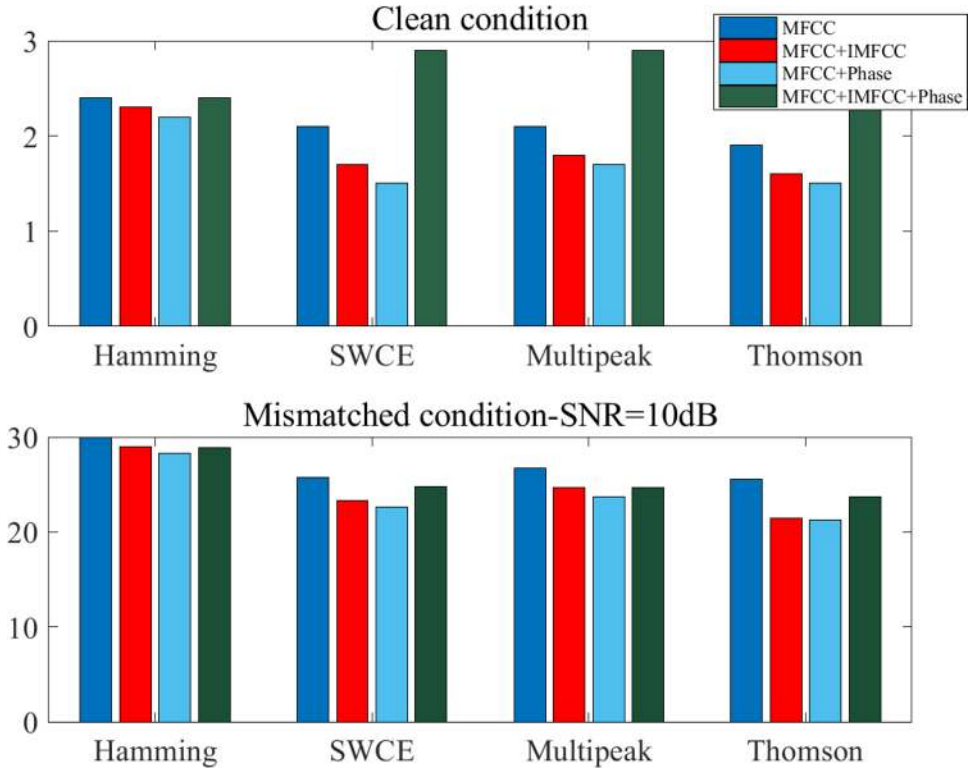

Fig. 17 Comparison performances of speaker verification fused systems in terms of EER (%)

trained in this experiment on a set of 4410 speech of TIMIT. We reduce the dimension of i-vector using Linear Discriminant Analysis (LDA) to 200. The length of the i-vector is normalized and followed by Gaussian PLDA as in [42].

Results of using Thomson multitaper with $k = 6$ in clean and match condition and $k = 4$ in noisy and mismatch condition is reported in Table 2. It is observed that the multitaper i-vector systems (i.e. the last column where $\gamma = 1$) in both clean and noisy conditions outperforms the Hamming i-vector baseline. The improvement of i-vector system is outstanding when the combination of multitaper normalized phase information and multitaper MFCC is applied. The maximum relative improvement obtained for Thomson multitapering method with normalized phase information in compare with Hamming i-vector baseline in clean and noisy conditions are 35.77% and 15.95%, respectively.

5 Summary

We proposed normalized multitaper phase-based system to improve the verification performance in both clean and noisy conditions compared with state-of-the-art single-taper Hamming-based baseline as well as unnormalize phase-based system. Combining the multitaper MFCCs with multitaper normalized phase information, we obtained the error reduction rate of 33.94%, 24.06%, and 29.86% in clean-match condition and 24.87%, 16.51%, and 20.32% in noisy-mismatch condition for Thomson, Multipeak, and SWCE, respectively, over Hamming MFCC-based method. Experimental results also demonstrate the effectiveness of our proposed method in comparison with multitaper MFCC system and multitaper phase-based system without normalization.

Table 2 Speaker verification results in term of EER (%) for the best performing GMM-UBM system using i-vector/PLDA back-end in noisy and clean condition. Bolded values indicate lowest EER (%)

		$\gamma = 0$	$\gamma = 0.1$	$\gamma = 0.3$	$\gamma = 0.5$	$\gamma = 0.7$	$\gamma = 0.9$	$\gamma = 1$
Hamming GMM-UBM	Clean	10.95	3.34	2.67	2.40	**2.21**	2.35	2.38
	10 dB	56.63	32.75	28.97	**28.34**	28.92	29.27	29.92
Hamming i-vector	Clean	7.89	2.96	2.45	2.18	**2.18**	2.29	2.35
	10 dB	54.94	28.50	21.44	**19.56**	21.12	25.76	26.63
Multitaper i-vector + phase	Clean	2.69	2.49	1.97	**1.40**	1.41	1.74	1.83
	10 dB	38.00	27.19	25.53	22.13	**16.44**	21.97	24.76

The MFCC complements, i.e. IMFCCs and phase-based features, usually are unable to outperform MFCCs in standalone systems since their ability to describe the acoustic space of a speaker is poorer than that of MFCC. The experimental results showed that the combination of the normalized multitaper phase information and MFCC was also very effective for noisy speech with SNR = 10 dB. The results showed that the type of multitaper was less important than the number of tapers. For all normalize multitaper phase-based systems, there were an optimal number of tapers that outperforms the Hamming-based system. Moreover, the exact choice of the number of tapers was not critical and the best results could be obtained when the number of tapers is between 2 and 6. However, experimental results on TIMIT dataset indicated that the proposed normalized Thomson phase-based system outperforms other multitaper-based and Hamming-based systems.

The proposed method also showed better performance improvement compared to the fusion of MFCC and high frequency cepstral feature (i.e. inverted-MFCC) with the same setup. According to the results, combining the MFCC, IMFCC, and proposed normalized phase information simultaneously would not add improvement over their fusion with MFCC separately. The largest relative improvements over Hamming-based baseline were obtained with the fusion of one of these complementary features with MFCCs.

Overall, multitapering methods for MFCC and normalized phase information extraction is a viable candidate for replacing the single-taper baseline MFCC. In the future, the performance of multitaper phase-based fused system will be studied with deep feature fusion approaches in feature level and with more focus on feature dimensionality reduction. The two fusion schemes, i.e. score level and feature level fusion, will be compared on larger database and with signals corrupted by different types of noise.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Kinnunen T, Li H (2010) An overview of text-independent speaker recognition: from features to supervectors. Speech Commun 52(1):12–40
2. Zamalloa M, Bordel G, Rodriguez LJ, Penagarikano M (2006) Feature selection based on genetic algorithms for speaker recognition. In 2006 IEEE Odyssey—the speaker and language recognition workshop, pp 1–8
3. Das RK, Mahadeva Prasanna SR (2018) Speaker verification from short utterance perspective: a review. IETE Tech Rev 35(6):599–617
4. Venturini A, Zao L, Coelho R (2014) On speech features fusion, alpha-integration Gaussian modeling and multi-style training for noise robust speaker classification. IEEE/ACM Trans Audio Speech Lang Process 22(12):1951–1964
5. Al-Kaltakchi MTS, Woo WL, Dlay SS, Chambers JA (2016) Study of fusion strategies and exploiting the combination of MFCC and PNCC features for robust biometric speaker identification. In: 2016 4th international conference on biometrics and forensics (IWBF), pp 1–6
6. Kishore KVK, Sharrefaunnisa S, Venkatramaphanikumar S (2015) An efficient text dependent speaker recognition using fusion of MFCC and SBC. In: 2015 international conference on futuristic trends on computational analysis and knowledge management (ABLAZE), pp 18–22
7. Paliwal KK, Alsteris L (2003) Usefulness of phase spectrum in human speech perception. In: Proceedings of Eurospeech'03, pp 2117–2120
8. Murty KSR, Yegnanarayana B (2006) Combining evidence from residual phase and MFCC features for speaker recognition. IEEE Signal Process Lett 13(1):52–55
9. Hegde RM, Murthy HA, Rao GVR (2004) Application of the modified group delay function to speaker identification and discrimination. In: Proceedings of IEEE international conference on acoustics, speech, and signal processing, 2004 (ICASSP'04), vol 1, pp I–517–20
10. Padmanabhan R, Sree Hari Krishnan P, Murthy HA (2009) Robustness of phase-based features for speaker recognition, Brighton, England, pp 2355–2358. Interspeech. Accessed 06 Apr 2016
11. Alam MJ, Kenny P, Stafylakis T (2015) Combining amplitude and phase-based features for speaker verification with short duration utterances. Accessed 06 Apr 2016
12. Nakagawa S, Wang L, Ohtsuka S (2012) Speaker identification and verification by combining MFCC and phase information. IEEE Trans Audio Speech Lang Process 20(4):10851095
13. Sahidullah M, Saha G (2013) A novel windowing technique for efficient computation of MFCC for speaker recognition. IEEE Signal Process Lett 20(2):149–152
14. Effects of phase on the perception of intervocalic stop consonants—ScienceDirect. Accessed 26 July 2018
15. Bahdanau D, Chorowski J, Serdyuk D, Brakel P, Bengio Y (2016) End-to-end attention-based large vocabulary speech recognition. In: 2016 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 4945–4949
16. Kim S, Hori T, Watanabe S (2017) Joint CTC-attention based end-to-end speech recognition using multi-task learning. In: 2017 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 4835–4839
17. Chorowski JK, Bahdanau D, Serdyuk D, Cho K, Bengio Y (2015) Attention-based models for speech recognition. In: Cortes C, Lawrence ND, Lee DD, Sugiyama M, Garnett R (eds) Advances in neural information processing systems, vol 28. Curran Associates, Inc., pp 577–585
18. Jain A, Sharma OP (2014) Evaluation of MFCC for speaker verification on various windows. In: International conference on recent advances and innovations in engineering (ICRAIE-2014), pp 1–6
19. Mottaghi-Kashtiban M, Shayesteh MG (2011) New efficient window function, replacement for the hamming window. IET Signal Process 5(5):499–505
20. Wang Y (2012) An effective approach to finding differentiator window functions based on sinc sum function. Circuits Syst Signal Process 31(5):1809–1828

21. Kinnunen T, Saeidi R, Sedlak F, Lee KA, Sandberg J, Hansson-Sandsten M, Li H (2012) Low-variance multitaper MFCC features: a case study in robust speaker verification. IEEE Trans Audio Speech Lang Process 20(7):1990–2001
22. Bakshi A, Kopparapu SK, Pawar S, Nema S (2014) Novel windowing technique of MFCC for speaker identification with modified polynomial classifiers. In: 2014 5th international conference confluence the next generation information technology summit (confluence), pp 292–297
23. Alam MJ, kinnunen T, Kenny P, Ouellet P, O'Shaughnessy D (2011) Multi-taper MFCC features for speaker verification using I-vectors. In: 2011 IEEE workshop on automatic speech recognition and understanding (ASRU), pp 547–552
24. Schuster A (1898) On the investigation of hidden periodicities with application to a supposed 26 day period of meteorological phenomena. Terr Magn 3(1):13–41
25. Reynolds DA, Quatieri TF, Dunn RB (2000) Speaker verification using adapted Gaussian mixture models. Digit Signal Process 10(13):19–41
26. Hansen JHL, Hasan T (2015) Speaker recognition by machines and humans: a tutorial review. IEEE Signal Process Mag 32(6):74–99
27. Thomson DJ (1982) Spectrum estimation and harmonic analysis. Proc IEEE 70(9):1055–1096
28. Riedel KS, Sidorenko A (1995) Minimum bias multiple taper spectral estimation. IEEE Trans Signal Process 43(1):188–195
29. Hansson M, Salomonsson G (1997) A multiple window method for estimation of peaked spectra. IEEE Trans Signal Process 45(3):778–781
30. Hansson-Sandsten M, andberg J (2009) Optimal cepstrum estimation using multiple windows. In: IEEE international conference on acoustics, speech and signal processing, 2009. ICASSP 2009, pp 3077–3080
31. Degottex G, Roebel A, Rodet X (2011) Phase minimization for glottal model estimation. IEEE Trans Audio Speech Lang Process 19(5):1080–1090
32. Virtanen T, Gemmeke JF, Raj B, Smaragdis P (2015) Compositional models for audio processing: uncovering the structure of sound mixtures. IEEE Signal Process Mag 32(2):125–144
33. Mowlaee P, Saeidi R, Christensen MG, Tan ZH, Kinnunen T, Franti P, Jensen SH (2012) A joint approach for single-channel speaker identification and speech separation. IEEE Trans Audio Speech Lang Process 20(9):2586–2601
34. Maia R, Stylianou Y (2014) Complex cepstrum factorization for statistical parametric synthesis. In: 2014 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 3839–3843
35. Mowlaee P, Saeidi R (2013) On phase importance in parameter estimation in single-channel speech enhancement. In: 2013 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 7462–7466
36. Mansouri A, Cardenas-Barrera J, Castillo-Guerra E (2015) A study on dimensions of feature space for text-independent speaker verification systems. In: 2015 IEEE 28th Canadian conference on electrical and computer engineering (CCECE), pp 1464–1469
37. Garofolo JS, Lamel LF, Fisher WM, Fiscus JG, Pallett DS (1993) DARPA TIMIT acoustic-phonetic continuous speech corpus CD-ROM. NIST speech disc 1-1.1. NASA STI/Recon technical report N, 93, 2. Accessed 24 Mar 2016
38. Poddar A, Sahidullah M, Saha G (2018) Speaker verification with short utterances: a review of challenges, trends and opportunities. IET Biom 7(2):91–101
39. The importance of phase in signals—IEEE Journals & Magazine. Accessed 26 July 2018
40. Beigi H (2011) Fundamentals of speaker recognition. Springer, Boston. Accessed 03 July 2015
41. Gallardo LF, Wagner M, Möller S et al (2014) I-vector speaker verification based on phonetic information under transmission channel effects. In: INTER-SPEECH, pp 696–700
42. Garcia-Romero D, Espy-Wilson CY (2011) Analysis of i-vector length normalization in speaker recognition systems. In: Twelfth annual conference of the international speech communication association