Circuits Syst Signal Process (2018) 37:5507–5539
https://doi.org/10.1007/s00034-018-0827-3

End Point Detection Using Speech-Specific Knowledge for Text-Dependent Speaker Verification

Ramesh K. Bhukya[1] · **Biswajit Dev Sarma**[1] · **S. R. Mahadeva Prasanna**[1]

Received: 24 November 2017 / Revised: 22 April 2018 / Accepted: 23 April 2018 /
Published online: 4 May 2018

Abstract This paper proposes a method using speech-specific knowledge to detect the begin and end points of speech under degraded condition. The method is based on vowel-like region (VLR) detection and uses both excitation source and vocal tract system information. Existing method for VLR detection uses excitation source information. Vocal tract system information from dominant resonant frequency is used to eliminate spurious VLRs in background noise. Foreground speech segmentation using excitation and vocal tract system information is carried out to remove spurious VLRs in the background speech region. Better localization of the end points is done using more detailed information about excitation source in terms of glottal activity to detect the sonorant consonants and missed VLRs. To include an unvoiced consonant, obstruent region detection is done at the beginning of the first VLR and at the end of last VLR. Detected begin and end points are evaluated by comparing with manually marked end points as well as by conducting the text-dependent speaker verification experiments. The proposed method performs better than some of the existing techniques.

Keywords End point detection · Vowel-like region · Glottal activity · Dominant resonant frequency · Foreground speech segmentation · Speech duration knowledge

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

Biswajit Dev Sarma
s.biswajit@iitg.ernet.in

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

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

1 Introduction

Speech communication is one of the most widely used methods of communication. Therefore, human beings always prefer to communicate with and send commands to machines by voice. Among all the applications involving in speech, speaker verification (SV) has expanded significantly over the years since its inception. SV is the technology that is used to authenticate the identity claim of a person based on his/her speech utterance. Depending on the text used during enrollment and testing phases, SV system can be classified into *text-dependent* and *text-independent* SV system [1,3,15]. In *text-dependent SV* (TDSV) system, the verification text is fixed during the enrollment and testing phase and is already known to the system. In a *text-independent SV* system, there are no such constraints used for enrollment and testing phase. In TDSV system, the system takes the user's voice sample and the identity claim as input to the system and decides whether the given input speech utterance belongs to the particular claimed user or not. In this paper, all studies are presented for a TDSV system suited for cooperative users specifically under degraded conditions like environmental noise and Babble noise.

The state-of-the-art SV systems provide good performance when the speech utterance is of high quality and free from any mismatch [17]. Such a speech utterance is treated as *clean speech* in the present work. However, in most practical operating conditions, the speech signal is affected by different degradations like background noise, background speech, sensor mismatch, channel mismatch, lip smacks, chirping noise, heavy breathing, and other environmental conditions, resulting in *degraded speech* that degrades the TDSV system performance significantly [11,23,53]. There are numerous techniques available for dealing with training and testing speech conditions due to degradation [29,35]. To deal with such degraded speech signals, the compensation is done at the signal level, feature level, model level, score level or all of them together. In this paper, we aimed at the signal-level compensation methods used for removing the effect of noise from the degraded speech utterances. The main goal is to develop a robust and accurate begin and end points detector which can improve the performance of the SV system. The accurate begin and end points detection gives faster response, as useful frames are passed for further processing.

Accurate begin and end points detection is very crucial in text- dependent speaker verification task [10,33,38,54]. In such systems, first begin and end points are detected, and then, feature vectors belonging to the speech region specified by the end points are considered for preparing the templates for training and testing. The testing template is compared with the reference templates using pattern matching technique like dynamic time warping (DTW) [10]. The most widely used end point detection method is based on the *voice activity detection* (VAD) [2,18,20], where voicing decision is made by considering the average energy as a threshold. Most of the existing end point detection methods use energy-based voice activity detection [41,46,51]. Energy-based techniques perform well for clean speech conditions and fail badly in degraded conditions [29]. Figure 1 shows speech signal for the utterance "Get into the hole of tunnels" with non-overlapping background noise [in (a)] and its short-term energy contour [in (b)]. The background noise includes both speech and non-speech background noise. The high energy in the background noise leads to incorrect detection of begin and

end points. In [13,14], pitch information is used for end point detection. However, pitch extraction is also a challenging task in degraded speech condition. In [50], a method for robust end point feature selection is proposed based on a self-adaptive approach. The periodicity-based VAD perform better compared to the conventional energy-based VAD [14]. The statistical model-based VAD was effectively developed based on the decision-directed parameter estimation for the likelihood ratio test which is robust to detect the speech regions under degraded conditions [50]. A better SV system was developed under degraded speech conditions by using speech, only from the less affected speech regions [41]. This robustness is achieved by exploiting the knowledge of vowel-like regions (VLRs) which are relatively high signal-to-noise-ratio (SNR) regions [28,29]. Another method using vowel onset point (VOP) was proposed to detect begin and end points in [54]. Although this VOP-based method is found to be better than the energy-based methods, there exist several drawbacks in the method. This end point detection is based on the VOP-based method developed in [36]. After detecting the VOPs, begin and end points were detected using the knowledge of first and last VOPs. In some cases, there may be a consonant at the beginning of the sentence and instead of detecting this consonant, begin point is marked 200 ms before the first VOP. Similarly, end point is marked 200 ms after the last VOP without estimating the length of the vowel. Therefore, the begin and end points detected in that way may not be correct, and some explicit consonant region detection may be more useful. Moreover, instead of using the last VOP, vowel end points (VEPs) can be used. An algorithm for vowel-like region (VLRs) detection by detecting VLR onset points (VLROPs) and VLR end points (VLREPs) was proposed in [29], and the method was further improved in [27,35]. VLRs include vowels, diphthongs and semivowels. Since objective is to detect speech end points, instead of using VOP, VLR can be used as it will provide both VLROP and VLREP. Figure 1a shows the manually marked VLROPs and VLREPs along with the speech signal. Although the begin point is near to the first VLROP, the end point is much deviated from the last VLREP of the speech region. The obstruent region at the end of the sentence must be detected to correctly detect the end point.

Another disadvantage of VOP-based method is the use of only excitation source information. Excitation source information normally captures the impulse-like excitation of the source. Sometimes impulse-like noise present in the signal is also captured by the instantaneous source information and detects them as spurious VLRs [43]. Therefore, vocal tract system information is required for eliminating the spurious detections. Moreover, VOPs are detected well when there is a consonant attached to it, for example, in case of an obstruent–vowel unit. It is because the signal strength changes suddenly at the onset of the vowel when there is a stop consonant or a fricative. In case of semivowel–vowel units, the VOP detection performance degrades [34,37]. Vowel end points (VEPs) are also detected well when there is a consonant at the end. In the absence of a consonant, the vowel energy decreases slowly and it becomes difficult to detect the VEP. Such vowel regions with very low strength can be detected by exploring more detailed information of the excitation source in terms of glottal activity. Glottal activity detection (GAD) will detect some other sonorant consonants along with the VLRs. But in this work, those sonorants are desirable as the final goal is to detect the end points of speech.

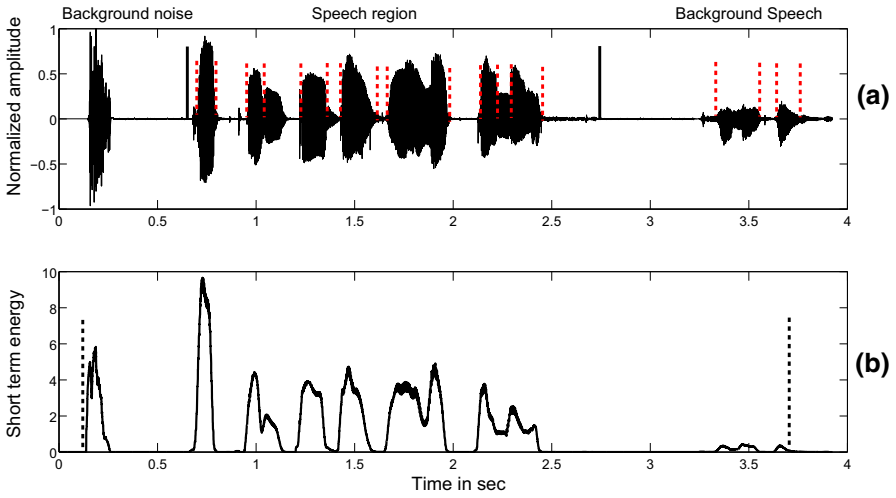

Fig. 1 (**a**) Speech signal containing non-overlapping background noise and background speech. Dark vertical lines show the ground-truth begin and end points. Dotted lines are manually marked vowel-like region onset and offset points in the speech region. (**b**) Short-term energy of the signal in (**a**). Dotted lines show the end points detected by the energy-based end points detection

In Fig. 1, it can be seen that the speech is degraded with background speech, and VLROPs and VLREPs are detected in the background speech regions as well. In such a scenario, a foreground speech segmentation module may be required to properly detect the begin and end points. It is also observed that the proposed method does not work where non-speech and background speech regions are overlapped with speech regions.

Taking all these issues into consideration, this work proposes a method to detect the begin and end points using some speech-specific knowledge. The proposed method uses both excitation source and vocal tract system information to reduce the spurious detection. Following is the step-by-step process of the proposed method.

- The vowel-like regions (VLRs) are detected using excitation source information.
- The vocal tract system information obtained from dominant resonant frequency (DRF) is used to reduce spurious detection in background noise.
- Glottal activity detection is used to detect the sonorant consonants present at the beginning and end of the speech signal, which cannot be detected by VLRs detection.
- Obstruent consonants are detected by using a sub-fundamental frequency filtering method.
- To eliminate background speech, a foreground speech segmentation module is used.
- Finally, speech duration knowledge is used to further refine the output of the begin and end points detection.

Depending on the background speech and the presence of different kinds of degradation, a poor system performance may be encountered. To improve such system, speech-specific knowledge is applied on the degraded speech so as to improve the per-

formance in a better way. Even though most of these are existing techniques, our contribution is in terms of exploiting the advantages of different individual techniques and combining them in a suitable manner for robust end point detection in a TDSV system.

The rest of the paper is organized as follows: Section 2 describes significance of different speech-specific knowledges for the detection of speech end points. Section 3 illustrates the procedure for robust end point detection. Experimental evaluation of the proposed end point detection (EPD) method is performed in Sect. 4. The use of the proposed EPD in text-dependent speaker verification system is described in Sect. 5. Section 6 evaluates the proposed text-dependent speaker verification system, and Sect. 7 summarizes the work and concludes.

2 Significance of Speech-Specific Knowledge for End Point Detection (EPD)

The task of end point detection involves marking the starting and end points of a speech utterance. The presence of background noise makes it difficult to locate the end points. In practical environment, there are different types of background noises present. Each background noise has their own type of characteristics. Therefore, looking the end point detection problem from speech knowledge point of view is more appreciable in practical environment. Some of the existing methods indirectly use speech-specific knowledge. However, those methods fail because the speech knowledge used in those systems is not sufficient. In this section, speech-specific knowledge required for eliminating the limitations of energy-based and VOP-based methods is described in detail.

2.1 Vowel-Like Region Onset and End Point Information

Vowels, semivowels and diphthongs are considered as the vowel-like regions (VLRs). VLRs are high SNR regions and easier to detect in degraded condition [27,30]. There are many methods in the literature for the VLR detection [27,45]. Most of these methods are based on detection of VLROP and VLREP. Knowledge of VLREPs can be used for detecting the speech end points, instead of using the last VOP. Vowels are of variable duration, and many languages have long and short version of the same vowel. Detecting the last VOP and marking the end point 200 ms after the VOP are not a good idea in such cases, rather the use of VLREP may be more accurate. The method described in [27] is used for VLRs detection in this work. Although VLROP and VLREP detection uses excitation source information only, it uses two different evidences to explore the information unlike the previous method [54], where VOPs are detected using only a single evidence. One VLROP evidence is obtained using the *Hilbert envelope* (HE) of *linear prediction* (LP) residual by processing the speech signal through the following steps : The speech frame is processed in blocks of 20-ms frame size with a frame shift of 10 ms. For each 20-ms frame size, 10th-order LP analysis is performed to estimate the *linear prediction coefficients* (LPCs) [22]. The speech frame is passed through the inverse filter to extract the LP residual. This excitation source characteristics are further enhanced by determining the HE of LP residual [32]. The detailed method used to detect the VLROP evidences using HE of LP residual is described below.

Method 1: VLROP evidence using HE of LP residual

→ Let $l(n)$ is the LP residual signal
→ Its analytic signal is $l_a(n)$ is given by

$$l_a(n) = l(n) + jl_h(n)\cdot$$

where, $l_h(n)$ is the Hilbert transform of $l(n)$
→ Let $h_l(n)$ be HE, which is the magnitude of $l_a(n)$

$$h_l(n) = |l_a(n)|$$

$$h_l(n) = \sqrt{l^2(n) + l_h^2(n)}$$

→ This HE of LP residual enhances information about glottal closure instants (GCIs)
→ The evidence is further enhanced by taking maximum value of the HE of LP residual for every 5-ms block with one sample shift
→ The smooth contour is convolved with a first-order Gaussian differentiator (FOGD) window of length 100 ms and standard deviation of one-sixth of window
→ The output of the convolution is the VLROP evidence using HE of LP residual

Low-frequency components are accommodated in the HE ($h_l(n)$), and relatively high-frequency components are accommodated in phase ($\Phi(n)$) of the analytic signal [47]. Hence, the HE of LP residual enhances speaker information about glottal closure instants (GCIs) [34,35].

Following steps are used for computing another evidence from the zero-frequency filtered signal (ZFFS). The output of zero-frequency filter (ZFF) consists of excitation information which is mainly impulses due to excitation. The rate of change of excitation strength is the main event at VLROP [35].

Method 2: VLROP evidence using ZFFS

The algorithmic steps for computing VLROP evidence from ZFFS are as follows:
→ Difference the speech signal $s(n)$

$$d(n) = s(n) - s(n-1)$$

→ Compute the output of a cascade of two ideal digital resonators at 0 Hz

$$y(n) = -\sum_{k=1}^{4} a_k y(n-k) + d(n)\cdot$$

where, $a_1 = 4$, $a_2 = -6$, $a_3 = 4$, $a_4 = -1$ and $d(n)$ is the differenced speech signal
→ Remove the trend, i.e.,

$$\hat{y}(n) = y(n) - \bar{y}(n)\cdot$$

where $\bar{y}(n) = \frac{1}{(2N+1)} \sum_{m=-N}^{N} y(n+m)$ and $2N+1$ corresponds to the average pitch period computed over a longer segment of speech
→ The signal ($\hat{y}(n)$) obtained after removing the trend is the ZFFS
→ The strength of excitation at the epochs is extracted by computing the first-order difference of ZFFS
→ The second-order difference of ZFFS contains change in the strength of excitation
→ This change is detected by convolving with a FOGD
→ The convolved output is called the VLROP evidence using ZFFS

Final VLROP evidence using the excitation source information is obtained by adding the two evidences and normalizing by the maximum value of sum. The location of peaks between two successive positive- to-negative zero crossings of the combined evidence represents the hypothesized VLROPs. To find the VLREP evidences, both the

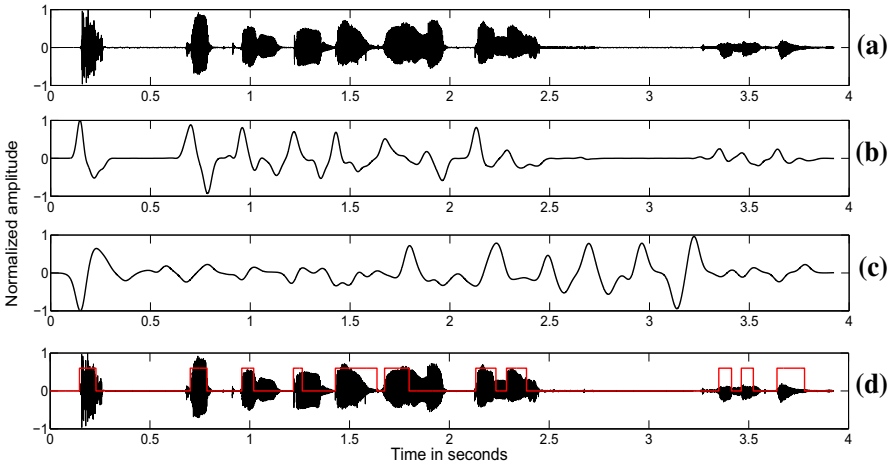

Fig. 2 (**a**) Speech signal containing non-overlapping background noise, (**b**) VLROP and (**c**) VLREP evidences obtained using excitation source information. (**d**) Detected VLRs marked using dark lines

smoothed HE of LP residual and second-order difference of ZFFS are convolved with FOGD from right to left as opposed to left to right in case of VLROPs and all other steps are kept same [25]. The locations of peaks between two successive positive-to-negative zero crossings of the combined VLREP evidence represent the hypothesized VLREPs.

The number of hypothesized VLROPs and VLREPs will be different due to independent detection, with misses and spurious detections occurring for both events. To get the most optimum sets of VLROP and VLREP, an algorithm is proposed using some speech-specific knowledge [27]. The algorithm forced the detection of missing cases in one evidence, if other evidence is sufficiently strong. The algorithm also reduces spurious detection of one event using knowledge of the other. Refined sets of VLROP and VLREP are used to obtain the VLRs.

Figure 2a shows the same speech signal as shown in Fig. 1a containing speech and non-speech background noise. Figure 2b–d shows the VLROP and VLREP evidences and detected VLRs, respectively. The non-speech background noise present in between 0 and 0.5 s may also have impulse-like characteristic due to which the region is detected as VLR. Some portion of speech background present in between 3 and 4 s are also detected as VLRs. These spurious detections will lead to wrong detection of the begin and end points. Therefore, some vocal tract system information should be explored to reduce the spurious detection.

2.2 Dominant Resonant Frequency Information

To remove missed spurious VLRs, dominant resonant frequency (DRF) can be used as vocal tract system information. DRF is the frequency which is resonated the most by the vocal tract [31]. DRF is computed from the Hilbert envelope of numerator group delay (HNGD) spectrum of zero-time windowed speech [12,56]. DRF is less than 1 kHz for vowel-like regions. Non-speech noise mostly contains high-frequency

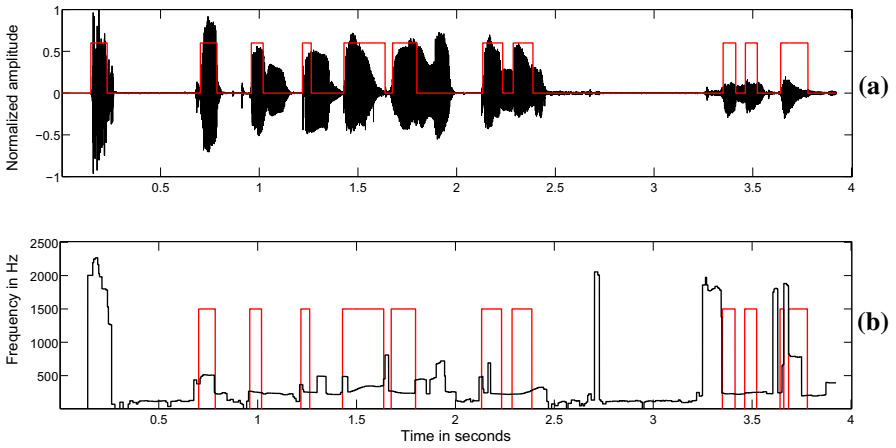

Fig. 3 (**a**) Speech signal with non-overlapping background noise. Hypothesized VLRs are shown using dark lines. (**b**) DRF contour of the speech along with refined VLRs. Spurious VLRs in non-speech background region are removed after using the DRF information

components. Hence, DRF is more than 1 kHz for non-speech noise. This knowledge can be used for identifying and removing the spurious VLRs.

Figure 3a shows speech with background noise along with the detected VLRs. The DRF contour is shown in Fig. 3b. The DRF values are low for VLRs and high for background noise. The VLRs where DRF value is more than 1 kHz are considered as non-speech region. In Fig. 2b, the refined VLRs are shown using dotted lines. It is seen that spurious VLRs in the non-speech background noise region in between 0 and 0.5 s is removed. Thus, DRF can be used for removing some of the spurious VLRs.

The procedure for detection of non-speech background noise using DRF is described below:

Method 3: Detection of non-speech background noise using DRF

→ Differenced speech signal $s[n]$ is multiplied two times with a zero time window ($w_1[n]$)

$$w_1[n] = \begin{cases} 0, & n = 0 \\ \dfrac{1}{4sin^2(\pi n/2N)}, & n = 1, 2, \ldots, N-1 \end{cases} \cdot$$

where N is window length. In this work, N is considered to be samples equivalent to 5 ms

→ Zero time windowing operation is performed to get better time resolution

→ The truncation effect at the end of the window is reduced by multiplying with a tapering window function $w_2[n]$ given by

$$w_2[n] = 4cos^2(\tfrac{\pi n}{2N}), n = 0, 1, \ldots N-1$$

→ Numerator group delay function ($g[k]$) of the resultant signal is obtained to nullify the loss in frequency resolution

$$g[k] = X_R[k]Y_R[K] + X_I[k]Y_I[K], k = 0, 1 \ldots, N-1 \cdot$$

where $X[k] = X_r[k] + jX_I[k]$ is the N-point discrete Fourier transform (DFT) of the windowed signal $x[n]$ and $Y[k] = Y_r[k] + jY_I[k]$ is the N-point DFT of the sequence $y[n] = nx[n]$

→ Final spectrum is obtained by taking Hilbert envelope of the differenced numerator group delay function

→ The frequency corresponding to the maximum value in the spectrum is considered as the DRF

→ The VLRs where DRF value is more than 1 kHz are declared as non-speech region

2.3 Foreground Speech Segmentation

The speech of the speaker who is close to the microphone sensor and directly communicates to a speech application is known as the foreground speech. The rest of the acoustic background captured by the sensor is known as the background degradation [5]. Background degradation contains both speech and non-speech noises. Non-speech noise can be removed by using DRF as described in the previous subsection. To remove the background speech, a foreground speech segmentation is necessary. A foreground speech segmentation algorithm was proposed in [5], which was further modified in [6]. The modified version of the foreground speech segmentation method is used in this work. The method uses excitation source and vocal tract system information. Excitation source information is extracted using ZFFS analysis, and vocal tract auditory features are extracted from the modulation spectrum energy.

Method 4: Detection of foreground speech segmentation

The foreground speech segmentation algorithm is described as follows:
→ Excitation source information is extracted using ZFF analysis
→ The ZFF signal is processed in blocks of 50 ms with shift of one sample
→ Normalized first-order autocorrelation coefficients and strength of excitation are extracted
→ The features show high value in the foreground speech regions compared to background speech region
→ Vocal tract articulatory features are extracted from the modulation spectrum energy using compressive Gamma chirp auditory filter
→ The linear combination of these features are further subjected to an end point detection algorithm to segment the foreground speech regions over a longer segment of speech

Figure 4a shows speech signal with background speech in the 3–4 s region. Figure 4b shows the final evidence obtained from the foreground segmentation algorithm along with the segmented boundaries (in dotted lines). With the help of foreground segmentation, the spurious VLRs in the background speech region can be removed and the VLRs present in the foreground speech region can be retained.

2.4 Glottal Activity Information

After detecting the VLRs, glottal activity information can be explored to add a sonorant consonant at the begin or at the end of speech utterance. GAD detects the VLRs as well as other sonorant consonant regions, such as voice bars and nasals. Therefore, better localization of voiced region can be done by using GAD. This GAD method also helps to minimize the VLRs miss detection rate.

A procedure for detection of voiced region using ZFF method is described in [8,24]. Detection of voiced speech involves detection of significant glottal activity in speech. Rest of the regions include unvoiced speech, silence as well as background noise. The method exploits the nature of ZFFS that the addition of a small amount of noise to the speech signal does not affect the zero crossings of ZFFS in the voiced region, whereas it leads to zero crossings at random locations in unvoiced and silence regions. The reason behind this is the glottal closure that takes part in producing the most significant impulse-like excitation in voiced speech. The epochs due to such impulse-

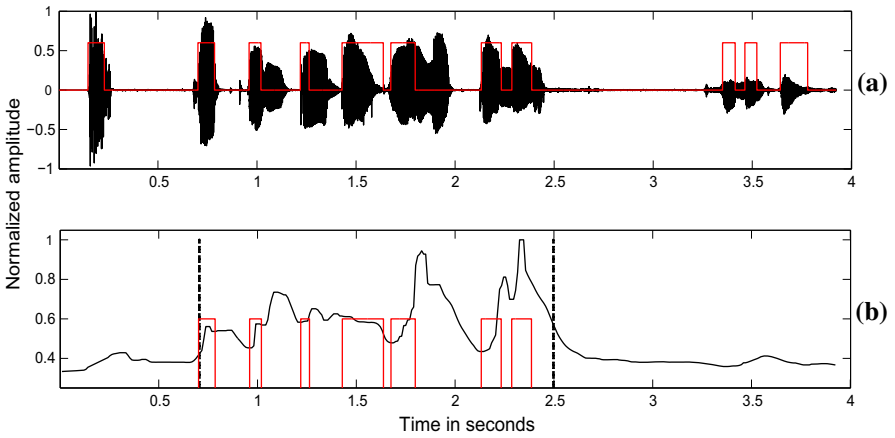

Fig. 4 (**a**) Speech signal with non-overlapping background noise. Detected VLRs are shown using dark lines. (**b**) Final evidence for foreground speech segmentation. Vertical dotted lines show the segmentation boundaries. Spurious VLRs in speech background can be removed using foreground speech segmentation

like excitations are robust to noise and can be located with high accuracy in the presence of degradation. On the other hand, unvoiced and silence region do not contain any significant excitation and result in zero crossings located at random instants. The drift in the epoch location after the addition of a little amount of noise is used to segment the voiced regions from rest of the regions. Additional spurious epochs are eliminated using the knowledge of pitch and jitter information. The detailed method for detection of voiced regions is as follows:

Method 5: Detection of voiced regions using GAD

Procedure for detection of voiced region using ZFF is described in the following way:
→ Obtain the epoch locations of speech using ZFF
→ Add 20 dB of white Gaussian noise to the clean speech
→ Obtain the epoch locations of noisy speech using ZFF
→ If the difference between the epoch locations obtained in the two cases is less than 2 ms, then retain it. otherwise discard the epoch
→ Calculate the instantaneous pitch period using the epoch locations and eliminate the epoch if its pitch period is more than 15 ms
→ Calculate the instantaneous jitter and eliminate the epoch if jitter is more than 1 ms

Figure 5 shows speech signal along with the hypothesized VLRs [in (a)] and detected glottal activity region [in (b)]. It is seen that there are some missed VLRs at the end of speech utterance (around 2.4 s). These missed VLRs are detected by the glottal activity detection algorithm. Further, there is one nasal sound in between 1 and 1.2 s which is not considered as VLR. GAD will be helpful if such sonorant is present at the beginning or at the end of speech utterance. Detection of such sonorant sound is required for appropriately detecting the end points.

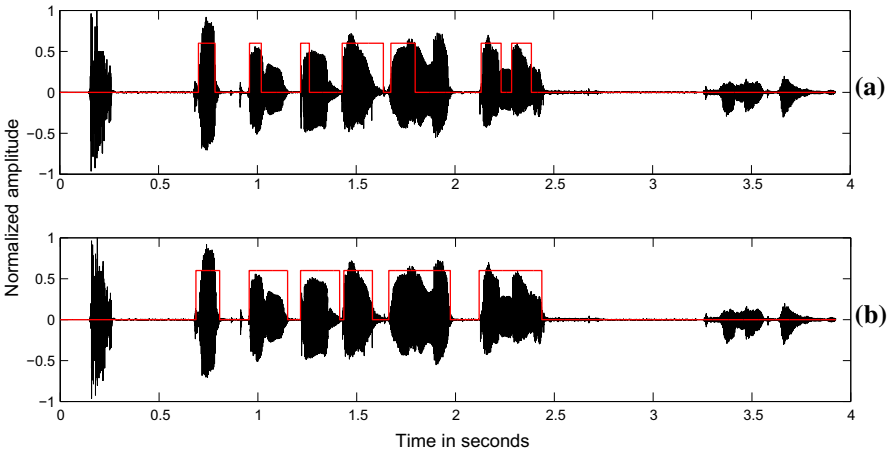

Fig. 5 Speech signal with hypothesized (**a**) VLRs and (**b**) glottal activity regions. Missed VLRs and the sonorant consonants are detected by using glottal activity detection

2.5 Obstruent Information

Stop consonants, fricatives and affricates are known as obstruents. There may be an obstruent consonant at the begin and end of speech utterance. The obstruents are not detected by the VLR detection. To include them in the speech region, an obstruent detection algorithm is required. In [44], a method is proposed to detect the dominant aperiodic regions in speech using sub-fundamental frequency filtering. Since aperiodic component is dominant in burst and frication region of most of the obstruents, the same method is used in this work to detect the obstruents. The method is a modification of the ZFF method. The steps involved in the method are described below.

Method 6: Detect the dominant aperiodic regions in speech

Procedure for detection of dominant aperiodic region is described in the following way:
→ Pass the speech signal through a 0-Hz resonator four times
→ Obtain the sub-fundamental frequency filtered signal by removing the trend four times using a window size of three pitch periods
→ Enhance the filtered signal in the dominant aperiodic region by squaring it
→ Smooth the enhanced signal using a 5-ms window
→ For further smoothing, locate the peaks in the signal using a peak picking algorithm and repeat the peak value till the next peak is obtained
→ A FOGD of 100 ms length is convolved with the signal to get the final evidence
→ A positive peak in the final evidence gives the probable onset of an obstruent
→ The first positive peak present within 100 ms before a VLROP is considered as the onset of obstruent and
→ The last negative peak present within 100 ms after the VLREP is considered as the offset of an obstruent

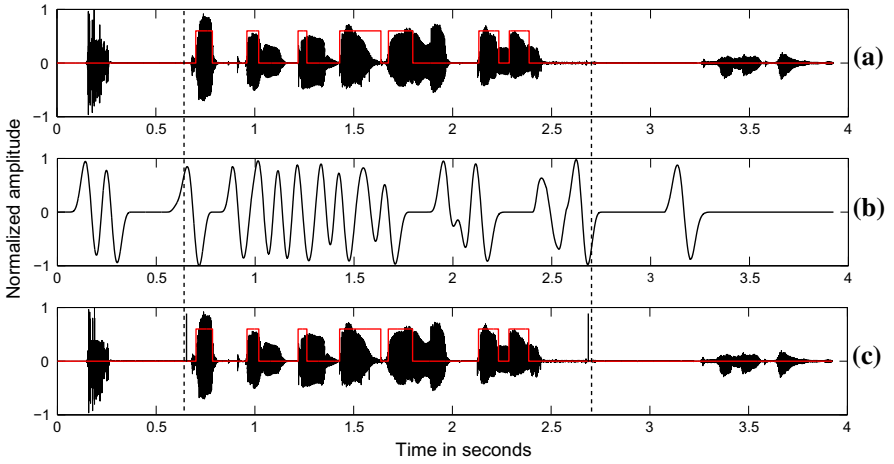

Fig. 6 (**a**) Speech with hypothesized VLRs after spurious removal. (**b**) Final obstruent evidence. (**c**) Speech with refined begin and end points. Vertical dark lines show the detected end points, and long vertical dotted lines show ground-truth end points. Burst region at the beginning and the frication region at the end of the signal are included after the refinement

Figure 6a shows a speech signal with hypothesized end points detected by the VLR detection method. Vertical dotted lines show the ground-truth end points. The first VLROP and last VLREP are much deviated from the ground-truth marking. Figure 6b shows the final obstruents evidence. The first positive peak present within 100 ms of first VLROP is considered as the refined begin point, and the last negative peak present within 100-ms region of the last VLREP is considered as the refined end point. In Fig. 6c, it can be observed that the end points are detected with more accuracy after using obstruent detection.

2.6 Speech Duration Knowledge

Sometimes database is collected from an unknown environment without any restriction on the testing condition. The subject is not restricted to speak the given sentence in a proper way, but left free to speak in a very casual way. In such cases, the subject may try to speak one or two words several times. For example instead of saying "*Lovely picture can only be drawn*", one may say "*Lovely [pause] lovely picture can only be drawn*". Sometimes one may repeat a word twice and can speak something which is not part of the sentence. When these extra words are uttered without any pause, it is not possible to do any further rectification. However, if these words are uttered at the beginning or at the end leaving some silence in between the actual utterance and the extra word, it is possible to remove the extra utterance by using speech duration knowledge (SDK). In normal speech, distance between two subsequent VLRs cannot be more than 300 ms [36]. If it is more than 300 ms, then it is not part of the given sentence. This knowledge will also help to remove other non-speech spurious VLRs which are detected 300 ms away from the speech utterance. Following procedure is followed to use the SDK.

Method 7: Detection of speech regions using speech duration knowledge

Procedure for detection of refined end point regions is described in the following way:
→ First, location of all the VLROPs are obtained
→ Average of the Euclidean distances from one VLROP to all other VLROPs are computed
→ The VLROP having the minimum average Euclidean distance is considered as the center (C) of the speech utterance

$$C = \arg\min_i \frac{1}{N-1} \sum_{j=1,\neq i}^{N} (V_i - V_j)^2.$$

where $1 \leq i \leq N$, $1 \leq j \leq N$, C is the detected center of speech utterance, N is the number of detected VLRs and V_i is the i^{th} VLROP in number of samples
→ Now, starting from the center of the speech, duration between two successive VLRs is calculated on either side until the duration is found to be greater than 300 ms or the peripheral VLR is reached
→ If the duration between two successive VLRs is found to be greater than 300 ms, then between the two VLRs, the one which is closure to C is declared as the peripheral VLR

3 Robust End Point Detection Using Speech-Specific Knowledge

The speech-specific knowledge described in the previous section can be systematically used to detect the end points in practical conditions containing both speech and non-speech degradation. Figure 7 shows the block diagram for the proposed method. First VLRs are detected from the speech signal. Spurious VLRs in the non-speech background are removed by using DRF information. Spurious VLRs in speech background are removed using foreground speech segmentation. Then, missed VLRs and sonorant consonants present at the begin or end are detected using GAD. The obstruent consonants present at the begin or end are detected, and end points of peripheral VLRs are modified. Finally, SDK is used to further remove the spurious VLRs. The VLROP of first VLR and VLREP of last VLR are declared as the begin and end points, respectively.

Figure 8 illustrates the proposed detection procedure. The same speech files used in the previous sections are used for the illustration. Figure 8a shows the speech signal with non-overlapping non-speech noise and speech background. First step is to detect the VLRs. Figure 8b shows the detected VLRs. Second step is to remove the non-speech noise using the DRF information. The spurious VLRs where the DRF is exceeding 1 kHz are removed, and refined VLRs are shown in Fig. 8c. Third step is to remove the speech background which is done using foreground speech segmentation. The detected boundaries and refined VLRs obtained after foreground speech segmentation are shown in Fig. 8d. In the next step, the glottal activity regions are detected. The detected glottal activity regions are added to VLRs and are shown in Fig. 8e. This includes some sonorant regions which are not detected by the VLR detection algorithm. Fifth step is to detect the obstruents at the beginning and at the end of the speech utterance. The modified begin and end points using obstruent evidence are shown in Fig. 8f. The arrow in Fig. 8f shows the center of speech utterance detected. Starting from this point, the duration between successive VLRs is computed in both sides. For this utterance, the inter-VLR time interval is less than 300 ms for all cases, and hence, no further modification is incorporated using the SDK knowledge.

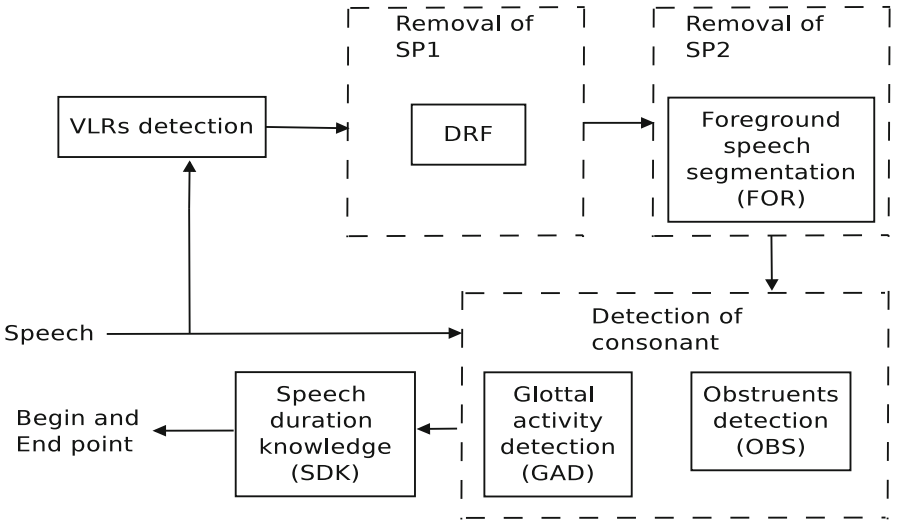

Fig. 7 Block diagram of proposed end point detection. SP1 and SP2 denote spurious VLRs in non-speech and speech background, respectively

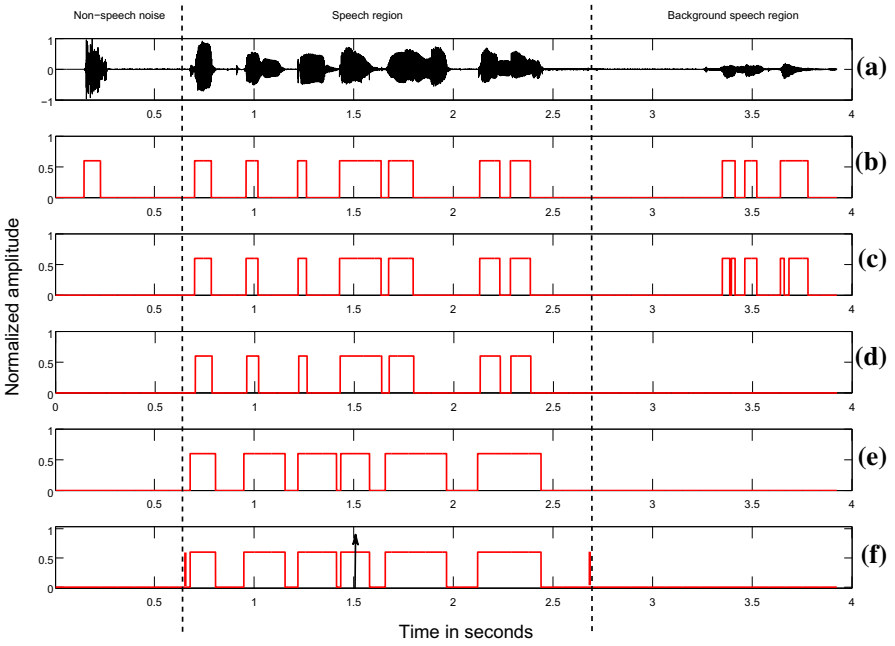

Fig. 8 Illustration of the begin and end point detection procedure. (**a**) Speech signal with non-speech noise and background speech. (**b**) Detected VLRs. (**c**) VLRs after removing the non-speech noise using DRF information. (**d**) VLRs after removing the speech background using foreground speech segmentation. (**e**) Detected glottal activity regions added to the VLRs. (**f**) Refined begin and end point using obstruent information. The arrow around 1.5 s shows the center C of the speech utterance. Duration between successive VLRs is less than 300 ms. Therefore, no further modification is made using SDK knowledge. Dotted line shows ground-truth manual marking

4 Experimental Evaluation of Proposed EPD

4.1 Databases

In order to evaluate the robustness of proposed algorithm in degraded speech conditions, we conducted the text-dependent SV experiments on the two databases, namely the IITG database [4,7] and the RSR2015 database [19]. The RSR2015 database is mainly designed for text-dependent speaker recognition with pre-defined pass phrases. This database contains 71 h of audio recording, recorded on a set of six portable devices including mobile phones and tablets, in a closed office environment. The speech data is recorded from 300 English speakers (143 females and 157 males) decorating the variety of accents spoken in Singapore. These 300 speakers were selected to be representative of the ethnic distribution of Singaporean population, with age ranging from 17 to 42 years old. Each of the speakers recorded 9 sessions; each session consists of 30 pre-defined pass phrases. In this work, three pass phrases are optionally have been selected for the experimental studies. The duration of the utterances varies between 3 and 5 s. Out of 9 utterances (sessions) for every pass phrase, 3 utterances are used for training, and remaining 6 utterances are used for testing. The pass phrases used in this study are shown below.

TD-1 : *"Only lawyers love millionaires"*.
TD-2 : *"I know I did not meet her early enough"*.
TD-3 : *"The Birthday party has cupcake and ice-cream"*.

As the RSR2015 database is collected in a closed office environment, it is considered as clean database. To evaluate the proposed method in degraded condition, the speech signal is corrupted with babble noise [52]. The energy level of the noise is scaled such that the overall *signal-to-noise ratio* (SNR) of the degraded speech varies from 20 to 0 dB, in steps of 5 dB.

The IITG database was collected from the course students with attendance as an application to address issues in the practical deployment of text-dependent SV system [4,7,16,21]. The speech biometric-based attendance system developed for course students of IITG was used on a regular basis for marking attendance [4,7,49]. The students used to call to a voice server-based toll-free number to mark attendance using fixed set of mobile handsets kept at the department office. In the enrollment phase, three predefined pass phrases were recorded for each enrolled speaker in an anechoic chamber. The data were recorded in three different sessions using a set of mobile handsets.

The three pass phrases from the IITG database are:

TD-1 : *"Don't ask me to walk like that"*.
TD-2 : *"Lovely pictures can only be drawn"*.
TD-3 : *"Get into the hole of tunnels"*.

During the testing phase, the speakers are prompted to utter one of the three predefined utterances randomly selected by the voice server. Collection of the testing data was continued for an entire academic semester from a population of 325 speakers (286 males and 39 females). On an average, 30 utterances are available per speaker

for each phrase. These speakers constitute an age group in between 25–35 years. The duration of the utterances of the database varies between 3 and 5 s. During collection of the testing database, the speakers could move freely within, and in and out of the open hall in the department office and also are allowed in free environment. This makes the testing data more practical because the data include background noise, background speech and other environmental conditions. Due to mismatches in sensor, style of speech and environment between the enrollment and the testing phases, the TDSV task becomes more challenging in IITG database. Hence, there is no need to add artificial noise in the IITG database for the assessment of the TDSV system.

4.2 Performance Evaluation

The performance of the end point detection (EPD) algorithm is evaluated in terms of two metrics : *Identification rate* (IR) and *Identification error* (IE) [48]. These utterances are marked manually for ground-truth begin and end points of the speech signal. These begin and end points are called as reference points. The begin and end points detected by applying different automatic detection algorithms are called as estimated points. Let l_b^{ref} and l_e^{ref} denote the midpoints (in ms) of the begin and end frames of a speech utterance/signal. All speech frames before the begin frame and after the end frame are considered as silence frames. Let l_b^{est} and l_e^{est} denote the midpoints (in ms) of the begin and end frames of the speech signal, as estimated by the EPD algorithm. If $l_b^{est} \in l_b^{ref} \pm 500$ ms, then the begin point is considered to be correctly estimated. Similarly, if $l_e^{est} \in l_e^{ref} \pm 500$ ms, then the end point is considered to be correctly estimated. We may define IR and IE as:

IR (%): Percentage of speech utterances for which the begin point (end point) is correctly estimated.
IE (ms): For every begin point (end point) that is correctly estimated, the absolute difference between the reference begin point (end point) and its estimate is calculated. In other words, $| l_b^{est} - l_b^{ref} |$ ($| l_e^{est} - l_e^{ref} |$) is calculated. The average value of the absolute difference for all the correctly estimated begin points (end points) is denoted as IE.

4.2.1 Performances on the RSR2015 Database Under Clean and Degraded Conditions

The performance of proposed begin and end point detection method is evaluated for clean speech using 300 speakers data from the RSR2015 database for three utterances *"Only lawyers love millionaires"*,*"I know I did not meet her early enough"*, and *"The Birthday party has cupcake and ice-cream"*. The begin and end points are marked manually to obtain the ground truths.

In Table 1, performance of automatically detected begin and end points is shown for TD-1 phrase of clean RSR2015 database. The proposed method is compared with different existing methods in terms of IR and IE. The evaluation is also carried out at different stages in the block diagram of the proposed method so that comparison

Table 1 Performance of begin and end point detection in terms of identification rate and identification error, evaluated on clean RSR2015 database

Technique		Energy	FSS	GAD	VLR	VLR + DRF	VLR + DRF + FSS	VLR + DRF + GAD	VLR + DRF + GAD + FSS	VLR + DRF + GAD + FSS + OBS	VLR + DRF + GAD + FSS + OBS + SDK
Method	Metric										
Begin	IR (%)	96.64	96.52	98.39	98.39	97.82	97.80	98.46	98.28	99.02	**99.17**
	IE (ms)	163	53	177	73	61	177	52	51	48	**39**
End	IR (%)	96.82	80.00	90.00	94.93	95.64	95.93	95.46	96.02	96.64	**98.78**
	IE (ms)	**69**	115	323	207	215	299	238	178	199	204

is made with the individual methods as well as with different possible combinations. It can be seen that the VLR-based detection gives around 1.75% improvement over energy-based detection for begin and end points. In case of GAD-based detection, performance increases by around 1.75% for begin points and decreases by around 7% for end points compared to energy-based detection. In case of FSS, the performance decreases for both begin and end points. It is also observed that performance of begin point detection increases for different combinations. For the proposed detection method which uses combination of all individual methods, the performance increases by around 2.5 and 2% for begin and end points, respectively. In terms of IE, the proposed method shows better performance compared to the individual methods for begin points. However, the IE is found to be very high in case of the end points.

To evaluate the robustness of the proposed begin and end point detection algorithm, the testing utterances of the RSR2015 database are corrupted with babble noise from NOISEX-92 database [52]. The SNR is varied from 20 to 0 dB in steps of 5 dB. In Table 2, the performance is shown for TD-1 phrase of the RSR2015 database. Due to degradation in the speech signal, the performance is quite low compared to the clean condition. The energy-based end point detection shows around 10% decrement in performance for 20 dB SNR. Similar decrement is observed for other individual methods. However, for the proposed method using combination of all the algorithms, an overall improvement is observed across all 5 levels of degradation. In some cases, we can observe that the individual method is performing better than the proposed method in terms of IR or IE or both. For example in case of 10 dB SNR, GAD is showing 99.28% IR and 298 ms IE for begin points, which is slightly better than the performance of the proposed method. However, for the GAD method with the same SNR, we can observe that the performance of the proposed method is significantly higher than the GAD for end point detection. Similarly, although FSS is performing better than the proposed method for end points, it is under performing with a significant margin for the begin points. On an average, the performance of the proposed method is found to be better than the individual methods and different intermediate combinations of the individual methods. Complementary advantages of the individual methods lead to improved performance for the proposed method. Results for TD-2 and TD-3 phrases of RSR2015 database are shown in Fig. 9 for both clean and degraded speech conditions, and similar trend is observed.

In Table 2, it can be observed that even though IR is better for the proposed methods, most of the times the IE in case of begin points is poorer compared to some of the individual methods. This is because of the method used in this work for computing IE, where only detected end points are considered. The end points beyond ± 500 ms of the reference mark are not considered for IE computation.

4.2.2 Performances on the IITG Database

Same experiments are reported for the IITG database which is more practical data in terms of session, sensor, background degradation and environmental conditions. Table 3 shows the performance of different detection algorithms on IITG database. It can be observed that for the practical data, the performance of the energy-based begin point detection is similar to that of the RSR2015 database. However, in case

Table 2 Performance of begin and end points detection in terms of identification rate (IR) and identification error (IE)

Technique			Energy	FSS	GAD	VLR	VLR + DRF	VLR + DRF + FSS	VLR + DRF + GAD	VLR + DRF + GAD + FSS	VLR + DRF + GAD + FSS + OBS	VLR + DRF + GAD + FSS + OBS + SDK
SNR (dB)	Method	Metrics										
20	Begin	IR (%)	84.46	96.07	83.57	82.14	96.07	**99.64**	98.04	97.86	98.75	96.78
		IE (ms)	165	73	147	155	73	**54**	224	218	224	147
	End	IR (%)	88.93	**93.39**	78.00	59.00	62.00	68.00	58.05	86.43	78.00	86.96
		IE (ms)	220	204	171	168	168	**166**	171	197	173	224
15	Begin	IR (%)	77.14	79.82	**99.82**	97.68	99.82	96.43	98.21	97.25	97.68	96.61
		IE (ms)	355	200	**97**	286	193	357	304	332	322	229
	End	IR (%)	49.82	**89.28**	80.00	50.00	61.14	68.9	77.50	76.16	74.1	77.14
		IE (ms)	354	199	**169**	170	177	171	174	184	175	248
10	Begin	IR (%)	60.00	71.61	99.28	96.25	95.72	**99.29**	96.96	96.98	96.42	96.25
		IE (ms)	376	**275**	298	303	406	298	345	352	363	310
	End	IR (%)	45.89	**79.46**	48.04	48.60	51.60	58.00	70.02	73.12	67.30	72.32
		IE (ms)	366	240	180	170	178	180	178	**168**	179	272

Table 2 continued

Technique			Energy	FSS	GAD	VLR	VLR + DRF	VLR + DRF + FSS	VLR + DRF + GAD	VLR + DRF + GAD + FSS	VLR + DRF + GAD + FSS + OBS	VLR + DRF + GAD + FSS + OBS + SDK
SNR (dB)	Method	Metrics										
5	Begin	IR (%)	59.00	68.75	**98.03**	95.35	95.71	95.17	97.14	96.81	96.43	95.17
		IE (ms)	376	**296**	440	292	396	419	392	399	407	381
	End	IR (%)	45.83	**72.85**	40.50	52.00	41.40	40.01	53.40	58.31	61.40	67.85
		IE (ms)	376	269	177	168	178	179	178	**159**	179	277
0	Begin	IR (%)	57.00	61.61	95.00	94.10	95.17	95.89	**96.25**	95.78	95.00	94.46
		IE (ms)	376	327	494	**271**	443	433	408	423	458	446
	End	IR (%)	45.75	**72.14**	24.8	53.90	27.90	29.10	34.60	46.28	55.90	58.92
		IE (ms)	366	266	169	**164**	180	180	171	192	181	276
20-0	Average	IR	61.38	78.50	74.71	72.9	72.65	74.98	78.02	82.49	82.09	**84.25**
	Begin and end	IE	333	235	234	**215**	239	244	255	263	266	281

The performance metrics are evaluated on the TD-1 utterance of the RSR2015 database. The testing utterances are corrupted by babble noise, with SNR varying from 20 to 0 dB in steps of 5 dB

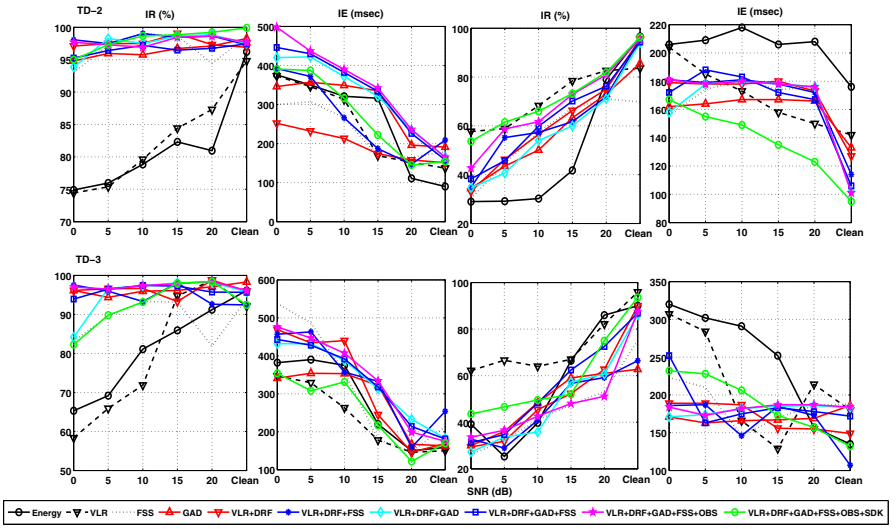

Fig. 9 Performance of the begin and end point detection using four individual methods and their six different combinations are summarized on TD-2 and TD-3 test utterances in terms of the IR and IE on RSR2015 database. The RSR2015 database is corrupted by babble noise, with SNR varying from 20 to 0 dB in steps of 5 dB

of end points, performance is around 10% less than the RSR2015 database. Among different individual algorithms, GAD shows the best performance for both begin and end points. The proposed method using combination of different individual methods shows significantly improved performance for both begin and end points detection. On an average, around 2 and 8% improvement is observed for begin and end points, respectively.

4.3 Analysis of Failure Cases

The end points which are not detected by the proposed method are analyzed. The major failure is due to the impulse-like non-speech background having higher low-frequency energy content. Due to impulse-like nature, both the VLR detection algorithm and GAD capture these regions as speech region. The DRF value computed from the HNGD spectrum is less than 1000 Hz and therefore cannot remove these spurious detections. Figure 10 explains the situation. Figure 10a shows a speech signal with non-overlapping background noise. Figure 10b shows the DRF contour. The non-speech background has DRF values less than 1000 Hz and are not removed by the proposed method.

In a few cases, the speaker is repeating some words without leaving sufficient space in between the repeated words. Speech duration knowledge is failing to remove spurious detection in such cases. The proposed EPD method is a combination of different algorithms, and every individual algorithm has limitations. Sometimes both VLR and GAD algorithms fail to detect some voiced speech regions. Similarly the

Table 3 Performance of begin and end points detection in terms of identification rate and identification error

Technique			Energy	VLR	FSS	GAD	VLR + DRF	VLR + DRF + FSS	VLR + DRF + GAD	VLR + DRF + GAD + FSS	VLR + DRF + GAD + FSS + OBS	VLR + DRF + GAD + FSS + OBS + SDK
SNR	Method	Metrics										
TD1	Begin	IR (%)	94.35	94.93	94.93	97.27	95.32	97.46	93.95	95.00	**97.85**	97.08
		IE (ms)	**37**	62	47	47	61	42	68	59	42	44
	End	IR (%)	87.13	77.78	76.80	88.11	78.55	93.17	91.42	92.00	93.56	**97.07**
		IE (ms)	**75**	126	115	97	123	96	148	86	80	79
TD2	Begin	IR (%)	95.79	96.78	95.06	98.17	95.25	99.08	87.56	89.76	**99.08**	96.16
		IE (ms)	**27**	30	50	48	73	37	83	63	31	30
	End	IR (%)	87.75	86.67	77.69	92.50	79.52	93.60	91.59	93.24	94.15	**96.89**
		IE (ms)	89	**76**	118	88	165	105	169	142	96	95
TD3	Begin	IR (%)	98.36	96.73	96.73	97.75	96.94	98.97	96.53	97.52	99.18	**99.39**
		IE (ms)	**32**	69	50	61	74	45	84	74	42	40
	End	IR (%)	86.73	83.06	81.22	90.61	83.47	92.86	95.51	94.28	94.28	**97.35**
		IE (ms)	**64**	134	89	84	135	79	135	113	68	66
–	Average	IR	91.68	89.33	87.07	94.06	88.18	95.86	92.76	93.63	96.35	**97.32**
	Begin and end	IE	**54**	83	78	71	105	67	115	90	60	59

The performance metrics are evaluated for the TD-1, TD-2 and TD-3 sentences of the IITG database. These testing utterances are collected in the practical environment

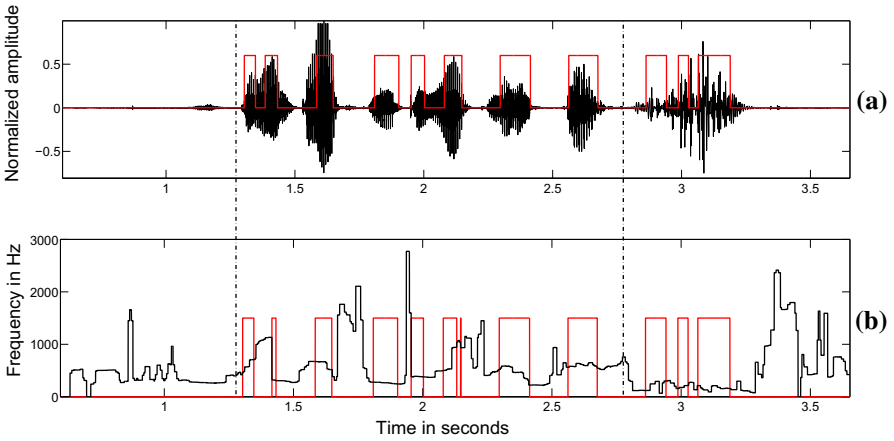

Fig. 10 (**a**) Speech signal with non-overlapping background noise along with hypothesized VLRs. (**b**) DRF contour of the speech along with refined VLRs. Two vertical dotted lines show the begin and end points. Spurious VLRs in non-speech background region are not removed after using the DRF information because the noise has DRF value less than 1000 Hz

sub-fundamental frequency filtering method may fail to detect some obstruent region. Foreground speech segmentation algorithm is threshold based and sometimes fails to detect background speech. Due to the failure in the individual level, the proposed algorithm fails giving a lesser identification rate in such cases.

5 Development of Text-Dependent Speaker Verification

The TDSV systems are usually based on template/model-sequence matching techniques in which the time axes of an input speech utterance and reference models of registered speakers are aligned, and the similarities between them are accumulated from the beginning to the end of utterance [9,10,26,40,55]. When the linguistic content of the utterance is known, speaker verification systems perform better. This is because such systems can better able to model the characteristics of specific phonetic content contained in the speech signal. To make a speaker verification system more suitable for practical applications, the proposed system should be designed such that it should be language, sensor independent and session independent and robust to background degradations.

From the commercial speech-based authentication systems as an application point of view, the enrolled user is free to provide the test speech at the testing phase with no constraints on the quality, recording conditions, duration, channel mismatch and other environmental noises. Accordingly, the performance of speaker verification systems is influenced by many of these possible variabilities. Among these variabilities, lexical content and channel variations are the most detrimental [19]. Compared to channel variability due to uncontrolled environmental factors, lexical variability can be manageable. In case of text-dependent scenario, lexical variability is not present. Matching

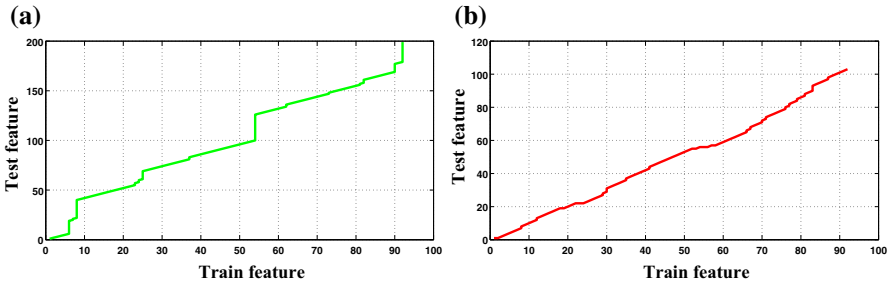

Fig. 11 Evidences for degraded and clean speech utterances for *Don't ask me to walk like that* (**a**) speech utterances containing non-overlapping background noise and background speech and showed DTW warping path deviated away from the regression line, (**b**) speech utterances containing clean speech region after detecting the begin and end points using algorithms and showed DTW warping path along the regression line

of short duration phrases during training and testing with high accuracy makes it an attractive option for commercial speech-based person authentication systems.

The accuracy of a speaker verification system depends on the match between the speaker model and test features [17]. The performance may improve after applying the proposed end point detection algorithm. Figure 11a, b shows the DTW warping path for an utterance before and after using the proposed end point detection algorithm. It can be seen that detection of end points brings the DTW warping path closure to the regression line.

In this work, a TDSV system is developed using 39-dimensional mel-frequency cepstral coefficient (MFCC) features and DTW as a template matching technique. The training phase of speakers includes 3 sessions of 3 different texts. Features of these training templates are extracted and kept as reference templates for further experiments. In the testing phase, the test data are processed in a similar way as the training data and MFCC features of these test data are extracted. DTW algorithm is used for matching the train and test templates.

The dynamic programming algorithm provides a many-to-many mapping between training and testing utterances while minimizing the overall accumulated distance. This minimum accumulated distance is called as dynamic time warping (DTW) distance [39,42]. The testing feature vectors are compared with reference to the claimed model reference template by the DTW algorithm. To find the best match between the train and test feature vectors, dynamic warping path minimizes the total distance between the train and test feature vectors using DTW algorithm. An efficient solution of this minimization is found by the dynamic programming due to its elastic distance measure. DTW algorithm computes the accumulated distance score between the reference model of the train template X and testing template T of different lengths by taking the DTW warping path as,

$$\mathrm{DTW}_{\psi}(X,T) = \sum_{l=1}^{M} \mathrm{DTW}(\psi_x(l), \psi_t(l))s(l)/S_{\psi} \tag{1}$$

where $\mathrm{DTW}_{\psi}(X, T)$ is the accumulated distance score between the test and reference template, $\mathrm{DTW}(\psi_x(l),\psi_t(l))$ is the shortest time spectral distortion, $s(l)$ is a non-negative warping path weighting coefficient and S_{ψ} is the dynamic time warping path normalizing factor. The test speech is matched with the claimed reference model using DTW algorithm as given in Eq. 1 to give the minimum distance score. The decision is taken with respect to a set of four cohort speakers by comparing the distance score against them. The cohort approach is mainly based on a subset of enrolled speaker's models which represents the closest score to the claimed speaker. This cohort selection is made in advance (offline), which consists closest model to the claimed speaker. The cohort-based method is faster than the traditional method of testing the model against the imposters, reducing the unnormalized equal error rate. This is because the utterance must be from a set of pre-defined phrases.

The performance evaluation of the TDSV system is based on two standard metrics—*Equal Error Rate* (EER) and *minimum Detection Cost Function* (mDCF). Let $\mathrm{S}c = \{\mathrm{s}c_1, \mathrm{s}c_2, \ldots, \mathrm{s}c_S\}$ represent the entire set of DTW scores obtained after matching the testing utterances vs. the training utterances. The set Sc is normalized so that $\{\mathrm{s}c_i \in [0, 1] \mid i = 1, 2, \ldots, S\}$. Let $\theta \in [0, 1]$ be the decision threshold, above which the claim of the test utterance against the speaker is accepted. Let S_G denote the number of genuine/true claims that have been accepted, S_I denote the number of imposter/false claims that have been rejected, S_M denote the number of genuine/true claims that have been rejected and S_F denote the number of imposter/false claims that have been accepted. Therefore, $S = S_G + S_I + S_M + S_F$. We may now define EER and DCF as follows:

EER (%) : For any given threshold, θ, the *Miss Rate* is given as $\mathrm{MR} = \frac{S_M}{S_G+S_M} \times 100\ \%$. The *False Alarm Rate* is given as $\mathrm{FAR} = \frac{S_F}{S_I+S_F} \times 100\ \%$. At a particular threshold, $\theta = \theta_0$, MR = FAR. This error is known as the EER. In other words, EER = MR = FAR, at $\theta = \theta_0$.

mDCF : For any given threshold, θ, the *Probability of Miss* is given as $P_M = \frac{S_M}{S_G+S_M}$. The *Probability of False Alarm* is given as $P_F = \frac{S_F}{S_I+S_F}$. Two parameters, C_M and C_F, assign costs to the event of a *miss* (a genuine claim rejected), and that of a *false alarm* (an imposter claim is accepted) respectively. Also, an *a priori* probability P_T is assigned, which assumes that out of all the test claims against a speaker, only a fraction P_T are genuine claims. The cost parameter of the mDCF, at any given θ, is given by,

$$C_{\theta} = C_M \times P_M \times P_T + C_F \times P_F \times (1 - P_T) \tag{2}$$

The mDCF is then given by,

$$\mathrm{mDCF} = \min_{\theta \in [0,1]} C_{\theta} \tag{3}$$

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

6 Experimental Evaluation of Proposed EPD in TDSV System

In this section, we present the experimental results for the TDSV systems implemented separately on the two databases, the RSR2015 and the IITG database. The RSR2015 is a clean database; therefore, to examine the robustness of the proposed begin and end point detection algorithm, the data are degraded by adding babble noise.

6.1 Performance on the Clean RSR2015 Database

The performance of proposed TDSV system is evaluated on the clean RSR2015 database. Table 4 presents the performance metrics of the individual algorithms and their combinations used for the begin and end points detection.

For the TD-1 utterance of the RSR2015 database [49], the energy-based and the VLR-based methods provide an EER of 7.59 and 4.89%, respectively. The proposed method without SDK (VLR + DRF + GAD + FSS + OBS) provides an EER of 4.8%, and the proposed method with SDK (VLR + DRF + GAD + FSS + OBS + SDK) provides an EER of 4.54% which is around 3% improved result compared to the energy-based detection. Among the individual methods, VLR-based method gives the best performance which is comparable to the performance of the proposed method. This result is expected in case of RSR2015 database, because except the channel effect, there is no other type of degradation present in the database. The telephonic handsets used for collecting the training and testing speech are same for maximum number of speakers and the non-speech regions are mostly silence regions. For such type of speech, the begin and end point detection can be performed accurately without much difficulty. Similar performances are observed for the TD-2 and TD-3 utterances as shown in Fig. 12.

6.2 Performances on the RSR2015 Database Under Degraded Condition

The testing utterances of the RSR2015 database are corrupted with babble noise of the NOISEX-92 database. Table 5 presents the performances of different methods for the TD-1 phrase. As can be seen from the table, the performance of TDSV decreases as the SNR increases. Different individual methods such as FSS, VLR and GAD show improved performance compared to the energy-based detection in case of all different levels of degradation. The proposed method gives the best performance among all different combinations. On an average, 4% improvement in EER across five levels of degradation is observed for the proposed method compared to the energy-based method. Experiments are performed for TD-2 and TD-3 phrases, and performances are shown in Fig. 12. As can be seen from the figure, TD-2 and TD-3 phrases show similar trend as TD-1 phrase in terms of EER.

Table 4 Performances of the TDSV systems using different end point detection methods evaluated on the TD-1 phrase of the clean RSR2015 database The TDSV systems is built using 39-dimensional MFCCs and DTW, and performance is evaluated in terms of EER and DCF

Technique		Energy	FSS	GAD	VLR	VLR + DRF	VLR + DRF + FSS	VLR + DRF + GAD	VLR + DRF + GAD + FSS	VLR + DRF + GAD + FSS + OBS	VLR + DRF + GAD + FSS + OBS + SDK
Method	Metric										
Clean	EER (%)	7.59	5.01	5.01	4.89	4.93	5.01	5.00	4.83	4.80	**4.54**
	DCF	0.0388	0.0347	0.0333	0.0343	0.035	0.0352	0.033	0.0332	0.033	**0.0307**

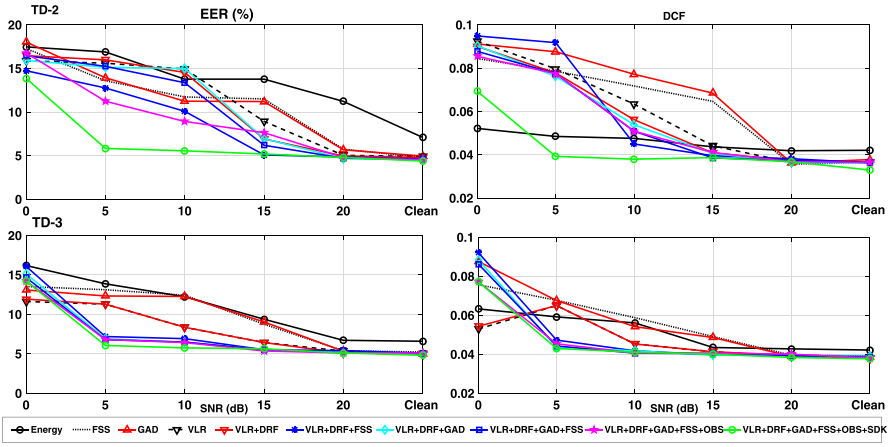

Fig. 12 Performances of the TDSV systems using different end point detection methods evaluated on the TD-2 and TD-3 phrases of the RSR2015 database. The testing utterances are corrupted by babble noise, with SNR varying from 20 to 0 dB in steps of 5 dB. The TDSV systems are built using 39-dimensional MFCCs and DTW, and performance is evaluated in terms of EER and DCF

6.3 Performance on IITG Database

Similar to the RSR2015 database, the performance of the TDSV system is evaluated on IITG database. The challenge in this case is to discriminate a large set of speakers, considering the testing conditions are quite different from the training conditions. Table 6 presents the performance metrics for the three phrases, namely TD-1, TD-2 and TD-3. VLR-, FSS- and GAD-based methods are giving 3–5% improvement in EER over the energy-based method. The EER is further reduced when different individual methods are combined in the proposed framework. Approximately 7% improvement in EER is observed for the proposed method compared to the energy-based method. Robustness of the proposed method in detecting the begin and end points is helping the TDSV system to achieve the improved performance.

7 Summary and Conclusions

This work proposes a begin and end point detection algorithm using some speech-specific information. The proposed algorithm aims at removing non-overlapping speech as well as non-speech background degradations. The method is vowel-like region detection based and uses vocal tract and source-related information to remove background noise. Dominant resonant frequency information is used to remove non-speech background, and foreground speech segmentation is performed to remove the speech background. To detect the consonants at the beginning and at the end of the VLRs, glottal activity and dominant aperiodic regions detection are performed. Further, speech duration knowledge is used to remove some spurious VLRs. The proposed algorithm is used to evaluate a text-dependent SV system. The TDSV system with the

Table 5 Performances of the TDSV systems using different end point detection methods evaluated on the TD-1 phrase of the RSR2015 database under degraded speech condition. The TDSV systems is built using 39-dimensional MFCCs and DTW, and performance is evaluated in terms of EER and DCF

Technique		Energy	FSS	GAD	VLR	VLR + DRF	VLR + DRF + FSS	VLR + DRF + GAD	VLR + DRF + GAD + FSS	VLR + DRF + GAD + FSS + OBS	VLR + DRF + GAD + FSS + OBS + SDK
SNR (dB)	Metrics										
20	EER (%)	8.17	6.78	6.85	4.96	5.42	4.98	5.08	5.37	5.08	**4.78**
	DCF	0.0394	0.0349	0.0351	0.0335	0.0329	**0.0319**	0.0335	0.0337	0.0341	0.0341
15	EER (%)	9.64	9.99	10.11	10.15	10.19	11.09	8.84	7.59	10.42	**4.83**
	DCF	0.0387	0.0546	0.0421	0.041	0.0546	0.0426	0.0381	**0.0341**	0.0428	0.0365
10	EER (%)	11.13	11.73	10.44	11.32	14.13	11.32	11.48	12.6	11.55	**5.22**
	DCF	0.0448	0.0657	0.0474	0.0426	0.0646	0.0575	0.0653	0.0648	0.0648	**0.0382**
5	EER (%)	12.57	13.36	12.45	17.14	16.85	17.14	17.89	17.68	17.68	**9.49**
	DCF	0.0488	0.0672	0.0534	0.0641	0.0645	0.0642	0.0725	0.0669	0.0669	**0.0402**
0	EER (%)	20.29	18.69	19.25	18.10	18.83	18.44	17.94	17.73	17.73	**16.85**
	DCF	**0.0493**	0.0727	0.0667	0.0785	0.0834	0.0642	0.084	0.0948	0.094	0.0649

Table 6 Performances of the TDSV systems using different end point detection methods evaluated on the three phrases of the practical IITG database. The TDSV systems is built using 39-dimensional MFCCs and DTW, and performance is evaluated in terms of EER and DCF

Technique		Energy	FSS	GAD	VLR	VLR + DRF	VLR + DRF + FSS	VLR + DRF + GAD	VLR + DRF + GAD + FSS	VLR + DRF + GAD + FSS + OBS	VLR + DRF + GAD + FSS + OBS + SDK
SNR	Metrics										
TD-1	EER (%)	13.19	8.69	9.50	8.64	7.78	8.10	7.70	7.61	6.96	**6.28**
	DCF	0.0422	0.0365	0.0367	0.0363	0.0355	0.0369	0.037	0.0343	0.0353	**0.0339**
TD-2	EER (%)	11.37	8.20	6.69	7.21	6.72	6.56	6.45	6.33	6.36	**6.20**
	DCF	0.0404	0.04	0.0338	0.0391	0.0352	0.0352	0.0343	0.0324	**0.0316**	0.0338
TD-3	EER (%)	18.42	9.74	9.84	10.61	8.35	10.87	9.38	10.53	9.07	**8.20**
	DCF	0.0526	0.0405	0.0395	0.0423	**0.0379**	0.0429	0.0386	0.0422	0.0426	0.0399

proposed method shows improved performance than the individual methods such as energy-, FSS-, GAD- and VLR-based end point detection.

The proposed end point detection method works when there is an issue of non-overlapping background speech and non-speech noise. In case of overlapping background noise, the temporal and spectral enhancement methods may be useful for robust TDSV system. Future work will be to use temporal and spectral enhancement methods to remove the overlapping background degradation followed by the use of speech-specific knowledge for removal of non-overlapping background degradation.

References

1. F. Bimbot, J.-F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-García, D. Petrovska-Delacrétaz, D.A. Reynolds, A tutorial on text-independent speaker verification. EURASIP J. Adv. Signal Process. **2004**(4), 101962 (2004)
2. S. E. Bou-Ghazale, K. Assaleh, A robust endpoint detection of speech for noisy environments with application to automatic speech recognition. in *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 4, pp. IV-3808 (2002)
3. J.P. Campbell, Speaker recognition: a tutorial. Proc. IEEE **85**(9), 1437–1462 (1997)
4. R.K. Das, S. Jelil, S.R.M. Prasanna, Development of multi-level speech based person authentication system. J. Signal Process. Syst. **88**(3), 259–271 (2017)
5. K.T. Deepak, B.D. Sarma, S.R.M. Prasanna, Foreground speech segmentation using zero frequency filtered signal. in *Interspeech 2012*, Sept (2012)
6. K .T. Deepak, S .R .M. Prasanna, Foreground speech segmentation and enhancement using glottal closure instants and mel cepstral coefficients. IEEE/ACM Trans. Acoust. Speech Lang. Process. **24**, 1204–1218 (2016)
7. S. Dey, S. Barman, R. K. Bhukya, R. K. Das, B. Haris, S. R. M. Prasanna, R. Sinha, Speech biometric based attendance system, in *2014 Twentieth National Conference on Communications (NCC)*, IEEE, pp. 1–6 (2014)
8. N. Dhananjaya, B. Yegnanarayana, Voiced/nonvoiced detection based on robustness of voiced epochs. IEEE Signal Process. Lett. **17**, 273–277 (2010)
9. T. Dutta, Dynamic time warping based approach to text-dependent speaker identification using spectrograms, in *Congress on Image and Signal Processing, CISP'08*, vol. 2. IEEE, pp. 354–360 (2008)
10. S. Furui, Cepstral analysis technique for automatic speaker verification. IEEE Trans. Acoust. Speech Signal Process. **29**, 254–272 (1981)
11. J. González-Rodríguez, J. Ortega-García, C. Martín, L. Hernández, Increasing robustness in GMM speaker recognition systems for noisy and reverberant speech with low complexity microphone arrays, in *Proceedings of Fourth International Conference on Spoken Language, 1996, ICSLP 96*, vol. 3, IEEE, pp. 1333–1336 (1996)
12. D. N. Gowda, in *Signal Processing for Excitation-Based Analysis of Acoustic Events in Speech*. Ph.D. Dissertation, Department of Computer Science and Engineering, IIT Madras (2011)
13. M. Hamada, Y. Takizawa, T. Norimatsu, A noise robust speech recognition system, in *The International Conference on Spoken Language Processing* (1990)
14. V. Hautamäki, M. Tuononen, T. Niemi-Laitinen, P. Fränti, Improving speaker verification by periodicity based voice activity detection, in *Proceedings of 12th International Conference on Speech and Computer (SPECOM2007)*, vol. 2, pp. 645–650 (2007)
15. M. Hébert, Text-dependent speaker recognition, in *Springer Handbook of Speech Processing*, Springer, pp. 743–762 (2008)
16. B. K. Khonglah, R. K. Bhukya, S. R. M. Prasanna, Processing degraded speech for text dependent speaker verification, in *International Journal of Speech Technology*, pp. 1–12 (2017)
17. T. Kinnunen, H. Li, An overview of text-independent speaker recognition: from features to supervectors. Speech Commun. **52**(1), 12–40 (2010)
18. L. Lamel, L. Rabiner, A. Rosenberg, J. Wilpon, An improved endpoint detector for isolated word recognition. IEEE Trans. Acoust. Speech Signal Process. **29**(4), 777–785 (1981)

19. A. Larcher, K.A. Lee, B. Ma, H. Li, Text-dependent speaker verification: classifiers, databases and RSR2015. Speech Commun. **60**, 56–77 (2014)
20. 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)
21. 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 *2016 Twenty Second National Conference on Communication (NCC)*, IEEE, pp. 1–6 (2016)
22. J. Makhoul, Linear prediction: a tutorial review. Proc. IEEE **63**(4), 561–580 (1975)
23. 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)
24. K .S .R. Murthy, B. Yegnanarayana, Epoch extraction from speech signals. IEEE Trans. Audio, Speech, Lang. Process. **16**(8), 1602–1613 (2008)
25. K.S.R. Murthy, B. Yegnanarayana, M.A. Joseph, Characterization of glottal activity from speech signals. IEEE Signal Process. Lett. **16**(6), 469–472 (2009)
26. R. Piyare, M. Tazil, Bluetooth based home automation system using cell phone, in *IEEE 15th International Symposium on Consumer Electronics (ISCE)*, IEEE , pp. 192–195 (2011)
27. 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)
28. 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)
29. G. Pradhan, Speaker verification under degraded conditions using vowel-like and nonvowel-like regions, Ph.D. Dissertation (2013)
30. G. Pradhan, S.R.M. Prasanna, Speaker verification under degraded condition: a perceptual study. Int. J. Speech Technol. (Springer) **14**(4), 405–417 (2011)
31. R. S. Prasad, B. Yegnanarayana, Acoustic segmentation of speech using zero time littering, in *Proceedings of INTERSPEECH*, pp. 2292–2296 Aug (2013)
32. S. R. M. Prasanna, B. Yegnanarayana, Detection of vowel onset point events using excitation source information, in *Proceedings of INTERSPEECH*, pp. 1133–1136, Sept (2005)
33. S.R.M. Prasanna, J.M. Zachariah, B. Yegnanarayana, Begin-end detection using vowel onset points, in *Workshop on Spoken Language Processing*, TIFR, Mumbai, India, Jan (2003)
34. S.R.M. Prasanna, B.V.S. Reddy, P. Krishnamoorthy, Vowel onset point detection using source, spectral peaks, and modulation spectrum energies. IEEE Trans. Audio Speech Lang. Process **17**(4), 556–565 (2009)
35. 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)
36. S. R. M. Prasanna, J. M. Zachariah, B. Yegnanarayana, Begin-end detection using vowel onset points, in *Workshop on Spoken Language Processing* (2003)
37. S. R. M. Prasanna, Event-based analysis of speech, Ph.D. Dissertation, Department of Computer Science and Engineering, IIT Madras (2004)
38. L.R. Rabiner, B.H. Juang, *Fundamentals of Speech Recognition* (Prentice-Hall, Upper Saddle River, 1993)
39. L.R. Rabiner, A.E. Rosenberg, S.E. Levinson, Considerations in dynamic time warping algorithms for discrete word recognition. J. Acoust. Soc. Am. **63**(S1), S79–S79 (1978)
40. K. Ramesh, S. R. M. Prasanna, R. K. Das, Significance of glottal activity detection and glottal signature for text dependent speaker verification, in *International Conference on Signal Processing and Communications (SPCOM), 2014*, IEEE, pp. 1–5 (2014)
41. G. Saha, S. Chakroborty, S. Senapati, A new silence removal and endpoint detection algorithm for speech and speaker recognition applications, in *Proceedings of the 11th national conference on communications (NCC)*, pp. 291–295 (2005)
42. H. Sakoe, S. Chiba, Dynamic programming algorithm optimization for spoken word recognition. IEEE Trans. Acoustics Speech Signal Process. **26**(1), 43–49 (1978)
43. B. D. Sarma, S. R. M. Prasanna, Analysis of spurious vowel-like regions detected by excitation source information, in *Indicon* (2013)
44. 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)
45. B. D. Sarma, P. S. Supreeth, S. R. M. Prasanna, Improved vowel onset and offset points detection using Bessel features, in *SPCOM* (2014)

46. M.H. Savoji, A robust algorithm for accurate endpointing of speech. Speech Commun. **8**, 45–60 (1989)
47. C.S.P. Secries, in *Time-Frequency Analysis: Theory and Applications, Series: Signal Processing Series* (Englewood Cliffs: Prentice-Hall, 1995)
48. R. Sharma, S.R.M. Prasanna, A better decomposition of speech obtained using modified empirical mode decomposition. Digit. Signal Process. **58**, 26–39 (2016)
49. 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)
50. J. Sohn, N.S. Kim, W. Sung, A statistical model-based voice activity detection. IEEE Signal Process. Lett. **6**(1), 1–3 (1999)
51. C. Tsao, R.M. Gray, An endpoint detection for LPC speech using residual look-ahead for vector quantization applications, in *IEEE International Conference on Acoustics, Speech, and Signal Processing* (Springer, Berlin, 1984), p. 1
52. 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)
53. L. P. Wong, M. Russell, Text-dependent speaker verification under noisy conditions using parallel model combination, in *Proceedings of (ICASSP'01). 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2001*, vol. 1, IEEE, pp. 457–460 (2001)
54. 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. Acoust. Speech Signal Process. **13**, 575–582 (2005)
55. 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)
56. B. Yegnanarayana, D.N. Gowda, Spectro-temporal analysis of speech signals using zero-time windowing and group delay function. Speech Commun. **55**, 782–795 (2013)