Circuits, Systems, and Signal Processing (2019) 38:5253–5288
https://doi.org/10.1007/s00034-019-01125-x

# Robust Methods for Text-Dependent Speaker Verification

Ramesh K. Bhukya[1] · S. R. Mahadeva Prasanna[1,2] · Biswajit Dev Sarma[3]

Received: 24 September 2018 / Revised: 18 April 2019 / Accepted: 19 April 2019 / Published online: 3 May 2019

**Abstract**
In this work, we explore various noise robust techniques at different stages of a *Text-Dependent Speaker Verification* (TDSV) system. A speech-specific knowledge-based robust end points detection technique is used for noise compensation at signal level. Feature-level compensation is done by using robust features extracted from *Hilbert Spectrum* (HS) of the *Intrinsic Mode Functions* obtained from *Modified Empirical Mode Decomposition* of speech. We also explored a combined temporal and spectral speech enhancement technique prior to the end points detection for enhancing speech regions embedded in noise. All experimental studies are conducted using two databases, namely the RSR2015 and the IITG database. It is found that the use of robust end points detection improves the performance of the TDSV system compared to the energy-based end points detection in both clean and degraded speech conditions. Use of noise robust HS features augmented with *Mel-frequency cepstral coefficients* further improves the performance of the system. It is also found that the use of speech enhancement prior to signal and feature-level compensation results in further improvement in performance for the low SNR cases. The final combined system obtained by using three robust methods provides a relative improvement from 6 to 25% in terms of the EER, on the RSR2015 database corrupted with Babble noise of varying strength and by around from 30 to 45% relative improvement on the IITG database.

**Keywords** End point detection · VLRs · Dominant resonant frequency · Glottal activity detection · Foreground speech segmentation · MEMD · IMFs · Hilbert spectrum · MFCCs · TDSV · DTW

## 1 Introduction

Speech is the natural mode of communication for human beings. The speech signal can be easily acquired which makes it a very attractive signal at low cost for the scientific community to use in different human–machine applications [49]. Presence

✉ Ramesh K. Bhukya
r.bhukya@iitg.ac.in

Extended author information available on the last page of the article

of speaker-specific information makes speech useful as biometric feature to recognize or authenticate a person. The biometric feature contains a multitude of information like the speakers age, height, emotion, style of speech delivery, accent, health and physiological disorders, identity, vocabulary usage, etc. [49]. There are many services and applications in which *speaker verification* (SV) could be used [12,15,39].

Among different applications involving in speech, SV has expanded remarkably over the years since its inception. SV refers to a technology that enables machines to recognize persons using his/her speech signal [34,35]. Based on the constraint imposed on the text content of the speech utterance used, SV systems can be classified into *text-independent SV* (TISV) and *text-dependent SV* (TDSV) system [23,31]. In a TDSV system, the text is fixed and the users have to utter the same text during training and testing. On the other hand, TISV does not put any restriction on text content of the speech utterance during training as well as testing speech. In TDSV system, the system takes the user's speech utterance and the identity claim as input to the system and decides whether the input speech utterance belongs to the claimed user or not. In this work, all experimental studies are presented for a TDSV system suited for the deployable systems in the practical environment and under degraded speech conditions.

To deal with degraded conditions, the compensation is done at the signal level, feature level, model level, score level or all of them. Compensation at the signal level involves detection of speech regions. In most of the methods in the literature, the signal is processed to detect the voiced regions by taking evidence from speech/non-speech frames [36]. Some of the old methods used energy, amplitude, zero-crossing rate, duration, linear prediction error energy, energy-based *voice activity detection* (VAD) and pitch for detection of speech regions [27]. These methods cannot distinguish between the true speaker's speech and other speaker's speech. Henceforth, statistical modeling methods like HMM, GMM-VAD, Laplacian–Gaussian model and gamma models were also used to detect speech regions during verification [21,32]. In [50], the authors used *glottal activity detection* (GAD) for detection of speech end points in TDSV system. Several other works in the literature detected *vowel-like regions* (VLRs) and used the detected regions for SV [46,48]. Authors in [45] proposed a method that uses independent processing of VLRs and *non-vowel-like regions* (non-VLRs) for achieving better SV performance under clean as well as degraded conditions. One of the recent method uses different speech-specific information for robust detection of speech end points [4]. This method used VLRs for detection of speech end points. Some spurious detection in the non-speech noise regions were removed using *dominant resonant frequency* (DRF) information [9], whereas some spurious detection in the speech background were removed using a *foreground speech segmentation* (FSS) algorithm [10]. The detected end points were further refined using glottal activity and dominant aperiodic region detection. These begin and end points detection method is used in the current work for signal-level compensation.

Feature-level compensation is done by using noise robust features along with the conventional features. The MFCCs are considered as the baseline features for various speech processing applications [12,34,45]. Most of the state-of-the-art SV systems also use the MFCCs derived exclusively from the magnitude spectrum of the speech utterance while neglecting its phase spectrum [49]. However, the phase spectrum of speech is equally critical to speech perception as the magnitude spectrum, and has found important use in many speech processing applications [1]. Moreover, the MFCCs may not always constitute the optimum features for all man-machine applications [24,55]. In fact, with this viewpoint, there have been many alternate avenue features introduced to the Mel filterbank, allowing to improve the performances in speech processing tasks [20]. To improve the feature extraction process, a standard strategy consisting of designing filterbanks using data-driven optimization procedure was introduced in [7]. In the same direction, different other approaches based on non-stationary data analysis techniques [9] and *Wavelet Transform* have been found to be useful in different speech processing applications [5,6,22,41,58]. In [54], an attempt had been made to explore new features for characterizing speakers, obtained from a nonlinear and non-stationary data analysis technique called *Empirical Mode Decomposition* (EMD) [26], and its variants called *Modified EMD* (MEMD) [53]. MEMD is a complete data-adaptive and AM-FM analysis-based technique which can decompose any real-world speech signal into its oscillatory or AM-FM components called the *Intrinsic Mode Functions* (IMFs). The objective in [54] was to investigate the data-adaptive filterbank nature of EMD/MEMD that could complement the Mel filterbank in the TISV task. In a recent work [52], we have an investigation on the effect of modifying the process of extracting IMFs with lesser mode-mixing, and better reflecting the higher frequency content of the speech. The same MEMD-based feature extraction method is used in this work for feature-level compensation.

In practical field deployable scenarios, the speech utterances are affected by different degradations like background noise, background speech, speech from other speakers, sensor and channel mismatch, emotional conditions and other environmental conditions, resulting in degraded speech. It has been noticed that the performance of the system falls significantly under such condition, especially when using MFCCs as the features [3,4,42,45,52]. These observations on the TDSV system highlight the issues related to the development of a system under such conditions and motivate a solution for achieving better system performance. In this work, modification is performed at two stages to improve the system performance under degraded conditions. In the first stage, a signal level compensation methods are used for removing the effect of noise from degraded speech utterances by applying the robust end point detection using speech-specific knowledge to detect the speech regions [4]. Then, in the second stage, the detected speech regions are passed through the *Hilbert spectrum* (HS), constructed from the IMFs obtained from MEMD.

Further, in this work, we also conduct experiments to combine another spectral processing method with the temporal processing method to obtain better noise reduction for improved system performance. The speech signal is spectrally enhanced using the method described in [33]. The enhanced speech signals are then passed through the robust *end point detection* (EPD) using speech-specific knowledge to detect the refined begin and end points of the speech utterances [4]. Finally, the speaker-specific features are extracted from the detected speech regions. The contribution of this paper is by combining the following three methods: (1) robust EPD using speech-specific knowledge, (2) robust features extracted from IMFs of the HS obtained from MEMD and (3) temporal and spectral techniques for speech enhancement. The popular RSR2015

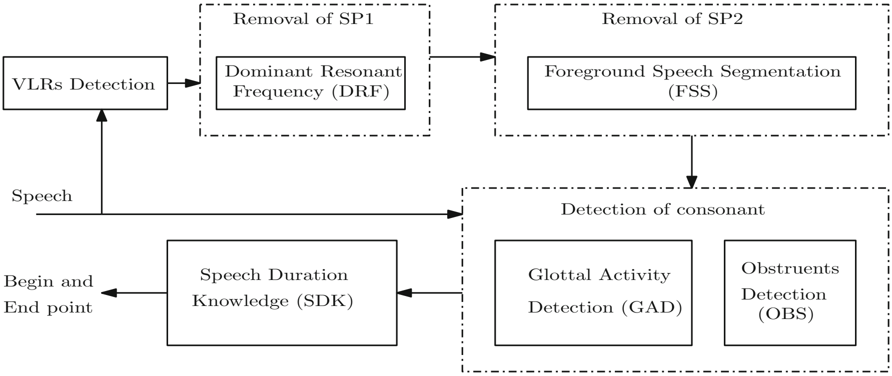
Removal of SP1
Dominant Resonant Frequency (DRF)
Removal of SP2
Foreground Speech Segmentation (FSS)
VLRs Detection
Speech
Detection of consonant
Glottal Activity Detection (GAD)
Obstruents Detection (OBS)
Speech Duration Knowledge (SDK)
Begin and End point

**Fig. 1** Block diagram of the robust EPD. SP1 and SP2 denotes spurious VLRs detected in non-speech and speech background, respectively [4]

database [34] and the IITG database [4,13,52] have been used in this work for conducting all the experiments.

The rest of the paper is organized as follows: Sect. 2 discusses the Robust End Point Detection. Section 3 discusses the robust features extracted from HS. Section 4 describes the speech enhancement techniques. Section 5 describes the experimental setup. Section 6 presents the experimental results and analysis. Finally, Sect. 7 summarizes and concludes this work.

## 2 Robust End Point Detection

Prior to extracting the features from the speech signal, robust EPD using speech-specific knowledge is performed to eliminate the silence regions and background noise regions at the beginning and end of the speech utterance. The EPD method is shown in Fig. 1. The method is based on VLRs detection. The idea is to detect the VLR onset and end points correctly so that begin and end points of speech can be searched near the onset point of the first VLR and the offset point of the last VLR, respectively. Some of the spurious VLRs are removed by using DRF information and a FSS algorithm. Some speech regions at the beginning and end of the speech region are detected using GAD and dominant aperiodic region detection.

### 2.1 Vowel-Like Region (VLR) Detection

Motivation behind the use of VLR for begin and end point detection is the high energy nature of the VLRs which makes them high *signal-to-noise ratio* (SNR) region, and therefore, they are less affected by noise degradation [48]. VLR detection involves detection of its begin and end points, namely, the VLR onset points (VLROPs) and VLR endpoints (VLREPs) [29,45,46].

The detection of VLROP and VLREP is performed by using excitation source information derived from *Zero Frequency Filtered Signal* (ZFFS) [43] and Hilbert envelope of LP residual of speech (HE of LP) [38,44,47]. The evidences acquired from these two methods are enhanced by adding amplitude envelope evidence [45], and the begin and end points are detected more accurately.

Evidence from HE of LP residual of speech is derived as follows: First HE of LP residual of speech is computed which enhances the excitation source information about glottal closure instants (GCIs). The excitation contour is smoothed by taking the maximum value of the HE of LP residual for each 5 ms block with one sample shift which is then convolved with a first-order Gaussian differentiator (FOGD) of length 100 ms and a standard deviation of one-sixth of the window. This convolved output is termed as VLROP evidence using excitation source information. Then, the evidence for VLREP is obtained by doing convolution from right to left, instead of left to right as in the case of VLROP.

Evidence from ZFFS is computed as follows: The first-order difference of the ZFFS preserves the signal energy around the impulse present at zero frequency and removes all other information. The first-order difference is also known as the strength of excitation at the epochs [45]. The second-order difference of ZFFS contains change in the strength of excitation. This change is detected by convolving with a 100 ms long FOGD utilizing a standard deviation of one-sixth of window length. The convolved output termed as VLROP evidence. The VLREP evidence is obtained by convolving from right to left.

The final VLROP and VLREP information is derived by adding the two evidences [45]. The combined evidence is then normalized by the maximum value of the sum. The locations of peaks between two successive positive to negative zero crossings of the combined evidence represent the hypothesized VLROP or VLREP.

### 2.2 Removal of Spurious VLRs in Non-speech Region Using DRF

The vocal-tract information are captured from the spectrum in the form of dominant resonances associated with the shape of the particular cavity in the vocal tract responsible for the production of the speech segment. These resonance peaks are called DRF. DRF is the frequency which is resonated most by the vocal tract. DRF is computed from the Hilbert envelope of numerator group delay spectrum of zero time windowed speech [2]. For VLRs, DRF value is mostly less than 1 kHz and the non-speech noises mostly contain high-frequency components. This knowledge is used for identifying and removing the spurious VLRs in the non-speech region [4]. The VLRs having DRF more than 1 kHz are removed from the output of the VLR detection.

### 2.3 Removal of Spurious VLRs in Background Speech Using FSS

To remove the background speech, a FSS is used. FSS algorithm was proposed in [11], which was further modified in [10]. In this paper, the modified version of the FSS method is used to remove the background speech [4]. The method uses both source and system information. Excitation source information is extracted using ZFFS analysis and vocal-tract system features extracted from the modulation spectrum energy. The spurious VLRs in the background speech region are removed, and the VLRs present in the foreground speech region are retained.

### 2.4 Glottal Activity and Obstruents Detection

Once the VLRs are detected, GAD can be explored to add a sonorant consonant at the begin or at the end of speech utterance. The GAD method proposed in [14] is used in this work. GAD detects VLRs as well as sonorant consonant regions. Therefore, it enables better localization of voiced regions and helps to minimize the misses in the VLR detection output. This, in turn, helps to detect the appropriate end points of the speech utterance.

There may be an obstruent consonant at the begin and end of the speech utterance. To include them in the speech region, obstruent detection is performed. Since the aperiodic component is dominant in burst and frication region of most of the obstruents, the dominant aperiodic region detection method proposed in [51] is used to detect the obstruents. The first VLROP and last VLREP are considered as the refined begin and end points and are detected with more accuracy after using obstruent detection.

### 2.5 Speech Duration Knowledge for Further Refining the End Points

Finally, *speech duration knowledge* (SDK) is used to further remove the spurious VLRs. In many cases, one may repeat the utterance twice and can talk something which is not part of the speech utterance. In these situations, it is not feasible to make any rectifications. However, if the user uttered at the begin or end leaving some silence region between actual speech utterance and the extra word. Then, there is a viable to discard the extra word using SDK. The SDK method for refining begin and end point detection is as follows:

First, identify all the locations of the VLROPs in the speech utterance and compute the average of the Euclidean distances from one VLROP to all other VLROPs. The VLROP having the minimum average Euclidean distance is marked as the center. Then, starting from the center of the speech, the duration between two successive VLRs is computed on either side until the peripheral VLR or the duration between two successive VLRs is greater than 300 ms [48]. At this point among the two VLRs, the one which is closer to center is declared as the peripheral VLR. All VLRs outside the peripheral VLRs are removed. Finally, the VLROP of first VLR and VLREP of last VLR are declared as the begin and end points of the speech utterance, respectively [4].

Figure 2 shows speech signal for the utterance “Lovely pictures can only be drawn” with background noise include both speech and non-speech background noise. Figure 2a shows the speech utterance with non-overlapping background noise and background speech. Figure 2b shows the VLROP and VLREP evidences and detected VLRs, respectively. The non-speech background noise present in between 0 and 0.5 s and background speech present in between 2.25 and 3.5 s are also detected as VLRs due to impulse-like characteristics. Second step is to remove such missed spurious VLRs, DRF information is used. If DRF is less than 1 kHz for VLRs and more than

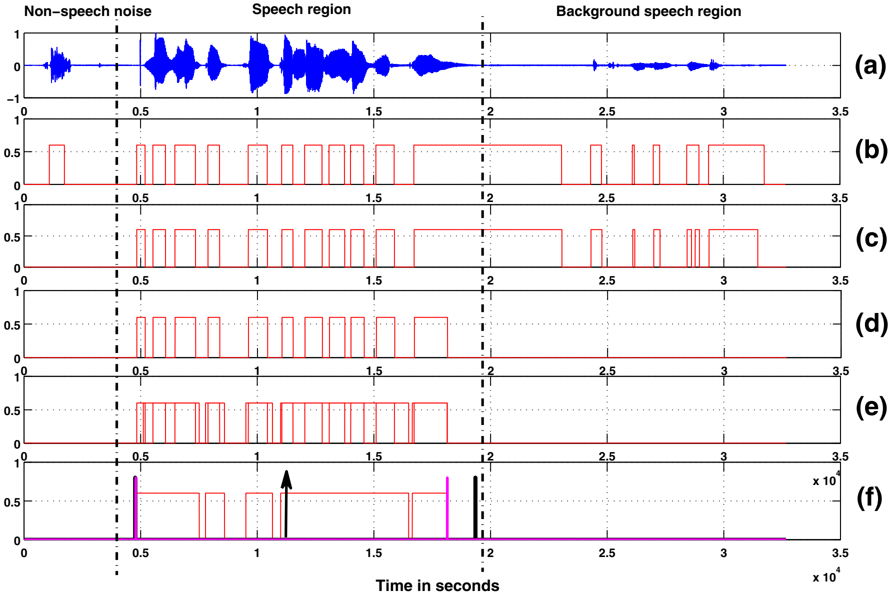

**Fig. 2** Illustration of the begin and end point detection procedure. **a** Speech signal containing non-overlapping background noise and background speech. **b** Detected speech regions of VLROP and VLREP are used to obtain the VLRs. **c** VLRs after discarding the background noise using DRF information. **d** VLRs after discarding the background speech using FSS. **e** Detected GADs added to the VLRs. **f** Refined begin and end points using obstruent information. The center *C* of the speech utterance shows the arrow around 1.2 s. The duration between successive VLRs are less than 300 ms or the peripheral VLR is reached. Finally, the VLROP of first VLR and VLREP of last VLR are detected using SDK knowledge. Dotted line shows the manually marked VLROP and VLREP points in the speech region

1 kHz are considered as non-speech region are removed and refined VLRs are shown in Fig. 2c. The third step is to remove the background speech which is obtained using FSS. With the help of FSS, the detected spurious VLRs in the background speech region are removed and refined VLRs are obtained after FSS are shown in Fig. 2d. In the next step, the glottal activity regions are explored to add a sonorant consonant at the begin or at the end of the speech utterance for better localization. The detected GADs are obtained for appropriately detecting the begin and end points as shown in Fig. 2e. Fifth step is to detect the obstruent consonant at the begin and end of speech utterance. The refined begin and end points are detected with more accuracy using obstruent evidences are shown in Fig. 2f. The arrow in Fig. 2f shows the center of detected VLRs of the speech utterance. Starting from the center of the speech, the duration between two successive VLRs are computed on either side until the duration is found to be greater than 300 ms or the peripheral VLR is reached. Thereafter, no further modification is incorporated using SDK knowledge. The dotted lines on either side show ground truth manual marking.

## 3 Robust Features from Hilbert Spectrum of MEMD

The EMD is a data-adaptive technique, which can decompose the speech signal into a finite number of components, called IMFs, without the need of any *a priori* basis [55]. Every speech signal has its unique and meaningful decomposition. Again, the signal changes dynamically and varies with resonant structure of the speech signal. The changed resonant structure distributed among its unique set of IMFs, which are obtained without any *a priori* basis. MEMD explored in different real-world applications [8,18,28,54]

### 3.1 Distribution of Mode-Mixing for MEMD

The mode-mixing makes the IMFs less narrowband, leading to less accurate estimation of their instantaneous frequencies and amplitude envelopes. Moreover, this makes it difficult to segregate or characterize a certain subset of the IMFs, as being useful for analysis, for a particular task. Hence, in this work, we utilize a recently proposed variants of EMD—the MEMD—which reduces mode-mixing in the IMFs [55].

Practically, the decomposition is stopped when a user-defined maximum $M$ number of IMFs, has been extracted. For a digital speech signal, $s(n)$.

$$s(n) = r_M(n) + \sum_{i=1}^{M} h_i(n) = \sum_{i=1}^{M+1} h_i(n) \tag{1}$$

where $h_i(n)$ represents the decomposition of the signal in its IMFs, and $r_M(n)$ the final residue, which is a trend-like signal and $M$ is the total number of IMFs extracted. In Fig. 3, the first 5 IMFs obtained from MEMD of a digital speech signal $s(n)$, taken from the RSR2015 database. This represents low- and high-frequency oscillations present at different instants within the same IMF or distributed among multiple IMFs. This process is called *mode-mixing*. These components spread across different IMFs at different instants of time, leading to a less accurate number of extrema and number of zero crossings differ by utmost one. The reliable IMFs of MEMD may be used to be better suited for AM-FM analysis. To avoid unnecessary generation and processing of higher-order IMFs, the decomposition reduced to a maximum of 10 components ($M = 9$), for the MEMD method [52].

### 3.2 Dynamic Changes of Instantaneous Frequencies and Amplitude Envelopes

Having obtained the IMF components, their center frequencies decreases as the order of IMF increases. The Hilbert transform is applied to each component to compute dynamically changing instantaneous frequency and amplitude envelope. The Hilbert transform, $H[x(t)]$, of a signal $x(t)$, is computed from Fourier transform.

$$x(t) \longleftrightarrow X(f), \quad \frac{1}{\pi t} \longleftrightarrow -j\,sgn(f) = \begin{cases} -j, & f > 0 \\ j, & f < 0 \end{cases},$$

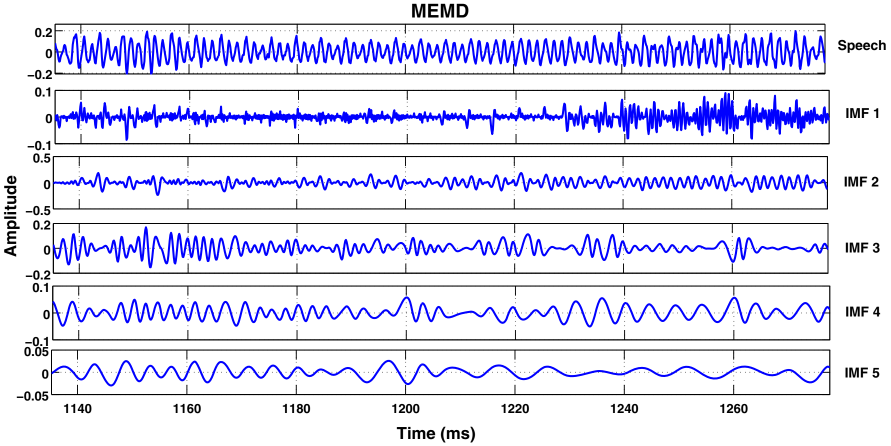

**Fig. 3** IMFs 1–5 generated from a speech signal, $s(n)$, using MEMD

$$H[x(t)] = x(t) * \frac{1}{\pi t} \longleftrightarrow -j\, sgn(f)X(f) = \begin{cases} -jX(f), & f > 0 \\ jX(f), & f < 0 \end{cases}$$

The instantaneous frequency function, $f(t)$, and amplitude envelope function, $A(t)$, is derived from the analytical signal, $x_A(t)$, which is free from any negative frequency Fourier components.

$$A(t) = |x_A(t)|, \quad \phi(t) = \arctan \frac{\Im\{x_A(t)\}}{\Re\{x_A(t)\}}, \quad f(t) = \frac{1}{2\pi}\frac{\mathrm{d}\phi(t)}{\mathrm{d}t} \tag{2}$$

Correspondingly, the *discrete Fourier transform* (DFT) method [49] is used for estimating the instantaneous frequency and amplitude envelope of any discrete-time signal, $x(n)$ [49]. Henceforth, if $A_k(n)$ and $f_k(n)$ represent the amplitude envelope and instantaneous frequency of $h_k(n)$, respectively. The time-frequency distribution of the energy envelope is the squared magnitude of the amplitude envelope. This formulation, when represented in a complete, compact and adaptive Fourier representation in terms of an image, is called the *Hilbert Spectrum* (HS) [17,25,26].

$$H(f,t) = \left\{A_k^2(n) \mid f_k(n), n\right\}, \quad k = 1, \ldots, K \leq M + 1 \tag{3}$$

Figure 4 represents the Hilbert spectra of the speech signal used in Fig. 3, obtained using MEMD, where fixed number of components ($K = M + 1 = 10$) are used. Readers are advised to refer to the soft-copy, rather than hard-print of the manuscript, for better visualization of the figure. Figure 4 shows the HS for a section of the speech utterance of the last few components, which are low-frequency trend-like waveforms, which are excluded from the spectrum, as they have high energy and obscure the image. This is evident from the spectrum, most of the energy in the spectrum lies within 100–700 Hz (particularly, one can observe more dots), which is the pitch frequency range, i.e., the frequency range of vibration of the vocal folds in the glottis (during

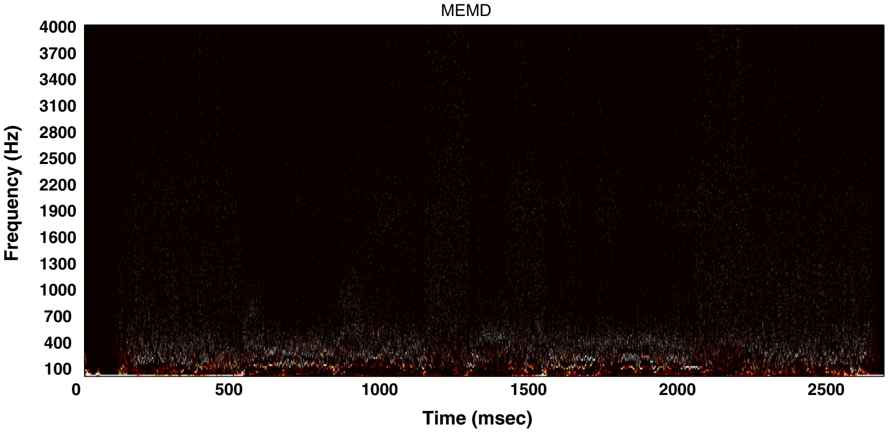

**Fig. 4** Hilbert spectrum of the speech signal used in Fig. 3, constructed from the IMFs of MEMD

the production of voiced speech) [52]. This spectrum can be post-processed to obtain instantaneous pitch frequency, which constitutes the HS, obtained from MEMD, for carrying speaker-specific information [55].

### 3.3 Characteristics of HS and MFCC Features

The Mel filterbank has a fixed structure, whereas the MEMD filterbank has an adaptive structure at every frame, both capture speaker-specific information using their specific methods. The HS represents the instantaneous frequencies, and amplitude envelopes of the IMFs are processed in short-time segments to generate features. These features capture different sort of speaker information for characterizing the speakers.

Figure 5 depicts the power spectra of the first 5 IMFs, corresponding to a 20-ms segment of the speech signal of Fig. 3. As is evident from the figure, the power spectra of the IMFs represent different portions of the speech spectrum, as if they have been band pass filtered. However, in this case, the entire process takes place in the time domain to manifest the characteristics of an adaptive filterbank. To illustrate the difference between this adaptive and the Mel filterbank, the center frequencies of a 22-filter Mel filterbank are plotted, in Fig. 6, along with the *mean frequencies* of the IMFs [53]. For a 20-ms segment of the speech signal, $s(n)$, the mean frequency of the corresponding segment of its $k$th IMF is obtained as:

$$F_k^m = \sum_{f=0}^{F_s/2} \frac{f \times S_k(f)}{\sum_{f=0}^{F_s/2} S_k(f)}, \quad k = 1, \ldots, M+1\,, \tag{4}$$

where $F_s = 8\,\text{kHz}$ is the *sampling frequency* of $s(n)$, and $S_k(f)$ is the power spectrum of 20-ms segment of $k$th IMF. $f$ represents the analog frequencies corresponding to the digital frequencies of the DFT spectrum of IMF. From Fig. 6, the MEMD and Mel

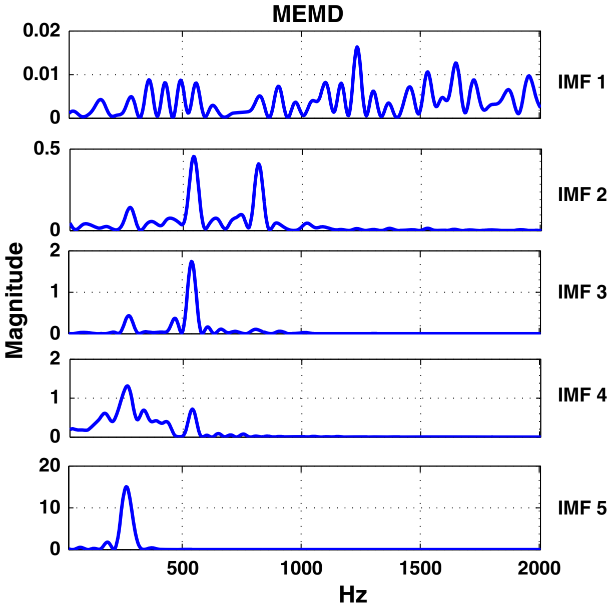

**Fig. 5** Squared magnitude spectra of each of IMFs 1–5 generated by MEMD of a 20-ms segment of the speech signal used in Fig. 3

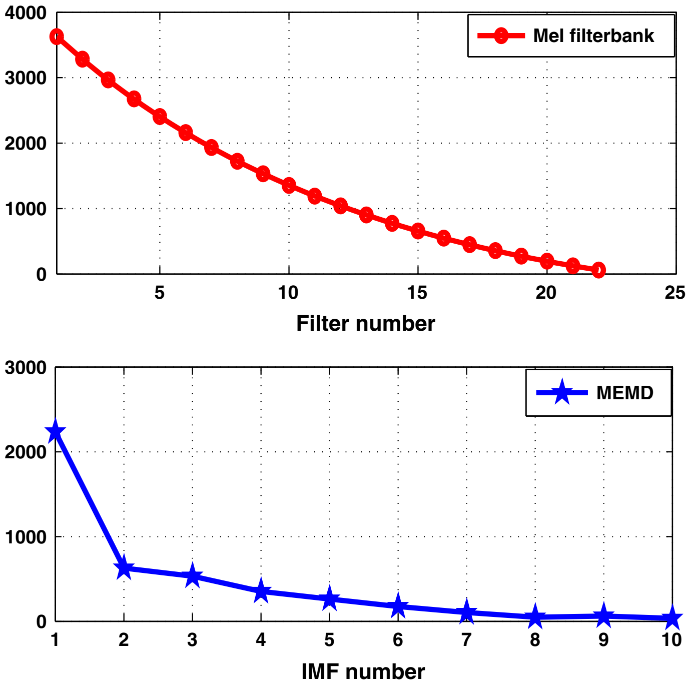

**Fig. 6** **a** Center frequencies of a 22-filter Mel filterbank; **b** Mean frequency of the IMFs derived from the speech signal of Fig. 3

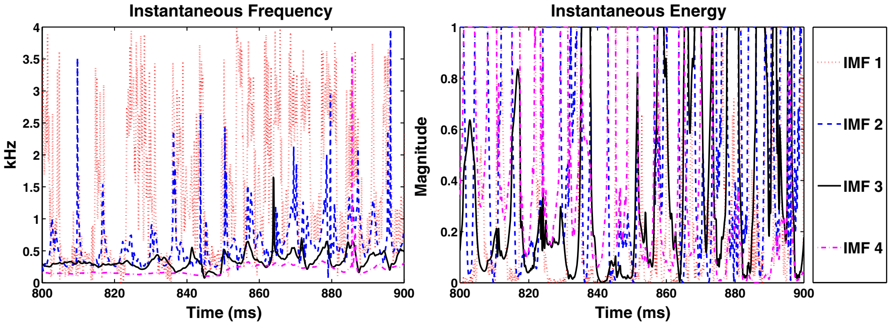

**Fig. 7** Instantaneous frequencies (kHz) and normalized Instantaneous energies of the first 4 IMFs (derived using MEMD) corresponding to the HS shown in Fig. 4

filterbank are different. The center frequencies are changes as the nature of the signal changes [57].

### 3.4 Instantaneous Frequencies and Energies of the Hilbert Spectrum

The features extracted from HS constitute the instantaneous energy envelopes and instantaneous frequencies (derived from the instantaneous phases) of the IMFs. For visual clarity, Fig. 7 shows the first 4 IMFs, corresponding to a 20 ms segment of speech from Fig. 3 is constructed. Thus, the instantaneous energy values are normalized at each time instant, and the frequencies are represented in kHz to reduce the dynamic range. Hence, it is beneficial to extract features from them, after some degree of averaging over the time segment. This reduces the feature space and allows the features derived to be concatenated with the MFCCs, which are obtained for every frame, after dividing the entire speech utterance into overlapping frames.

However, the instantaneous frequencies and energies need to be redistributed. If no mode-mixing is present, then, at every speech frame, the first IMF, $h_1(n)$, would produce the highest instantaneous frequency. The second highest frequency would be given by $h_2(n)$, the third highest by $h_3(n)$, and the fourth highest by $h_4(n)$. Based on this, the four instantaneous frequencies at every time instant are sorted in descending order of frequency. These sorted instantaneous frequencies and energies are then rearranged. Figure 8 shows the rearranged IMFs as shown in Fig. 7, and contains speaker-specific information, which may be used for extracting features for the task of TDSV.

### 3.5 Robust Feature Extraction from the Instantaneous Frequencies and Energies

Based on the preceding discussion, different features are extracted from the IMFs of the speech. Let $K \leq M + 1$ be the number of IMFs from which the instantaneous frequencies, $\{f_k(n), \quad k = 1, \ldots, K\}$, and the instantaneous energies, $\{|A_k(n)|^2, \quad k = 1, \ldots, K\}$ are extracted. From 20 ms frame size and with a frame shift of 10 ms, $s^j(n)$ represent the $j$th frame of speech signal $s(n)$. Then, $h_k^j(n)$ represents the $j$th frame

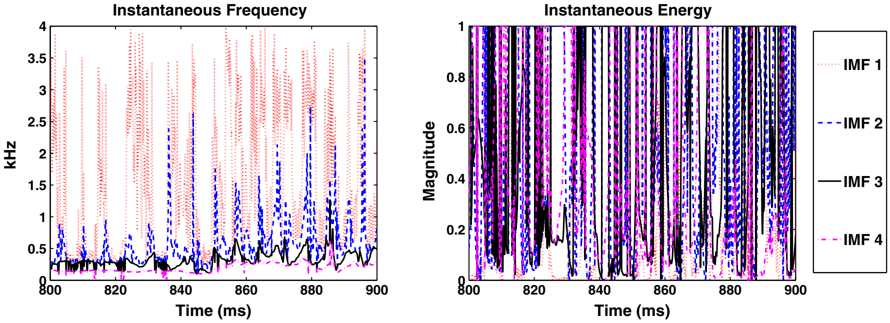

**Fig. 8** Instantaneous frequencies (kHz), sorted in decreasing order, and the corresponding rearranged normalized instantaneous energies of the first 4 IMFs (derived using MEMD) corresponding to Fig. 7

of the $k$th IMF of $s(n)$. Correspondingly, $f_k^j(n)$ and $|A_k^j(n)|^2$ are the instantaneous frequencies and energies, respectively, of $h_k^j(n)$, extracted using HS. Let $N_f$ be the number of samples in a 20 ms frame. The following features are extracted from each speech frame. It has been shown that the first few IMFs of the speech signal show the vocal-tract resonances of the speech utterance. In [52], the IMFs obtained from HS have been utilized to estimate the vocal-tract resonances produced by the different cavities that are formed when the speaker utters. These cavities depends on the sound produced and physical structure of the vocal tract, which is never be same for two different speakers. Hence, the first few IMFs are used for the task of TDSV system to carry important speaker-specific information. More specifically, HS constitutes different frequency bands and their corresponding energy bands to represent useful speaker characteristics that can complement the MFCCs.

**Mean Instantaneous Frequency:** It is derived as:

$$F_k^j = \frac{1}{N_f} \sum_{n=0}^{N_f-1} f_k^j(n)/1000, \quad k = 1, \ldots, K \leq M+1 \tag{5}$$

$$F_K^j = [F_1^j, F_2^j, \ldots, F_K^j]^{\mathrm{T}}, \quad F_K = \{\varGamma\{F_K^j\}, \forall j \in \mathbb{N}\} \tag{6}$$

The $F_K$ feature is used to capture the dominant frequencies of the different frequency bands of the IMFs obtained from MEMD. This feature may be expected to carry speaker-specific cues for characterizing the speakers.

**Absolute Deviation of Instantaneous Frequency:** It is derived as:

$$\Delta F_k^j = \frac{1}{N_f} \sum_{n=0}^{N_f-1} \left| \frac{f_k^j(n)}{1000} - F_k^j \right|, \quad k = 1, \ldots, K \leq M+1 \tag{7}$$

$$\Delta F_K^j = [\Delta F_1^j, \Delta F_2^j, \ldots, \Delta F_K^j]^{\mathrm{T}}, \quad \Delta F_K = \{\varGamma\{\Delta F_K^j\}, \forall j \in \mathbb{N}\} \tag{8}$$

For a speech signal, some frequency bands of the HS show large variations, whereas other frequency bands show comparatively steady in the HS obtained from MEMD.

The $\Delta F_K$ feature is used to capture these variations, spreads or widths of the first $K$ frequency bands arranged in the decreasing order of frequencies. This $\Delta F_K$ feature might be useful in characterizing the speaker information in a better way.

**Correlation between Instantaneous Frequencies:** It is derived as,

$$\sigma F_k^j = \sum_{n=0}^{N_f-1} \left[ \{f_k^j(n) - F_k^j\}\{f_{k+1}^j(n) - F_{k+1}^j\} \right], \quad k = 1, \ldots, K-1 \leq M \tag{9}$$

$$\sigma F_{K-1}^j = [\sigma F_1^j, \sigma F_2^j, \ldots, \sigma F_{K-1}^j]^{\mathrm{T}}\ , \sigma F_{K-1} = \{\Gamma\{\sigma F_{K-1}^j\}, \forall j \in \mathbb{N}\} \tag{10}$$

There is a possibility that certain frequency bands may vary in a particular manner. On the other hand, a different pair of nearest frequency bands may be closely related with each other. For such a speech utterance, $\sigma F_{K-1}$ feature capture the dependence or correlation between successive frequency bands of the HS. The $K$ frequency bands and their $K-1$ successive frequency bands correlation values are obtained. Therefore, $\sigma F_{K-1}$ inter-band relations could serve as an useful speaker-specific cues.

**Mean Instantaneous Energy:** It is derived as,

$$|A_k^j(n)|^2 := \frac{|A_k^j(n)|^2}{\max\limits_k |A_k^j(n)|^2}, \quad k = 1, \ldots, K \leq M+1 \tag{11}$$

$$E_k^j = \frac{1}{N_f} \sum_{n=0}^{N_f-1} |A_k^j(n)|^2, \quad k = 1, \ldots, K \leq M+1 \tag{12}$$

$$E_K^j = [E_1^j, E_2^j, \ldots, E_K^j]^{\mathrm{T}}, \quad E_K = \{\Gamma\{E_K^j\}, \forall j \in \mathbb{N}\} \tag{13}$$

The $E_K$ feature denotes the average energy (amplitude envelope) of the different frequency bands. For the first $K$ frequency bands, there are equivalent number of energy bands. Therefore, $E_K$ represents the mean value of these $K$ energies at different frequencies. Hence, $E_K$ feature could be useful in discriminating the speakers.

**Absolute Deviation of Instantaneous Energy:** It is derived as,

$$\Delta E_k^j = \frac{1}{N_f} \sum_{n=0}^{N_f-1} |\ \{|A_k^j(n)|^2 - E_k^j\}\ |, \quad k = 1, \ldots, K \leq M+1 \tag{14}$$

$$\Delta E_K^j = [\Delta E_1^j, \Delta E_2^j, \ldots, \Delta E_K^j]^{\mathrm{T}}, \quad \Delta E_K = \{\Gamma\{\Delta E_K^j\}, \forall j \in \mathbb{N}\} \tag{15}$$

A particular speaker can emit certain frequencies at varying strengths, whereas other speaker can only emit other frequencies at specific strengths. One can observe large energy variations in certain frequency bands (energy bands), whereas other frequency bands (energy bands) may be relatively steady. Therefore, $\Delta E_K$ feature represents the variation of energy in different frequency bands. These cues may be served as an important speaker-specific cue.

**Correlation between Instantaneous Energies:** It is derived as,

$$\sigma E_k^j = \sum_{n=0}^{N_f-1} \left[ \{|A_k^j(n)|^2 - E_k^j\}\{|A_{k+1}^j(n)|^2 - E_{k+1}^j\} \right], \quad k = 1, \ldots, K-1 \leq M \tag{16}$$

$$\sigma E_{K-1}^j = [\sigma E_1^j, \sigma E_2^j, \ldots, \sigma E_{K-1}^j]^{\mathrm{T}}, \quad \sigma E_{K-1} = \{\Gamma\{\sigma E_{K-1}^j\}, \forall j \in \mathbb{N}\} \tag{17}$$

The $\sigma E_{K-1}$ feature captures the relation between successive frequency bands in terms of their energy variations. Hence, the increase in energy in a particular energy band affects its succeeding energy band. Therefore, these energy variations may be speaker specific and hence serve as an important feature for each speaker.

**Instantaneous Energy weighted Instantaneous Frequency:** It is derived as

$$\Upsilon_k^j = \sum_{n=0}^{N_f-1} \frac{|A_k^j(n)|^2}{\sum_{k=0}^{K} |A_k^j(n)|^2} f_k^j(n), \quad k = 1, \ldots, K \leq M+1 \tag{18}$$

$$\Upsilon_K^j = [\Upsilon_1^j, \Upsilon_2^j, \ldots, \Upsilon_K^j]^{\mathrm{T}}, \quad \Upsilon_K = \{\Gamma\{\Upsilon_K^j\}, \forall j \in \mathbb{N}\} \tag{19}$$

For different speakers, the correlation may vary. To capture such correlation, $\Upsilon_K$ feature merges the information contained in instantaneous frequencies and energies. The first $K$ frequency bands (in decreasing order of frequency) and their corresponding $K$ energy bands can be weighted. Therefore, the $\Upsilon_K$ feature may be expected to capture speaker-specific cues [52].

Figure 9 represents the block diagram of the feature extraction process used for the TDSV system. The features are extracted from the instantaneous frequencies and energies. Along with HS features, the conventional 39-dimensional MFCCs, 51-dimensional Ext. MFCCs, and the $cG_K$ [52,54,57] features are also derived from every speech frame. These features are used with the expectations that their combinations with MFCCs would significantly enhance the capability of the TDSV system.

## 4 Speech Enhancement Techniques for TDSV System

The main concern of the work presented in this section is to use a combined temporal and spectral enhancement method for enhancing speech under degraded conditions. This method can be successfully used for identifying and enhancing speech-specific components from the degraded speech. The temporal processing method involves identification and enhancement of speech-specific components present at the gross and fine levels [33]. The evidences obtained by using gross level components are as follows: first the gross level components are used to detect high SNR regions using the sum of the 10 largest peaks in the DFT spectrum which represents the vocal-tract information. The second evidence is obtained from the smoothed HE of LP residual of speech representing the excitation source information. The third evidence is the modulation spectrum which represents the suprasegmental information of speech. The origin of

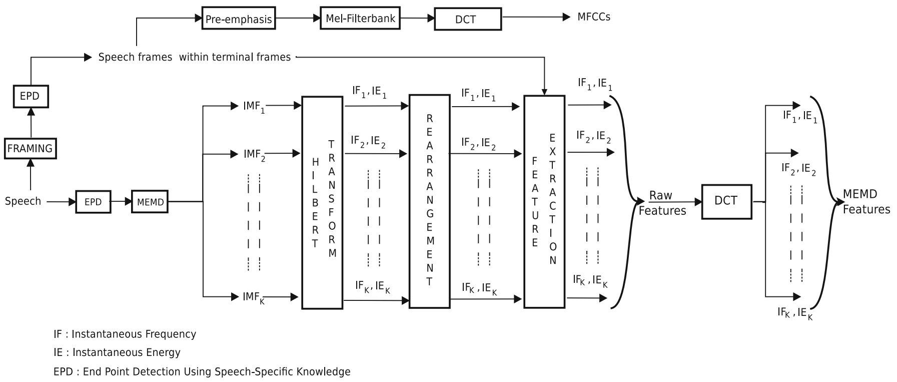

**Fig. 9** Feature extraction process for the TDSV system

these three approaches are different and hence combining them together improves the robustness and detection accuracy as compared to individual processing methods [33]. The gross weight function $w_g(n)$ computed from these three evidences are summed up together, normalized and nonlinearly mapped using the mapping function [33].

$$w_g(n) = \frac{1}{1 + e^{-}\lambda(s_i(n) - T)} \tag{20}$$

where $\lambda$ is the slope parameter and $w_g(n)$ is the nonlinearly mapped values of the normalized sum $s_i(n)$ and $T$ is the average value of $s_i(n)$. The gross weight function is obtained by computing the deviation between spectrally processed speech w.r.t direct speech.

The fine level components are identified using the knowledge of the instants of significant excitation which mostly correspond to the epoch locations [33]. Therefore, using HE of LP residual one can extract the robust epoch locations from the speech signal. From HE of the LP residual perspective, an approximate location of instants is sufficient because the enhancement is commonly achieved by emphasizing the residual signal in the speech regions around the instants of significant excitation. The speech regions around the instants of significant excitation are used as fine level evidence. The epoch locations are used for obtaining the fine weight function [33]. The region around the instants of significant excitation are convolved with a Hamming window which has a 3 ms temporal duration. Therefore, the fine weight function is derived as:

$$w_f(n) = \left(\sum_{k=1}^{N_K} \delta(n - i_k)\right) * h_w(n) \tag{21}$$

where $N_k$ is the total number of epochs located in speech frame, $i_k$ is estimated location of speech.

A total weight function $w(n)$ is obtained by multiplying the gross weight function $w_g(n)$ with the fine weight function $w_f(n)$ which is represented as:

$$w(n) = w_g(n) * w_f(n) \tag{22}$$

The temporally processed speech can be obtained by synthesizing as follows:

$$S_k(z) = \frac{R_w(z)}{1 + \sum_{n=1}^{M} a_n z^{-n}} \tag{23}$$

where $S_k(z)$ represents the temporally processed speech and $R_w(z)$ is the weighted LP residual used to excite the time-varying all-pole filter derived from the degraded speech to generate the enhanced speech and $a_n$ is the LP filter coefficients.

The temporally processed speech is further subjected to improve the vocal-tract characteristics at the spectral domain. Whereas the temporal processing approach enhances the speech region around the instants of significant excitation, spectral processing approach enhances speech-specific components and suppresses the noise components in the spectral domain.

The short-term magnitude of the degradation and degraded speech are estimated. The *minimum mean square error of log-spectral amplitude* (MMSE-LSA) estimator is applied to the magnitude spectra for obtaining the enhanced spectra from the degraded speech [16]. The spectral gain function for the MMSE-LSA estimator is expressed as follows:

$$H(n) = \frac{\xi_n}{1 + \xi_n} exp\left(\frac{1}{2}\int_{\upsilon_n}^{\infty} \frac{e^{-x}}{x} \mathrm{d}x\right) \tag{24}$$

where

$$\upsilon_n = \frac{\xi_n}{1 + \xi_n}\gamma_n \tag{25}$$

where $\xi_n$ and $\gamma_n$ are *a priori* SNR and *a posteriori* SNR, respectively.

The enhanced magnitude spectra and degraded phase spectra are then combined to produce an estimate of clean speech, and the overlap-add method is used for the re-synthesis in the time domain. The re-synthesized speech is the enhanced speech, and this enhancement may improve the TDSV system performance under degraded speech and challenging test conditions.

## 5 Experimental Setup

The proposed work is directed toward addressing issues related to practically deployable systems. It is important to consider proper databases for the experimental study. In order to evaluate the robustness of the proposed framework under degraded speech conditions, we conducted the TDSV system experiments on two databases, namely the RSR2015 database [34] and the IITG database [4,13,30,37,52]. The databases are described below.

### 5.1 Databases

The RSR2015 database is one of the largest publicly available and popular databases mainly outlined for TDSV with pre-defined speech utterances. The RSR2015 database is recorded so as to give the speech community an adequately expansive data set from the gender-balanced set of speakers. Each of the speakers recorded 9 sessions, each session is made of 30 pre-defined speech utterances. In this proposed framework, three speech utterances are selected for the experimental studies. For each speaker, out of the 9 speech utterances (sessions) for every speech utterance, 3 speech utterances are used for training and remaining 6 speech utterances are used for testing the system performance. The selected three speech utterances are $\mathbf{TD-1}$: *"Only lawyers love millionaires"*, $\mathbf{TD-2}$: *"I know I did not meet her early enough"* and $\mathbf{TD-3}$: *"The Birthday party has cupcake and ice-cream"*.

To evaluate the proposed framework under degraded speech condition, the clean speech utterances are corrupted with Babble noise to generate the noise mixed speech [56]. The energy level of the noise is scaled such that overall SNR of the degraded speech files is maintained at 20–0 dB in steps of 5 dB. The performance of the system is evaluated on both clean and synthetically degraded RSR2015 database.

The IITG database was collected from the undergraduate and post-graduate course students with a speech biometric-based attendance system as an application to address the issues in the practical deployment of TDSV system [4,13,30,37,52]. During the enrollment phase, three pre-defined speech utterances were recorded for each enrolled course student in an anechoic chamber. The three speech utterances used for the attendance system from the IITG database are $\mathbf{TD-1}$: *"Dont ask me to walk like that"*, $\mathbf{TD-2}$: *"Lovely pictures can only be drawn"* and $\mathbf{TD-3}$: *"Get into the hole of tunnels"*.

During the collection of the testing database, the course students could move freely within, and in and out of the open hall in the department office, in the corridor, in front of the classroom and also in the free environment. The data collection process was more practical because the database includes background noise, background speech, and other environmental conditions. The recorded data contains mismatches between the enrollment and the testing phases, in terms of the sensors, mobile handsets, style of speech and environment. Due to these conditions, the text-dependent SV task becomes more challenging in IITG database. Henceforth, there is no compelling reason to add artificial noise in the IITG database for the evaluation of the TDSV system.

### 5.2 Feature Extraction

During the training and testing process, the given speech signal is processed in frames of 20 ms duration at a 10 ms frame shift. Prior to extracting the features, EPD is performed to eliminate the background speech/background noise regions at the beginning and end of the given speech utterance. The speech utterance between the detected end points are considered and used for the feature extraction. For each 20 ms frame size Hamming windowed speech frame, MFCCs are computed using 22 logarithmically spaced Mel-filterbanks. The features extracted from every frame are stacked and then *mean subtraction and variance normalization* is performed, for every feature dimension. This extracted feature set is then used for the task of TDSV system, using the technique of *Dynamic Time Warping* (DTW) [4,19,49,52]. DTW is a popularly well known feature matching algorithm, which optimizes the distance between two feature sets while maintaining the temporal relation between them. Hence, given a training speech utterance corresponding to a speech sentence (TD-1 / TD-2 / TD-3), and a testing speech utterance for the same speech sentence, the corresponding feature sets obtained from the two utterances may be compared by using DTW matching algorithm, resulting in a DTW score. A lesser DTW score indicates higher similarity between the two utterances, and vice versa.

In this work, the seven different feature sets obtained from IMFs derived from HS of a given speech utterance are used independently, and in combination with the feature set corresponding to 39-dimensional MFCCs, for the TDSV task, using DTW. The dimensions of each of the 7 feature sets are varied, by varying the number of IMFs from $K = 10$ to 2, in steps of 2, to observe the importance of each IMF. This can be done in order to find out the range of IMFs which are useful for the TDSV task. From every speech frame, two different sets of MFCCs are obtained. For extracting the MFCCs from a speech frame, the speech frame is first passed through the pre-emphasis filter and then 22-logarithmically spaced filters (Mel filterbank) are applied on its DFT spectrum. Then, 39-dimensional MFCCs and an extended 51-dimensional MFCCs (Ext. MFCCs) feature vector are obtained from each speech frame. The standard 39-dimensional MFCCs feature vector comprises of the first 13 cepstral (excluding the 0th coefficient), the first 13 $\Delta$ cepstral, and the first 13 $\Delta\Delta$ cepstral coefficients. The Ext. MFCCs feature vector comprises of the first 17 cepstral (excluding the 0th coefficient), the first 17 $\Delta$ cepstral, and the first 17 $\Delta\Delta$ cepstral coefficients. Comparing the system performances of the Ext. MFCCs with that of the 39-dimensional MFCCs enables us to notice how useful the higher dimensions of the MFCCs are in the TDSV system task. Also, it empowers us to ascertain the combinations of the seven different feature sets obtained from HS, with the 39-dimensional MFCCs. Apart from the aforementioned feature sets, the *refined Sum Log-Squared Amplitude* feature vector, $cG_K$, is obtained from each speech frame using its raw IMFs of the given speech signal as described in [55]. However, considering this enormity of our database, we derive the $cG_K$ feature for every frame of the speech utterance. Further, we refine the feature vector using DCT [54].

### 5.3 Performance Metrics

The performance of proposed TDSV system is evaluated on the RSR2015 and the IITG databases, based on two standard metrics—*Equal Error Rate* (EER) and *minimum Detection Cost Function* (*m*DCF) [4,52,55]. Let $D_c = D_{c1}, D_{c2}, \ldots, D_{cS}$ represent the entire set of the DTW scores obtained by verifying the claims of the testing speech utterance against the training speech utterance. The set $D_c$ score is normalized so that $D_{c_i} \epsilon [0, 1] | i = 1, 2, .., S$. Let $\xi^D \epsilon [0, 1]$ is the decision threshold, above which the claim of the test speech utterance against the train speech utterance is considered valid, i.e., both test and train speech utterances are considered to belong to the same speaker. Let $D_G$, $D_F$ and $D_I$, $D_M$ denote the number of genuine/true claims, imposter/false claims that have been accepted and the number of imposter/false claims, genuine/true claims that have been rejected. The DTW scores of genuine and imposter together is considered as, $D = D_G + D_I + D_M + D_F$. We may now define the evaluation parameters EER and *m*DCF as follows:

**EER (%)** This assigns cost to the event of a wrong classification, and also takes into consideration the probability that a score is genuine (claim is genuine) or not. For any given threshold value, $\xi^D$, the *Probability of Miss* is given as $P_M = \frac{D_M}{D_G + D_M} \times 100\%$. The *Probability of False Alarm* is given as $P_F = \frac{D_F}{D_I + D_F} \times 100\%$. Then, at a particular threshold value, $\xi^D = \xi_0^D$, Then $P_M = P_F$. This error is known as the EER. Then, EER $= P_M \times 100\% = P_F \times 100\%$.

***m*DCF** In this work, for calculating the *m*DCF, two parameters, $C_M$ and $C_F$, assign minimum costs to the event of a *miss* (a genuine claim rejected) and that of a *false alarm* (an imposter claim is accepted), respectively. There is also one more parameter, an *a priori* probability, $P_T$, which is assigned to the cost function. It is assumed that out of all the test claims, only a fraction $P_T$ number of speakers are genuine claims. Then, for any given threshold, $\xi^D$, the cost parameter of the *m*DCF is given by,

$$C_{\xi^D} = C_M \times P_M \times P_T + C_F \times P_F \times (1 - P_T) \tag{26}$$

Then *m*DCF is given by,

$$mDCF = \min_{\xi^D \in [0,1]} C_{\xi^D} \tag{27}$$

In this work, $C_M = 10$, $C_F = 1$, and $P_T = 0.01$ are considered, for both the databases.

## 6 Results and Analysis

In this section, we present the experimental results and analysis, in terms of EER and *m*DCF, for the TDSV system implemented separately on the RSR2015 and the IITG databases.

### 6.1 Performances Obtained When Robust EPD Followed by Extraction of MEMD Features

The performances of the seven HS features derived from MEMD, the conventional 39-dimensional MFCCs features, the Ext. 51-dimensional MFCCs, the $cG_K$ features and their different combinations are evaluated. The effectiveness of the robust EPD algorithm with different features is shown by comparing the performance with the energy-based EPD.

#### 6.1.1 Performances on the RSR2015 Database Under Clean Conditions

The performance metrics of the proposed methods for the TDSV system is evaluated on the clean RSR2015 database. Table 1 presents the performance metrics for each of the seven extracted experimental features, derived using all the 10 IMFs. For comparison, the performance metrics for MFCCs, Ext. MFCCs and cGk features are also shown. As it is observed from the table, for all the three speech utterances, the performance for MFCCs is better than the rest of the features. For the TD-1 speech utterance, the 39-dimensional MFCCs provide an EER of 6.37%. The 51-dimensional Ext. MFCCs shows improvement as compared to the standalone MFCCs, with an EER of 4.93%. This improved performance of Ext. MFCCs over 39-dimensional MFCCs in TDSV task was observed in one of our previous energy-based EPD work reported in [52]. In Table 1, the numbers shown in the brackets are the results obtained by using energy-based EPD and the numbers shown without brackets are the results obtained by using the speech-specific knowledge-based robust end points detection method. It can be seen that the performance for the robust end points detection method is significantly better than the energy-based method. For all the three utterances, standalone MFCC features and Ext. MFCCs are far better than cGk and seven experimental HS features. Having observed the performances of the standalone features, we need to proceed toward combining them with the MFCCs and evaluating the effect of the combinations on the TDSV system.

Table 2 presents the performance metrics of the feature combinations for the TD-1 sentence. The seven features extracted from HS and $cG_K$ feature are augmented with MFCCs, and the number of IMFs from which the features are extracted are reduced by changing K values from 10 to 2, in steps of 2. It can be observed from the table that the best performance of the combination of the features are obtained for $K \leqslant 4$. The combination of different HS features and $cG_K$ features are compared to the standalone MFCCs, an absolute improvement in EER of around 0.5–1.0% are obtained at $K = 4$. The additional 51-dimensions of the Ext. MFCCs also significantly improves the EER as compared to the standard MFCCs. Similar trends are noticed for $m$DCF values in each case. Figure 10 shows the *Detection Error Tradeoff* (DET) curves for different feature combinations, considering $K = 4$. It can be observed from the figure that the Ext. MFCCs outperforms the MFCCs and MFCCs+$cG_K$ combination [40]. The MFCCs+HS combinations outperform the standard MFCCs for most of the cases. However, they are not always competitive with the 51-dimensional Ext. MFCCs. The results obtained using the speech-specific knowledge-based robust end points detection method are compared with the results obtained using energy-based VAD method, shown in brackets. It is observed that the results are improved after using the robust end points detection method. Figure 11 presents the performance metrics of different feature combinations for the sentences TD-2 and TD-3, respectively. In the case of TD-2 and TD-3 sentences, the Ext. MFCCs marginally improves the system performance after applying the robust EPD ([4]) and also very less improvement is observed in HS feature combinations as compared to our previous work reported in [52]. However, the MFCCs+$cG_K$ and MFCCs+HS feature, again, obtain best performances at K=4. This study clearly suggests that HS features and $cG_K$ can naturally complement the MFCCs performances, under clean conditions. The additional 51-dimensional Ext. MFCCs may not always improve the system performance. Moreover, the use of robust end points detection method further improves the performance of the system. It is clearly observed that only the first few IMFs are useful in characterizing the speakers which is similar to the observations reported in [52,55].

**Table 1** The performance of the TDSV system evaluated using Robust EPD and energy-based EPD results are shown in brackets, using 39-dimensional MFCCs, 51-dimensional Ext. MFCCs, $cG_K$ feature and the proposed seven HS features derived from MEMD, using all the IMFs ($K = 10$), on the RSR2015 database

| Technique | | Mel filterbank | | MEMD filterbank | MEMD-Hilbert spectrum | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Feature | | MFCCs 39 dim | Ext. MFCCs 51 dim | $cG_{10}$ | $F_{10}$ | $E_{10}$ | $\sigma F_9$ | $\sigma E_9$ | $\Delta F_{10}$ | $\Delta E_{10}$ | $\Upsilon_{10}$ |
| Sentence | Metric | | | | | | | | | | |
| TD-1 | EER (%) | 6.37 (7.59) | 4.93 (6.22) | 13.22 (27.71) | 6.83 (15.48) | 11.76 (12.57) | 8.81 (11.65) | 12.14 (13.14) | 13.39 (22.61) | 21.67 (19.33) | 16.88 (16.22) |
| | mDCF | 0.0375 (0.0388) | 0.035 (0.0378) | 0.0382 (0.0555) | 0.0341 (0.0454) | 0.0516 (0.0428) | 0.0294 (0.0421) | 0.0317 (0.0433) | 0.0432 (0.0513) | 0.0469 (0.0484) | 0.0494 (0.046) |
| TD-2 | EER (%) | 5.9 (6.29) | 4.9 (6.4) | 11.23 (23.16) | 6.42 (13.33) | 10.81 (16.75) | 8.27 (9.02) | 8.64 (8.85) | 8.57 (12.63) | 12.87 (13.43) | 7.71 (11.84) |
| | mDCF | 0.0442 (0.0421) | 0.0367 (0.0423) | 0.0437 (0.0546) | 0.0432 (0.0474) | 0.0546 (0.0497) | 0.0452 (0.044) | 0.0453 (0.0439) | 0.0359 (0.0469) | 0.0457 (0.0472) | 0.0419 (0.0463) |
| TD-3 | EER (%) | 5.87 (6.6) | 4.94 (6.25) | 11.89 (17.5) | 6.67 (11.4) | 6.45 (8.18) | 7.33 (15.17) | 8.56 (15.93) | 7.03 (7.83) | 8.39 (9.9) | 7.26 (10.19) |
| | mDCF | 0.039 (0.0422) | 0.0382 (0.0418) | 0.0414 (0.058) | 0.039 (0.0492) | 0.0336 (0.0444) | 0.0411 (0.0544) | 0.0393 (0.0555) | 0.0301 (0.0441) | 0.0367 (0.0469) | 0.0378 (0.0475) |

The numbers shown in brackets are obtained by using energy-based VAD and without brackets are obtained by using speech-specific knowledge-based robust EPD method

**Table 2** The performance of the TDSV system using robust EPD followed by MEMD feature extraction

| Technique | | Mel filterbank | | MEMD filterbank | MEMD-Hilbert spectrum | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Feature | | MFCCs 39 dim | Ext. MFCCs 51 dim | MFCCs + $cG_K$ | MFCCs + $F_K$ | MFCCs + $E_K$ | MFCCs $+\sigma F_{K-1}$ | MFCCs + $\sigma E_{K-1}$ | MFCCs + $\Delta F_K$ | MFCCs + $\Delta E_K$ | MFCCs + $\Upsilon_K$ |
| $K$ | Metric | | | | | | | | | | |
| 10 | EER (%) | 6.37 (7.59) | **4.93** (6.22) | 8.72 (8.65) | 6.9 (6.48) | 6.82 (8.17) | 10.46 (9.36) | 7.33 (8.82) | 8.08 (8.41) | 7.88 (8.4) | 7.72 (7.11) |
| | mDCF | 0.0375 (0.0388) | **0.035** (0.0378) | 0.0413 (0.0398) | 0.037 (0.038) | 0.0375 (0.0392) | 0.042 (0.0402) | 0.0385 (0.0397) | 0.0383 (0.0396) | 0.037 (0.0394) | 0.0372 (0.0385) |
| 8 | EER (%) | – | – | 7.28 (7.34) | 6.99 (6.54) | 6.86 (7.07) | 8.85 (9.7) | 7.61 (8.54) | 6.69 (7.58) | 7.98 (7.39) | 7.9 (7.41) |
| | mDCF | – | – | 0.0346 (0.0387) | 0.0353 (0.038) | 0.0353 (0.0383) | 0.0391 (0.0404) | 0.039 (0.0395) | 0.0357 (0.0389) | 0.0378 (0.0385) | 0.0367 (0.0388) |
| 6 | EER (%) | – | – | 5.96 (6.96) | 6.44 (6.54) | 6.68 (7.56) | 7.96 (8.27) | 6.69 (7.02) | 6.77 (6.95) | 7.76 (7.04) | 6.7 (7.21) |
| | mDCF | – | – | 0.0295 (0.0384) | 0.0344 (0.038) | 0.0359 (0.0387) | 0.0383 (0.0393) | 0.0367 (0.0382) | 0.0359 (0.0384) | 0.039 (0.0382) | 0.0336 (0.0386) |
| 4 | EER (%) | – | – | **5.15** (**5.03**) | **6.18** (6.57) | **6.23** (6.62) | 6.67 (6.91) | **6.28** (6.39) | 6.88 (7.28) | 6.45 (6.71) | **6.27** (7.2) |
| | mDCF | – | – | **0.0249** (0.0368) | **0.0351** (0.0381) | **0.0361** (0.0379) | **0.0338** (0.0381) | **0.0353** (0.0377) | **0.0389** (0.0386) | **0.0349** (0.038) | **0.0349** (0.0386) |
| 2 | EER (%) | – | – | 5.87 (6.66) | 6.67 (6.96) | 7.3 (6.93) | 7.41 (6.85) | 6.75 (7.7) | 6.78 (6.75) | 6.67 (7.42) | 7.09 (7.27) |
| | mDCF | – | – | 0.0269 (0.0381) | 0.0348 (0.0384) | 0.0358 (0.0382) | 0.0357 (0.0381) | 0.0367 (0.0388) | 0.0367 (0.0382) | 0.0353 (0.0386) | 0.0346 (0.0386) |

The results are shown for 39-dimensional MFCCs, 51-dimensional Ext. MFCCs, and the combinations of the 39-dimensional MFCCs with the $cG_K$ feature and each of the seven HS experimental feature obtained from MEMD. The performance metrics are evaluated for the TD-1 speech utterance of the RSR2015 database. The dimensions of the MEMD features are varied by changing $K$ from 10 to 2, in steps of 2. The numbers shown in brackets are obtained by using energy-based VAD and without brackets are obtained by using speech-specific knowledge-based robust EPD method

Bold values represent the improved results as compared to the existing methods

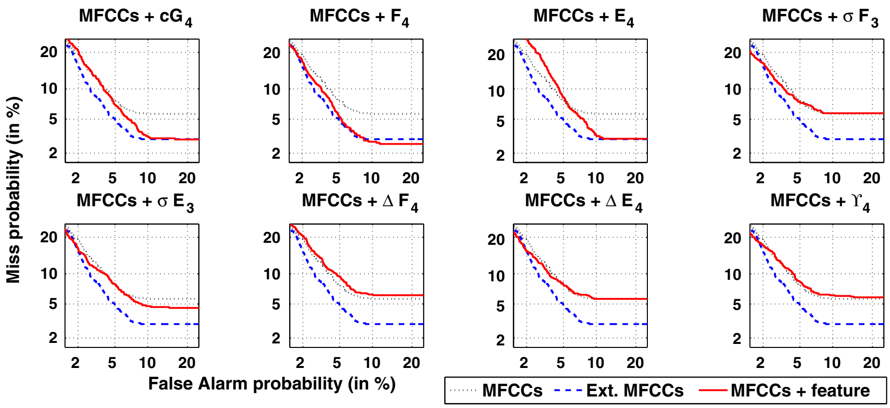

**Fig. 10** The DET curves of the 39-dimensional MFCCs, 51-dimensional Ext. MFCCs, and the combinations of the MFCCs with the $cG_K$ feature and each of the seven experimental features. The DET curves are plotted for the TD-1 sentence of the RSR2015 database. The dimensions of the $cG_K$ and seven HS features are fixed at $K = 4$

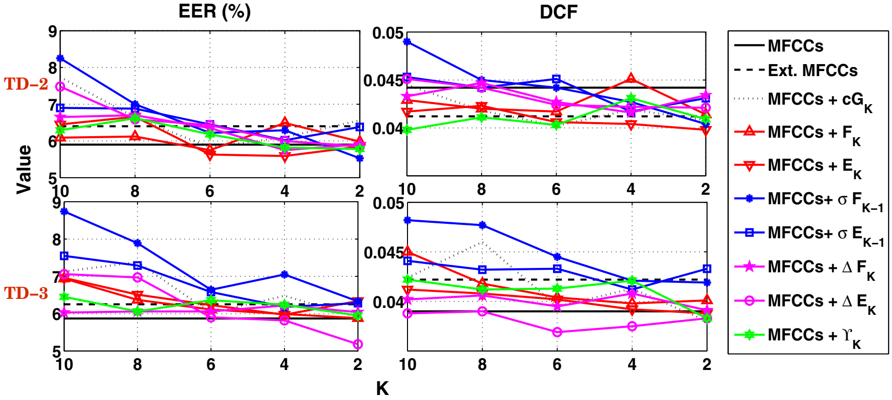

**Fig. 11** The performance of the TDSV system using robust EPD followed by MEMD feature extraction. The results are shown for the 39-dimensional MFCCs, 51-dimensional Ext. MFCCs, and the combinations of the 39-dimensional MFCCs with the $cG_K$ feature and each of the seven experimental features. The performance metrics are evaluated for the TD-2 and TD-3 sentences of the RSR2015 database. The dimensions of the $cG_K$ and seven HS features are varied by changing $K$ from 10 to 2

### 6.1.2 Performances on the RSR2015 Database Under Degraded Conditions

Having demonstrated the performances of the features and their combinations under clean conditions, we can now evaluate how robust the features are to external interference. For this reason, the testing speech utterances are corrupted with Babble noise [56] prior to extracting feature sets from the RSR2015 database. Table 3 presents the performances of the HS features and $cG_K$ feature combination with MFCCs for the TD-1 sentence. As observed from the clean speech, the best performances are observed at $K = 4$ and therefore this value is used for all the experiments conducted under degraded conditions. This can be noticed from the table that the performance

**Table 3** The performance of the TDSV system using robust EPD followed by MEMD feature extraction

| Technique | | Mel filterbank | | MEMD filterbank | MEMD-Hilbert Spectrum | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Feature | | MFCCs 39 dim | Ext. MFCCs 51 dim | MFCCs + $cG_4$ | MFCCs + $F_4$ | MFCCs + $E_4$ | MFCCs + $\sigma F_3$ | MFCCs + $\sigma E_3$ | MFCCs + $\Delta F_4$ | MFCCs + $\Delta E_4$ | MFCCs + $\Upsilon_4$ |
| SNR | Metric | | | | | | | | | | |
| 20 dB | EER (%) | 6.81 (8.1) | 6.55 (11.91) | 6.46 (6.55) | 6.33 (6.71) | 6.44 (6.67) | 6.5 (6.98) | 6.92 (6.48) | 6.3 (7.34) | 6.76 (7.4) | **6.25** (7.84) |
| | mDCF | 0.0442 (0.0394) | **0.0332** (0.0425) | 0.0415 (0.038) | 0.0444 (0.0382) | 0.0447 (0.0379) | 0.0452 (0.0382) | 0.0474 (0.0378) | 0.0453 (0.0387) | 0.0453 (0.0385) | 0.0429 (0.0391) |
| 15 dB | EER (%) | 7.23 (9.26) | 6.9 (12.47) | 6.93 (11.13) | 7.12 (7.77) | **5.94** (6.95) | 7.49 (7.64) | 6.08 (7) | 7.34 (7.37) | 7.37 (7.56) | 6.63 (8.06) |
| | mDCF | 0.0464 (0.0404) | 0.0453 (0.0429) | **0.0408** (0.0418) | 0.0483 (0.039) | 0.0439 (0.0382) | 0.0472 (0.0387) | 0.0462 (0.0382) | 0.0465 (0.0387) | 0.0463 (0.0387) | 0.0434 (0.0393) |
| 10 dB | EER (%) | 7.5 (11.23) | 7.38 (12.9) | 11.23 (13.22) | 7.26 (13.23) | **6.89** (10.31) | 8.1 (10.27) | 7.35 (10.29) | 8.32 (9.89) | 7.13 (10.34) | 7.35 (10.48) |
| | mDCF | 0.0488 (0.042) | 0.046 (0.0433) | **0.0418** (0.0436) | 0.0476 (0.0436) | 0.0442 (0.0409) | 0.0488 (0.0409) | 0.0473 (0.0409) | 0.0494 (0.0408) | 0.0485 (0.041) | 0.0485 (0.0413) |

**Table 3** continued

| Technique | | Mel filterbank | | MEMD filterbank | MEMD-Hilbert Spectrum | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Feature | | MFCCs 39 dim | Ext. MFCCs 51 dim | MFCCs + $cG_4$ | MFCCs + $F_4$ | MFCCs + $E_4$ | MFCCs + $\sigma F_3$ | MFCCs + $\sigma E_3$ | MFCCs + $\Delta F_4$ | MFCCs + $\Delta E_4$ | MFCCs + $\Upsilon_4$ |
| SNR | Metric | | | | | | | | | | |
| 5 dB | EER (%) | 13.42<br>(12.51) | 10.73<br>(20.93) | **10.21**<br>(14.18) | 11.05<br>(14.35) | 11.25<br>(12.6) | 10.32<br>(12.37) | 12.52<br>(12.59) | 10.32<br>(13.31) | 14.09<br>(13.51) | 13.56<br>(13.14) |
| | mDCF | 0.048<br>(0.0431) | 0.0453<br>(0.0499) | 0.0444<br>(0.0443) | **0.0441**<br>(0.0445) | 0.0487<br>(0.0428) | 0.0476<br>(0.0427) | 0.0447<br>(0.0428) | 0.0479<br>(0.0436) | 0.0459<br>(0.0436) | 0.0458<br>(0.0435) |
| 0 dB | EER (%) | 22.93<br>(20.29) | **16.9**<br>(23.29) | 18.43<br>(21.04) | 18.76<br>(21.6) | 20.04<br>(20.42) | 21.56<br>(20.76) | 22.22<br>(20.1) | 20.86<br>(22.17) | 21.04<br>(20.68) | 18.71<br>(22.03) |
| | mDCF | 0.0476<br>(0.0493) | 0.0485<br>(0.0522) | 0.051<br>(0.05) | 0.0527<br>(0.0505) | **0.0453**<br>(0.0493) | 0.0486<br>(0.0496) | 0.0489<br>(0.049) | 0.0556<br>(0.051) | 0.0464<br>(0.0495) | 0.0488<br>(0.0508) |

The results are shown for the 39-dimensional MFCCs, 51-dimensional Ext. MFCCs, and the combinations of the 39-dimensional MFCCs with the $cG_K$ feature and each of the seven HS experimental feature obtained from MEMD. The performance metrics are evaluated for the TD-1 speech utterance of the RSR2015 database. The testing utterances are corrupted with Babble noise, with SNR varying from 20 to 0 dB, in steps of 5 dB. The dimensions of the MEMD features are kept constant, with $K = 4$. The numbers shown in brackets are obtained by using energy-based VAD and without brackets are obtained by using speech-specific knowledge-based robust EPD method

Bold values represent the improved results as compared to the existing methods

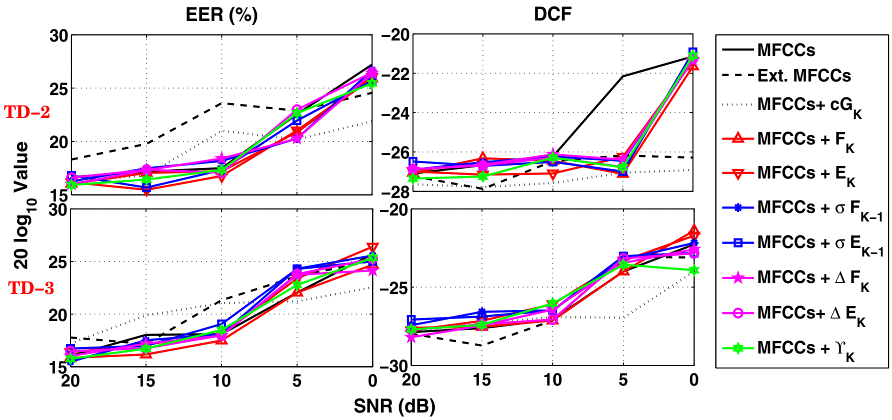

**Fig. 12** The performance of the TDSV system using robust EPD followed by MEMD feature extraction. The results are shown for the 39-dimensional MFCCs, 51-dimensional Ext. MFCCs, and the combinations of the 39-dimensional MFCCs with the $cG_K$ feature and each of the seven experimental features. The performance metrics are evaluated for the TD-2 and TD-3 sentences of the RSR2015 database. The testing utterances are corrupted by Babble noise, with SNR varying from 20 to 0 dB. The dimensions of the $cG_K$ seven features are kept constant, with $K = 4$

of the TDSV system degrades significantly with the increase in SNR levels. The 51-dimensional Ext. MFCCs feature show slightly better performance as compared to the MFCCs. Most of the feature combinations (MFCCs + HS) shows improved performance compared to the MFCCs alone. Numbers within the brackets are the results obtained using energy-based end points detection. Clearly, use of robust end points detection in the proposed method is giving improved results compared to the energy-based end points detection, for different levels of degradation.

Figure 12 presents the performance metrics for the feature combinations for the TD-2 and TD-3 sentences. The performances under different SNR levels have been plotted for better representation of the improvement in the system performance. This can be noticed in the figure that similar to the observations made in Table 3, in most of the cases, the combined features are showing improved performance compared to the MFCCs alone.

### 6.1.3 Performances on the IITG Database

So far we have evaluated the speech corrupted by artificially inserted Babble noise. Now, in this section, we present the performance of the system using speech utterances which are naturally affected by background speech, background noise, telephone channel, interferences from other speakers and other environmental conditions. For this purpose, we considered the IITG database for evaluating robustness of the proposed system under practical conditions. This system was used for practical deployment at the institute level for marking attendance as an application.

Table 4 presents the performance metrics of the MFCCs+HS features and the MFCCs+$cG_K$ feature combination for each of the three speech utterances. Taking characteristics of the speaker-specific cues from the examinations made for the

**Table 4** The performance of the TDSV system using robust EPD followed by MEMD feature extraction

| Technique | | Mel filterbank | | MEMD filterbank | MEMD-Hilbert spectrum | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Feature | | MFCCs 39 dim | Ext. MFCCs 51 dim | MFCCs + $cG_4$ | MFCCs + $F_4$ | MFCCs + $E_4$ | MFCCs + $\sigma F_3$ | MFCCs + $\sigma E_3$ | MFCCs + $\Delta F_4$ | MFCCs + $\Delta E_4$ | MFCCs + $\Upsilon_4$ |
| Sentence | Metric | | | | | | | | | | |
| TD-1 | EER (%) | 10.81 (13.19) | 9.86 (12.84) | 10.48 (10.91) | **9.42** (11.46) | 10.96 (10.62) | 11.77 (10.41) | 12.76 (10.25) | 11.23 (11.83) | 12.68 (10.68) | 11.3 (11.48) |
| | mDCF | 0.0416 (0.0422) | 0.0459 (0.0412) | 0.0411 (0.035) | 0.0423 (0.0368) | 0.0423 (0.0339) | 0.0429 (0.0332) | 0.0443 (0.0327) | 0.043 (0.0379) | **0.041** (0.0341) | 0.0419 (0.0368) |
| TD-2 | EER (%) | 11.21 (11.37) | 10.65 (11.65) | 12.56 (11.56) | 10.89 (11.3) | 11.85 (10.33) | 14.59 (10.71) | 13.45 (9.86) | **9.1** (11.1) | 10.86 (10.86) | 10.38 (10.38) |
| | mDCF | 0.0406 (0.0404) | 0.0413 (0.0415) | 0.0404 (0.0412) | 0.0433 (0.0403) | 0.0441 (0.0366) | 0.0449 (0.038) | 0.0435 (0.0349) | **0.04** (0.0395) | 0.0417 (0.0385) | 0.0415 (0.037) |
| TD-3 | EER (%) | 16.87 (18.42) | 15.93 (17.93) | **14.69** (17.96) | 15.88 (19.26) | 16.37 (17.86) | 16.82 (17.48) | 17.45 (17.35) | 15.7 (17.3) | 17.25 (17.9) | 15.92 (17.93) |
| | mDCF | 0.0491 (0.0526) | 0.0526 (0.0513) | 0.0497 (0.0514) | **0.045** (0.0551) | 0.0518 (0.0509) | 0.0544 (0.0498) | 0.0538 (0.0494) | 0.0485 (0.0495) | 0.0511 (0.051) | 0.0512 (0.0513) |

The results are shown for the 39-dimensional MFCCs, 51-dimensional Ext. MFCCs, and the combinations of the 39-dimensional MFCCs with the $cG_K$ feature and each of the seven HS experimental feature obtained from MEMD. The performance metrics are evaluated for the TD-1, TD-2 and TD-3 speech utterances of the IITG database. The testing utterances are collected in the practical scenarios. The dimensions of the MEMD features are kept constant, with $K = 4$. The numbers shown in brackets are obtained by using energy-based VAD and the results shown without brackets are obtained by using speech-specific knowledge-based robust EPD method

Bold values represent the improved results as compared to the existing methods

**Table 5** The performance of the TDSV system using enhanced 39-dimensional MFCC features

| Sentence<br>Dimension | Metric | Clean<br>39 | 20 dB<br>39 | 15 dB<br>39 | 10 dB<br>39 | 5 dB<br>39 | 0 dB<br>39 |
|---|---|---|---|---|---|---|---|
| TD-1 | EER (%) | 4.83 (6.18) | 5.12 (6.25) | 5.21 (5.94) | 6.24 (6.89) | 11.36 (10.32) | 17.36 (18.71) |
| | mDCF | 0.0386 (0.0351) | 0.0378 (0.0429) | 0.0402 (0.0439) | 0.0421 (0.0442) | 0.0653 (0.0479) | 0.0756 (0.0488) |
| TD-2 | EER (%) | 4.77 (5.53) | 5.16 (6.3) | 6.11 (5.94) | 7.73 (7.35) | 13.15 (14.09) | 19.05 (21.71) |
| | mDCF | 0.0367 (0.0404) | 0.042 (0.0453) | 0.0432 (0.0439) | 0.0498 (0.0473) | 0.0674 (0.0684) | 0.0772 (0.088) |
| TD-3 | EER (%) | 4.85 (5.18) | 5.49 (5.94) | 5.84 (6.43) | 7.52 (7.49) | 10.26 (12.65) | 15.79 (18.31) |
| | mDCF | 0.0384 (0.0383) | 0.0396 (0.0425) | 0.0429 (0.042) | 0.0525 (0.044) | 0.0634 (0.0631) | 0.073 (0.0721) |

The performance metrics are evaluated for the TD-1, TD-2 and TD-3 speech utterances of the RSR2015 database in terms of the EER and mDCF. The experimental analysis done on the clean and testing utterances are corrupted with varying strength from 0 to 20 dB, in steps of 5 dB. The numbers shown in brackets are obtained by MFCCs+HS best performing results and without brackets are obtained by using speech enhancement followed by robust EPD method

RSR2015 database, the dimensions of the $cG_K$ and different HS features remain fixed by setting $K = 4$. From the table, one can notice that the 51-dimensional Ext. MFCCs and MFCCs+$cG_K$ feature combination provide the improved performance of the system with respect to the MFCCs for the TD-1 and TD-3 speech utterances, but MFCCs+$cG_K$ feature combination for the TD-2 speech utterance case show slightly degraded performance as compared to the MFCCs and Ext. MFCCs. The MFCCs + HS feature combinations inconsistently provide the improved performance, for all the three speech utterances. The best performances are spread across different feature combinations, and hence, fusion of various low-dimensional features could be useful. It is also observed that the robust end points detection is helping in most of the cases by improving the EER compared to the energy-based end points detection. In the table, results obtained by using energy-based end points detection are shown inside brackets.

### 6.2 Performances Obtained by Using Speech Enhancement Followed by Robust EPD Method

In this experiment, speech enhancement is performed before the end points detection and feature extraction stage. The performance on the RSR2015 database is presented not only under clean speech conditions but also under Babble noise of varying strength. Table 5 presents the performances of the RSR2015 database under clean and different noise levels. Numbers inside the bracket show the best results from Table 3. These results were obtained without performing speech enhancement and are shown in Table 5 for comparison.

**Table 6** The performance metrics are evaluated for the TD-1, TD-2 and TD-3 speech utterances of the IITG database

| Sentence<br>Metric | TD-1<br>39 dim | TD-2<br>39 dim | TD-3<br>39 dim |
|---|---|---|---|
| EER (%) | 6.82 (9.42) | 7.38 (9.1) | 8.54 (14.69) |
| mDCF | 0.0331 (0.0423) | 0.0343 (0.04) | 0.0398 (0.0497) |

The testing utterances are collected in the practical scenarios. The results are evaluated in terms of EER and mDCF. The numbers shown in brackets are obtained by MFCCs+HS best performing results and without brackets are obtained by using speech enhancement followed by robust EPD method

From the table, one can observe that introduction of the temporal and spectral enhancement on the speech utterances provides slightly better and comparable performances, under various conditions. Table 5 show the experimental results that consistently providing improved results as compared to the MFCCs augmented with HS features and MFCC standalone features. Table 6 shows the same experiments for the IITG database. The IITG database is already affected by environmental noise, telephone channel and interference from other speakers, and hence noise is not added artificially. Similar trend in performance is observed for the IITG database as well. Speech enhancement improves the performance of the TDSV system significantly.

### 6.3 Performances Obtained by Using Speech Enhancement Followed by Robust EPD and Extraction of MEMD Features

The final combined system can be obtained by using all three robust methods, where robust EPD is performed on the enhanced speech and the robust MEMD features are extracted from the detected speech regions. Performances are evaluated on the RSR2015 and the IITG databases.

Table 7 shows the results of the final combined system obtained by using three robust methods applied on the speech utterances in a sequential manner. The obtained results shows improvement in performances compared to the performances with the signal- and feature-level compensation techniques shown in Table 3. The improvements are consistently observed for the low SNR cases namely, 0 dB and 5 dB SNR. For 10–20 dB SNR, improvements are observed in some of the cases. Similar trend in performances are observed for TD-2 and TD-3 speech sentences as shown in Fig. 13.

Table 8 shows the same set of features and same experiments for the IITG database. Similar trend in the system performance is observed for the IITG database as well, which can be observed by comparing the results in Table 8 with the corresponding results shown in Table 4. The combination of three robust methods improves the performance of the TDSV system for most of the cases.

## 7 Summary and Conclusion

This work focuses on using robust techniques for TDSV system under degraded speech and challenging test conditions. The process of TDSV system includes several stages

**Table 7** The performance of the TDSV system using speech enhancement followed by robust EPD and extraction of MEMD features

| Technique | | Mel filterbank | | MEMD filterbank | MEMD-Hilbert spectrum | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Feature | | MFCCs 39 dim | Ext. MFCCs 51 dim | MFCCs + $cG_4$ | MFCCs + $F_4$ | MFCCs + $E_4$ | MFCCs + $\sigma F_3$ | MFCCs + $\sigma E_3$ | MFCCs + $\Delta F_4$ | MFCCs + $\Delta E_4$ | MFCCs+ $\Upsilon_4$ |
| SNR | Metric | | | | | | | | | | |
| 20 dB | EER (%) | 7.3 | 5.37 | **5.08** | **5.08** | 5.12 | 5.37 | **5.08** | **5.08** | 5.51 | 5.83 |
| | mDCF | 0.0353 | 0.0341 | 0.0341 | **0.033** | 0.0338 | 0.0341 | **0.033** | 0.0341 | 0.038 | 0.0365 |
| 15 dB | EER (%) | 10.37 | 7.59 | 11.55 | 11.36 | 10.75 | 10.85 | 8.84 | 11.55 | 8.45 | **7.01** |
| | mDCF | 0.0571 | 0.0337 | 0.0428 | 0.0428 | 0.066 | 0.0428 | **0.0381** | 0.0428 | 0.044 | 0.0441 |
| 10 dB | EER (%) | **10.55** | 12.6 | 13.36 | 11.55 | 12.43 | 11.55 | 11.48 | 13.36 | 10.75 | 11.22 |
| | mDCF | 0.0704 | 0.0648 | 0.054 | 0.0535 | 0.0522 | 0.0649 | 0.0635 | 0.054 | 0.066 | **0.0482** |
| 5 dB | EER (%) | 15.68 | 17.68 | 17.68 | 16.57 | 16.18 | 16.18 | 17.89 | 17.68 | 16.53 | **15.49** |
| | mDCF | 0.0725 | 0.0669 | 0.0648 | 0.0647 | 0.0707 | 0.0707 | 0.0725 | 0.0648 | 0.0617 | **0.0602** |
| 0 dB | EER (%) | 17.24 | 17.73 | 17.73 | **17.17** | 21.64 | 21.64 | 17.94 | 17.73 | 17.68 | 20.14 |
| | mDCF | 0.0864 | 0.0748 | **0.0669** | 0.0758 | 0.0781 | 0.0781 | 0.084 | **0.0669** | 0.082 | 0.0777 |

Results are shown for 39-dimensional MFCCs, 51-dimensional Ext. MFCCs, and the combinations of the 39-dimensional MFCCs with the $cG_K$ feature and each of the seven HS experimental feature obtained from MEMD. The performance metrics are evaluated for the TD-1 speech utterance of the RSR2015 database. The testing utterances are corrupted with Babble noise, with SNR varying from 20 to 0 dB, in steps of 5 dB. The dimensions of the MEMD features are kept constant, with $K = 4$

Bold values represent the improved results as compared to the existing methods

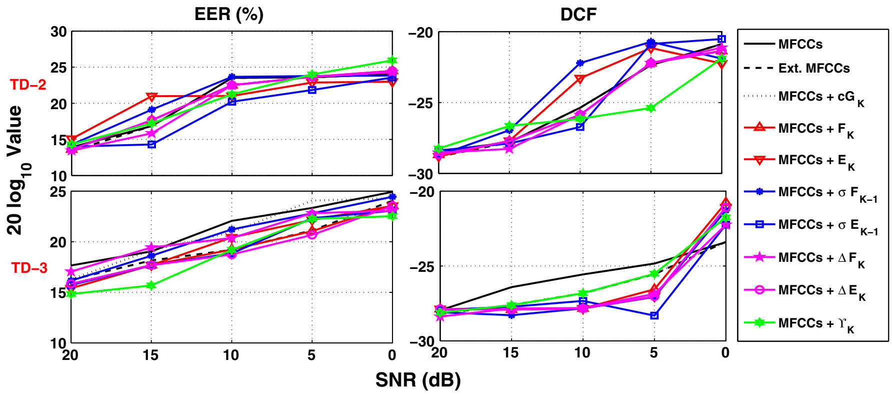

**Fig. 13** The performance of the TDSV system using speech enhancement followed by robust EPD and extraction of MEMD features. Results are shown for 39-dimensional MFCCs with the c$G_K$ feature and each of the seven experimental features. The performance metrics are evaluated for the TD-2 and TD-3 sentences of the RSR2015 database. The testing utterances are corrupted by Babble noise, with SNR varying from 20 to 0 dB. The dimensions of the c$G_K$ seven features are kept constant, with $K = 4$

such as, pre-processing, feature extraction, modeling and decision. In this work, several robust methods are explored in different stages of the speaker verification system. In the pre-processing stage, a robust end points detection using speech-specific knowledge is used instead of energy-based VAD. Similarly, in the feature extraction stage, robust features extracted from HS of IMFs obtained from MEMD are used in addition to the conventional MFCC features. We observed that when we perform the robust end points detection and use the robust HS features as additional features in the TDSV system, there is improvement in performance as compared to the standalone MFCCs extracted from the speech regions detected by energy-based VAD. The seven features in combination with the MFCCs, i.e., MFCCs + HS features show improved performance compared to MFCCs alone in different levels of degradation. This improvement is observed for HS features obtained from low dimensions of IMFs as expected, and the best performances are spread across different feature combinations with MFCCs.

Moreover, we also explored a combined temporal and spectral technique for speech enhancement. The enhanced speech utterances are passed through robust EPD to obtain the refined begin and end points. The results obtained after performing speech enhancement show slightly better performance compared to the best performance obtained by using MFCCs augmented with HS features and standalone MFCCs, without performing speech enhancement. Finally, we used all three methods in a sequential manner, where robust EPD is performed on the enhanced speech and then MEMD features are extracted from the regions between the detected end points. The combined method significantly improves the system performance for the test utterances with 0 dB, 5 dB and 10 dB SNRs. On the RSR2015 database, the proposed method provides the relative improvement of EER by 25%, 23% and 6%, for 0 dB, 5 dB and 10 dB, SNR cases, respectively. On the IITG database provides the relative improvement of EER from 30 to 45%. Performance is improved in terms of mDCF as well. On the RSR2015 database, the combined method shows relative improvement in the mDCF by 8%,

**Table 8** The performance of the TDSV system using speech enhancement followed by robust EPD and extraction of MEMD features

| Technique | | Mel filterbank | | MEMD filterbank | MEMD-Hilbert spectrum | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Feature | | MFCCs 39 dim | Ext. MFCCs 51 dim | MFCCs + $cG_4$ | MFCCs + $F_4$ | MFCCs + $E_4$ | MFCCs + $\sigma F_3$ | MFCCs + $\sigma E_3$ | MFCCs + $\Delta F_4$ | MFCCs + $\Delta E_4$ | MFCCs + $\Upsilon_4$ |
| Sentence | Metric | | | | | | | | | | |
| TD-1 | EER (%) | 9.5 | 8.69 | 6.82 | 7.07 | 6.73 | 6.28 | 6.46 | **6.12** | 6.18 | 7.07 |
| | mDCF | 0.0367 | 0.0365 | 0.0341 | 0.0341 | 0.0339 | 0.0339 | 0.0333 | **0.0309** | 0.0331 | 0.0341 |
| TD-2 | EER (%) | 6.45 | 6.69 | 6.26 | 6.43 | 6.23 | 6.2 | **6.01** | 6.91 | 6.26 | 6.43 |
| | mDCF | 0.0343 | 0.0338 | 0.0336 | 0.0341 | 0.0338 | 0.0331 | 0.0334 | 0.0354 | 0.0346 | **0.0328** |
| TD-3 | EER (%) | 9.84 | 9.74 | 8.35 | 9.69 | 9.38 | 8.17 | **7.63** | 8.2 | 8.35 | 9.67 |
| | mDCF | 0.0395 | 0.0405 | **0.0379** | 0.0401 | 0.0386 | 0.038 | 0.0387 | 0.0399 | **0.0379** | 0.0397 |

Results are shown for 39-dimensional MFCCs, 51-dimensional Ext. MFCCs, and the combinations of the 39-dimensional MFCCs with the $cG_K$ feature and each of the seven HS experimental feature obtained from MEMD. The performance metrics are evaluated for the TD-1, TD-2 and TD-3 speech utterances of the IITG database. The testing utterances are collected in the practical scenarios. The dimensions of the MEMD features are kept constant, with $K = 4$

Bold values represent the improved results as compared to the existing methods

23% and 14%, for 0 dB, 5 dB and 10 dB, SNR cases, respectively, whereas the mDCF is relatively improved by 24% on the IITG database.

## References

1. L.D. Alsteris, K.K. Paliwal, Further intelligibility results from human listening tests using the short-time phase spectrum. Speech Commun. **48**(6), 727–736 (2006)
2. Y. Bayya, D.N. Gowda, Spectro-temporal analysis of speech signals using zero-time windowing and group delay function. Speech Commun. **55**(6), 782–795 (2013)
3. H. Beigi, *Speaker Recognition: Advancements and Challenges* (INTECH Open Access Publisher, London, 2012)
4. R.K. Bhukya, B.D. Sarma, S.R.M. Prasanna, End point detection using speech-specific knowledge for text-dependent speaker verification. Circuits Syst. Signal Process. **37**(12), 5507–5539 (2018)
5. G. Biagetti, P. Crippa, L. Falaschetti, S. Orcioni, C. Turchetti, An investigation on the accuracy of truncated DKLT representation for speaker identification with short sequences of speech frames. IEEE Trans. Cybern. **47**(12), 4235–4249 (2017)
6. G. Biagetti, P. Crippa, L. Falaschetti, S. Orcioni, C. Turchetti, Speaker identification in noisy conditions using short sequences of speech frames. In: *International Conference on Intelligent Decision Technologies* (Springer, 2017), pp. 43–52
7. H. Boril, P. Fousek, P. Pollák, Data-driven design of front-end filter bank for Lombard speech recognition. In: *Ninth International Conference on Spoken Language Processing* (2006)
8. A. Bouchikhi, A.O. Boudraa, Multicomponent AM-FM signals analysis based on EMD-B-splines ESA. Signal Process. **92**(9), 2214–2228 (2012)
9. C. Charbuillet, B. Gas, M. Chetouani, J. Zarader, Optimizing feature complementarity by evolution strategy: application to automatic speaker verification. Speech Commun. **51**(9), 724–731 (2009)
10. K.T. Deepak, S.R.M. Prasanna, Foreground speech segmentation and enhancement using glottal closure instants and mel cepstral coefficients. IEEE/ACM Trans. Audio Speech Lang. Process. **24**(7), 1205–1219 (2016)
11. K.T. Deepak, B.D. Sarma, S.R.M. Prasanna, Foreground speech segmentation using zero frequency filtered signal. In: *Thirteenth Annual Conference of the International Speech Communication Association* (2012)
12. N. Dehak, P.J. Kenny, R. Dehak, P. Dumouchel, P. Ouellet, Front-end factor analysis for speaker verification. IEEE Trans. Audio Speech Lang. Process. **19**(4), 788–798 (2011)
13. S. Dey, S. Barman, R.K. Bhukya, R.K. Das, B.C. Haris, S.R.M. Prasanna, R. Sinha, Speech biometric based attendance system. In: *National Conference on Communications* (2014)
14. N. Dhananjaya, B. Yegnanarayana, Voiced/nonvoiced detection based on robustness of voiced epochs. Signal Process. Lett. IEEE **17**(3), 273–276 (2010)
15. G.R. Doddington, M.A. Przybocki, A.F. Martin, D.A. Reynolds, The NIST speaker recognition evaluation—overview, methodology, systems, results, perspective. Speech Commun. **31**(2), 225–254 (2000)
16. Y. Ephraim, D. Malah, Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. IEEE Trans Acoust Speech Signal Process **33**(2), 443–445 (1985)
17. P. Flandrin, Some aspects of huangs empirical mode decomposition, from interpretation to applications. In: *International Conference of Computational Harmonic Analysis CHA*, vol. 4 (2004)
18. P. Flandrin, P. Gonçalves, G. Rilling, EMD equivalent filter banks, from interpretation to applications, in *Hilbert-Huang Transform and Its Applications. Interdisciplinary Mathematical Sciences*, ed. by N.E. Huang, S.S.P. Shen (World Scientific Publishing, Singapore, 2005), pp. 57–74
19. S. Furui, Cepstral analysis technique for automatic speaker verification. IEEE Trans. Acoust. Speech Signal Process. **29**(2), 254–272 (1981)
20. T. Ganchev, N. Fakotakis, G. Kokkinakis, Comparative evaluation of various MFCC implementations on the speaker verification task. Proc. SPECOM **1**, 191–194 (2005)
21. S. Gazor, W. Zhang, A soft voice activity detector based on a Laplacian–Gaussian model. IEEE Trans. Speech Audio Proces. **11**(5), 498–505 (2003)
22. F. Gianfelici, G. Biagetti, P. Crippa, C. Turchetti, Multicomponent AM-FM representations: an asymptotically exact approach. IEEE Trans. Audio Speech Lang. Process. **15**(3), 823–837 (2007)

23. M. Hébert, Text-dependent speaker recognition, in *Springer Handbook of Speech Processing*, ed. by J. Benesty, M.M. Sondhi, Y.A. Huang (Springer, 2008), pp. 743–762
24. R.S. Holambe, M.S. Deshpande, *Advances in Non-linear Modeling for Speech Processing* (Springer Science & Business Media, Berlin, 2012)
25. N.E. Huang, Empirical mode decomposition and Hilbert spectral analysis (1998), https://ntrs.nasa.gov/search.jsp?R=19990078602
26. N.E. Huang, S.S. Shen, *Hilbert–Huang transform and Its Applications*, vol. 5 (World Scientific, Singapore, 2005)
27. J.C. Junqua, B. Reaves, B. Mak, A study of endpoint detection algorithms in adverse conditions: incidence on a DTW and HMM recognizer. In: *Second European Conference on Speech Communication and Technology* (1991)
28. K. Khaldi, A.O. Boudraa, A. Komaty, Speech enhancement using empirical mode decomposition and the Teager–Kaiser energy operator. J. Acoust. Soc. Am. **135**(1), 451–459 (2014)
29. A.N. Khan, B. Yegnanarayana, Vowel onset point based variable frame rate analysis for speech recognition. In: *Proceedings of 2005 International Conference on Intelligent Sensing and Information Processing, 2005* (IEEE, 2005), pp. 392–394
30. B.K. Khonglah, R.K. Bhukya, S.R.M. Prasanna, Processing degraded speech for text dependent speaker verification. Int. J. Speech Technol. **20**(4), 839–850 (2017)
31. T. Kinnunen, H. Li, An overview of text-independent speaker recognition: from features to supervectors. Speech Commun. **52**(1), 12–40 (2010)
32. H. Kremer, A. Cohen, T. Vaich, Voice activity detector (VAD) for hmm based speech recognition. In: *Proceedings of ICSPAT* (1999)
33. P. Krishnamoorthy, S.R.M. Prasanna, Enhancement of noisy speech by temporal and spectral processing. Speech Commun. **53**(2), 154–174 (2011)
34. A. Larcher, K.A. Lee, B. Ma, H. Li, Text-dependent speaker verification: classifiers, databases and RSR2015. Speech Commun. **60**, 56–77 (2014)
35. K.A. Lee, A. Larcher, H. Thai, B. Ma, H. Li, Joint application of speech and speaker recognition for automation and security in smart home. In: *INTERSPEECH* (2011), pp. 3317–3318
36. Q. Li, J. Zheng, A. Tsai, Q. Zhou, Robust endpoint detection and energy normalization for real-time speech and speaker recognition. IEEE Trans. Speech Audio Process. **10**(3), 146–157 (2002)
37. D. Mahanta, A. Paul, R.K. Bhukya, R.K. Das, R. Sinha, S.R.M. Prasanna, Warping path and gross spectrum information for speaker verification under degraded condition. In: *22nd National Conference on Communication (NCC)* (IEEE, 2016), pp. 1–6
38. J. Makhoul, Linear prediction: a tutorial review. Proc. IEEE **63**(4), 561–580 (1975)
39. S. Marinov, H.I. Skövde, Text dependent and text independent speaker verification systems. Technology and applications. Overview article (2003)
40. A. Martin, G. Doddington, T. Kamm, M. Ordowski, M. Przybocki, The DET curve in assessment of detection task performance. Technical report, National Institute of Standards and Technology, Gaithersburg MD (1997)
41. N. McLaughlin, J. Ming, D. Crookes, Speaker recognition in noisy conditions with limited training data. In: *2011 19th European Signal Processing Conference* (IEEE, 2011), pp. 1294–1298
42. J. Ming, T.J. Hazen, J.R. Glass, D.A. Reynolds, Robust speaker recognition in noisy conditions. IEEE Trans. Audio Speech Lang. Process. **15**(5), 1711–1723 (2007)
43. K.S.R. Murty, B. Yegnanarayana, M.A. Joseph, Characterization of glottal activity from speech signals. IEEE Signal Process. Lett. **16**(6), 469–472 (2009)
44. A. Paul, D. Mahanta, R.K. Das, R.K. Bhukya, S. Prasanna, Presence of speech region detection using vowel-like regions and spectral slope information. In: *2017 14th IEEE India Council International Conference (INDICON)* (IEEE, 2017), p. 15
45. G. Pradhan, S.R.M. Prasanna, Speaker verification by vowel and nonvowel like segmentation. IEEE Trans. Audio Speech Lang. Process. **21**(4), 854–867 (2013)
46. S.R.M. Prasanna, G. Pradhan, Significance of vowel-like regions for speaker verification under degraded conditions. IEEE Trans. Audio Speech Lang. Process. **19**(8), 2552–2565 (2011)
47. S.R.M. Prasanna, B. Yegnanarayana, Detection of vowel onset point events using excitation information. In: *Ninth European Conference on Speech Communication and Technology* (2005)
48. S.R.M. Prasanna, J.M. Zachariah, B. Yegnanarayana, Begin-end detection using vowel onset points. In: *Workshop on Spoken Language Processing* (2003)

49. L.R. Rabiner, R.W. Schafer et al., Introduction to digital speech processing. Found. Trends® Signal Process. **1**(1–2), 1–194 (2007)
50. K. Ramesh, S.R.M. Prasanna, R.K. Das, Significance of glottal activity detection and glottal signature for text dependent speaker verification. In: *2014 IEEE International Conference on Signal Processing and Communications (SPCOM)* (2014), pp. 1–5
51. B.D. Sarma, S.R.M. Prasanna, P. Sarmah, Consonant-vowel unit recognition using dominant aperiodic and transition region detection. Speech Commun. **92**, 77–89 (2017)
52. R. Sharma, R.K. Bhukya, S.R.M. Prasanna, Analysis of the Hilbert spectrum for text-dependent speaker verification. Speech Commun. **96**, 207–224 (2018)
53. R. Sharma, S.R.M. Prasanna, A better decomposition of speech obtained using modified empirical mode decomposition. Digit. Signal Process. **58**, 26–39 (2016)
54. R. Sharma, S.R.M. Prasanna, R.K. Bhukya, R.K. Das, Analysis of the intrinsic mode functions for speaker information. Speech Commun. **91**, 1–16 (2017)
55. R. Sharma, L. Vignolo, G. Schlotthauer, M.A. Colominas, H.L. Rufiner, S.R.M. Prasanna, Empirical mode decomposition for adaptive AM-FM analysis of speech: a review. Speech Commun. **88**, 39–64 (2017)
56. A. Varga, H.J. Steeneken, Assessment for automatic speech recognition: II. noisex-92: a database and an experiment to study the effect of additive noise on speech recognition systems. Speech Commun. **12**(3), 247–251 (1993)
57. J.D. Wu, Y.J. Tsai, Speaker identification system using empirical mode decomposition and an artificial neural network. Expert Syst. Appl. **38**(5), 6112–6117 (2011)
58. B. Yegnanarayana, S.R.M. Prasanna, J.M. Zachariah, C.S. Gupta, Combining evidence from source, suprasegmental and spectral features for a fixed-text speaker verification system. IEEE Trans. Speech Audio Process. **13**(4), 575–582 (2005)

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

## Affiliations

### Ramesh K. Bhukya[1] · S. R. Mahadeva Prasanna[1,2] · Biswajit Dev Sarma[3]

S. R. Mahadeva Prasanna
prasanna@iitg.ac.in

Biswajit Dev Sarma
biswajit@kaliberlabs.com

[1] Electro Medical and Speech Technology Laboratory, Department of Electronics and Electrical Engineering, Indian Institute of Technology Guwahati, Guwahati 781039, India

[2] Department of Electrical Engineering, Indian Institute of Technology Dharwad, Dharwad 580011, India

[3] Bay Area Advanced Analytics India (P) Ltd., a Kaliber.AI company, Guwahati 781039, India