International Journal of Speech Technology (2018) 21:279–291
https://doi.org/10.1007/s10772-018-9505-x

Application of non-negative frequency-weighted energy operator for vowel region detection

Ramakrishna Thirumuru[1] · Anil Kumar Vuppala[1]

Received: 22 November 2017 / Accepted: 29 March 2018 / Published online: 10 April 2018

Abstract

In this paper, a novel technique has been proposed for the vowel region detection from the continuous speech using an envelope of the derivative of the speech signal, which is a non-negative, frequency-weighted energy operator. The proposed vowel region detection method is implemented using a two-stage algorithm. The first stage of vowel region detection consists of speech signal analysis to detect vowel onset points (VOP) and vowel end-points (VEP) using an instantaneous energy contour obtained from the envelope of the derivative of a speech signal. The VOPs and VEPs are spotted using the peak-finding algorithm based upon the first order Gaussian differentiator. The next stage consists of removal of spurious vowel regions and the correction of hypothesized VOP and VEP locations using combined cues obtained from the uniformity of epoch intervals and strength of the excitation of the speech signal. Performance of the proposed method for detecting vowel regions from the speech signal is evaluated using TIMIT acoustic-phonetic speech corpus. The proposed approach resulted in significantly high detection rate and less false alarm rate compared to the state-of-the-art methods in both clean and noisy environments.

Keywords Vowel onset point · Vowel end-point · Instantaneous energy contour · Envelope-derivative of the speech signal · Uniformity of epoch intervals · Strength of the excitation

1 Introduction

Vowels are primary units of the sound system of a language (Johnson 2004). These are produced by periodic impulse like excitation and possess high energy, periodicity and longer duration (Deller et al. 1993; Fant 1971; Stevens 2000). Vowel region detection is a task of identifying vowel occurrences with precise boundary markings. These boundary markings are termed as vowel onset point (VOP) and vowel end-point (VEP). VOP is the time instant at which vowel region begins and VEP can be considered as the time instant at which vowel region ends in a continuous speech. The vowel regions detection is an important step in many speech processing applications. These include automatic speech recognition (ASR), speaker verification, smart audio filtering, recognition of CV units for emotion conversion, determining the duration of vowels in forensic applications, speech rate manipulation in speech synthesis, cochlear implants, and multimedia synchronization (Hermes 1990; Donaldson et al. 2013; Juneja and Espy-Wilson 2008; Rao and Yegnanarayana 2009; Pradhan and Prasanna 2013; Prasanna and Pradhan 2011; Rose 2003).

Modern digital signal processing algorithms have been used for detecting acoustic events such as fricative, voice onset time for stops, VOPs and VEPs in a continuous speech (Vydana and Vuppala 2016; Dumpala et al. 2016; Saha et al. 2016; Hansen et al. 2010). The knowledge of crucial acoustic events such as VOP and VEP of the speech can be integrated into the automatic statistical speech processing systems to improve the overall performance. Using these landmarks as speech characteristic cues, speech systems are developed to process externally degraded speech (Salomon et al. 2004). The landmark detection is used as a front-end framework of speech systems that can supplement existing statistical based speech processing systems (Schutte and Glass 2005; Glass 2003). In the literature, the problem of identifying vowel regions under the context of landmark detection is studied through

✉ Ramakrishna Thirumuru
ramakrishna.thirumuru@research.iiit.ac.in

Anil Kumar Vuppala
anil.vuppala@iiit.ac.in

[1] Speech Processing Lab, KCIS, International Institute of Information Technology, Hyderabad (IIIT-H), Hyderabad, India

the extraction of distinctive features (Liu 1996). The vowel regions were marked using VOPs derived from excitation source information of speech signal (Prasanna and Yegnanarayana 2005). VOP detection methods are based on rising slope of spectral amplitude in the magnitude spectrum of the speech signal. An alternative method is proposed by combining evidence from the excitation source, spectral peaks and modulation spectrum for the detection of VOP (Prasanna et al. 2009). An improved VOP detection for vowel region extraction is proposed based on spectral energy present in glottal closure regions of speech signal (Vuppala et al. 2012; Vuppala and Rao 2013). The VOPs were corrected using uniformity of epoch intervals (Vuppala et al. 2012). In another attempt, a method was proposed to detect VEPs based on falling and rising of the slope of spectral energy (Yadav and Rao 2013). The region between VOP and VEP is considered as a vowel region. Recently a technique was proposed on improvements in detection of VOP and VEP using three-class classifier with front-end feature extraction technique (Kumar et al. 2017). This approach exploits spectral and temporal characteristics of the excitation source information of the speech signal. A vowel detection method is proposed based on temporal objective contour generated from the speech signal and spectrally processed to obtain vowel landmarks (Kashani et al. 2017).

Inspired by the previous works related to the VOP and VEP detection and treating a vowel region in between the VOP and VEP, an alternative method is presented for the detection of vowel regions using an envelope of the derivative of the speech signal. The proposed method is carried out in two stages. In the first stage, the landmarks such as VOP and VEPs are detected with a known characteristic that vowels exhibit high sonority and loudness. The landmarks are obtained from the envelope of the derivative of the speech signal. This technique produces optimum instantaneous energy contour with a good temporal resolution to localize the occurrence of the acoustic events. This energy measure includes both amplitude and frequency information of the speech signal, so that landmarks are detected based on spectral content intensity variation around VOPs and VEPs. In the second stage, two cues namely uniformity of epochs and strength of the excitation (SoE) of the speech signal are used to eliminate spurious vowel regions along with the correction of onset point and end-point locations of the vowels.

Rest of the paper is organized in the following manner: Sect. 2 describes the baseline methods for the vowel region detection. In Sect. 3, an envelope of the derivative of the signal and its properties are discussed. The proposed method for vowel region detection is described in Sect. 4. The performance of the proposed method evaluated using TIMIT acoustic-phonetic speech corpus is discussed in Sect. 5. Section 6 describes the summary and conclusions of this work.

2 Baseline methods

Significant research has been carried out on detection of VOPs and VEPs from the continuous speech and a few studies are targeted towards detecting vowel regions. Among these, two state-of-the-art methods for vowel region detection schemes have been considered for comparing the performance of the proposed method. The recent works (Kumar et al. 2017; Kashani et al. 2017) with a range of performance scores for vowel region detection on TIMIT speech corpus have been used for the evaluation of the proposed method. These methods are referred as method I and method II in this paper. In method I, vowel region detection was carried using different acoustic modeling approaches using the combination of mel-frequency cepstral coefficients and excitation source features. Among these approaches, it is reported that the subspace GMM–HMM with discriminative training using boosted maximum mutual information produced superior performance. In method II, vowels are detected using a perceptually-enhanced spectrum matching. It explores a new model based on proposed components called matched filters. Matched filters are extracted by applying a series of perceptually-based processing operations to the speech spectra of the voiced frames. MFs are subjected to different factors leading to the variation in the speech spectra. An acoustic space representing two effective factors, namely phonetic context and speaker identity is modeled. Then, vowel and consonant MFs are conditioned to this context-speaker acoustic space.

In addition to these methods, the state-of-the-art techniques for the detection of VOPs and VEPs have been selected to formulate vowel region detection methods. Therefore, these methods are also considered as a baseline methods of this work. The experimental results obtained by the proposed method were used to compare with the baseline methods based on vowel region overlap criteria with the ground truth. In this paper, the formulated baseline methods are referred as COMB method and FGCI method respectively. In this regard, two VOP and a VEP detection techniques (Prasanna et al. 2009; Vuppala and Rao 2013; Yadav and Rao 2013) were selected to implement these baseline vowel region detection methods. In COMB method, vowel regions are detected from the cues obtained using the combination of the excitation source, spectral peaks, and modulation spectrum of the speech signal. The spectral energy contour around the glottal closure instants (GCIs) is used as an evidence for detecting VOPs and VEPs in FGCI method. These methods have been described in the following subsections.

2.1 Vowel region detection using COMB method

In this method, the combined evidence for the detection of boundary markings of a vowel region is derived from the three shreds of evidence derived from the excitation source, spectral peaks and modulation spectrum. Speech is produced by the excitation of the time-varying vocal tract system with a quasi-periodic signal. The time-varying excitation information is context-dependent in terms of voicing, level of voiced energy, and associated periodicity. Linear prediction (LP) residual corresponds to the excitation source information useful in voice analysis of a speech signal. It is extracted using LP analysis (Makhoul 1975). The time-varying dynamics in the excitation characteristics are overspread in the LP residual due to its bipolar nature (Ananthapadmanabha and Yegnanarayana 1979). So, Hilbert envelope of LP residual is estimated, which is unipolar. The smoothened Hilbert envelope of the LP residual is obtained by convolving with Hamming window of 50 ms. This evidence is considered for the VOP and VEP detection and enhanced using first-order difference operator. These acoustic events are detected based on the nature of the gradient of the output signal.

This method also seeks to characterize the periodic and high-intensity variation of acoustic amplitude as a function of time during vowel production in a speech signal. A 256-point DFT of the speech signal with 20 ms duration with 50% overlap produces amplitude spectrum. The sum of ten largest spectral peaks is selected from the first 128 points and plotted as a function of a time. It represents the spectral energy contour of the speech signal. The reason for selecting ten largest peaks is that, they characterize gross level information of the vocal-tract shape (Prasanna et al. 2009). The VOP and VEP can be observed as a significant changes in a complementary manner in this time-varying signal. The locations of the VOPs and VEPs are enhanced in the spectral energy signal using first order difference. Thus obtained spectral energy is used to supplement the first evidence in the process.

A slowly varying temporal envelope of a speech signal can be represented by using amplitude modulation spectrum or simply modulation spectrum. The modulation spectrum of speech is dominated by the low-frequency components. VOP and VEP detection using modulation spectrum energy is carried out in the following sequence of steps. The speech signal is passed through approximately 18 trapezoidal critical bandpass filters between 0 and 4 kHz. An amplitude envelope of the signal is computed using half wave rectification and low pass filtering on all bands. Amplitude envelope signals are down-sampled to 80 and normalized by the average envelope of that channel, measured over entire utterance. The modulations of the normalized envelope signals are analyzed by computing DFT over 250 ms with an overlap of 5% in order to capture dynamic properties of the signal. The 4–16 Hz components are added together across all critical bands to derive modulation spectrum energy. Thus obtained signal is enhanced and processed to obtain third evidence for detecting VOPs and VEPs. The combined method uses three independent and complementary evidence to derive a single combined evidence by adding three shreds of evidence sample by sample. The combined evidence is convolved with first-order Gaussian difference operator having 100 ms length and 25 ms standard deviation. Spurious peaks are eliminated by thresholding. The VOPs and VEPs are marked at peaks and valleys of the convolved output (Prasanna et al. 2009; Yadav and Rao 2013). The regions between a VOP and a VEP is hypothesized as a vowel region. Experimental result for the COMB method is shown in Fig. 1. From the Fig. 1f, it is noted that spurious vowel regions are detected producing higher false alarm (FA) rate in this method.

2.2 Vowel region detection using FGCI method

The time instants at which glottal signal produce high energy during the production of voiced speech are referred as GCIs. The vocal tract is completely isolated from trachea and lungs during glottal closure phase. Spectrum estimation during glottal closure phase will be more accurate as true vocal tract resonances are present during this period. The spectral energy is computed around the GCIs has been used as an evidence to detect VOP and VEP. Firstly, the GCIs are detected using zero frequency filtering (ZFF) technique (Yegnanarayana and Murty 2009). Around these GCIs, formants are computed for 30% of speech samples using group delay function. This formant energy of speech signal is computed as the sum of first three formants, and it is plotted as a function of time. This formant energy contour is smoothed using mean smoothing window of 50 ms and enhanced using first-order difference operator. Significant changes in the spectral characteristics, present in the enhanced signal are detected by convolving the same with first-order Gaussian differentiator operator having 100 ms length and 25 ms standard deviation. After eliminating the spurious peaks, positive and negative peaks of this signal represent locations of VOP and VEP respectively (Vuppala and Rao 2013; Yadav and Rao 2013). The region between VOP and VEP is considered as a vowel region. Figure 2 demonstrates the experimental result for a test utterance using FGCI method. Figure 2a–f refers to a continuous speech utterance, sum of the first three formant peaks, mean smoothed evidence contour, enhanced evidence using first-order difference operator, VOP and VEP markings using FOGD operator and prediction respectively. From the Fig. 2, it is noted that the result produced in this method is closely comparable with the COMB method.

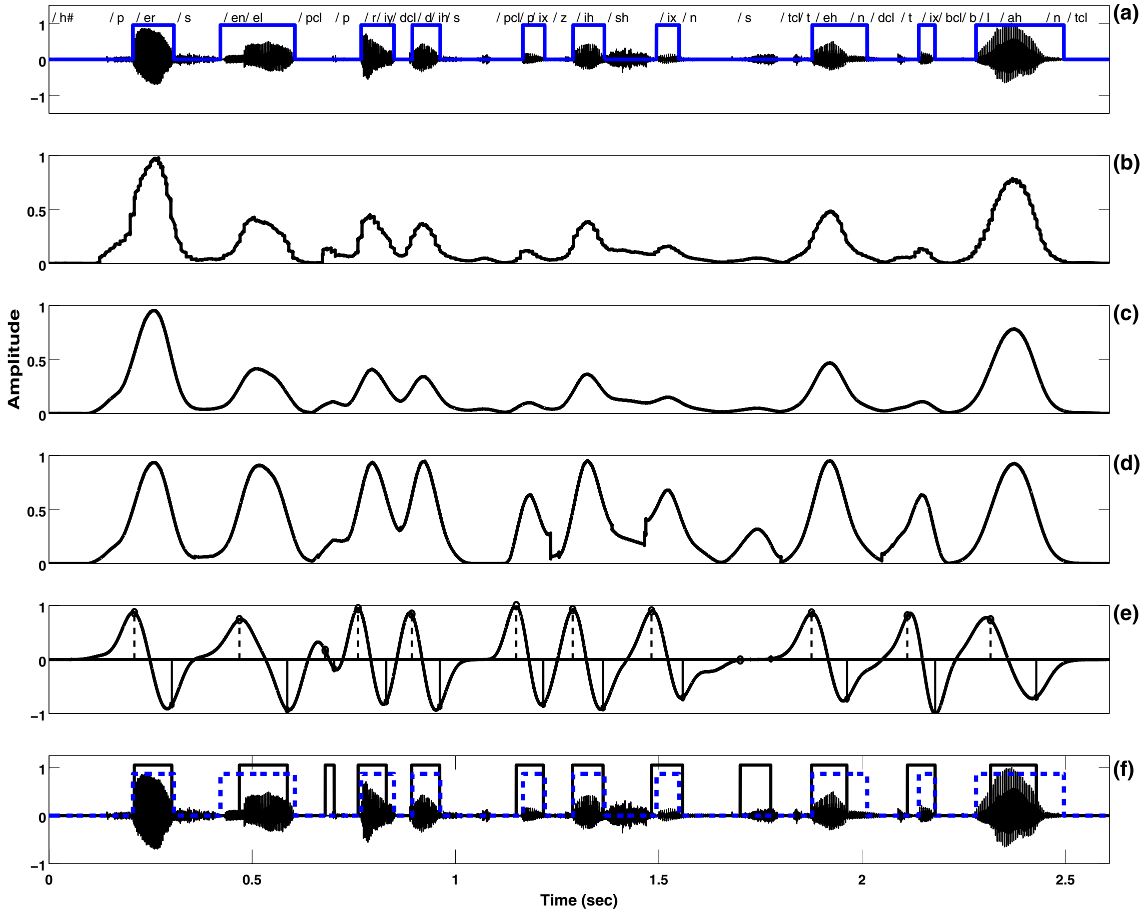

Fig. 1 Detection of vowel region using COMB method for a speech utterance /"Personal predispositions tend to blunt"/. **a** Speech signal with vowel boundaries marked as per TIMIT acoustic-phonetic speech corpus. **b** Combined evidence. **c** Mean smoothed evidence contour. **d** Enhanced evidence using first order difference operator. **e** Hypothesized VOPs and VEPs for the speech signal. **f** Prediction and the speech signal with ground truth (dashed line)

3 Non-negative frequency-weighted energy measure

Most of the natural signals change with time. Speech is also natural signal and the source of changes in speech signal is due to the change in the shape of a vocal tract that enhances or attenuates individual spectral components. The most important phonetic information is embedded in these changes and it reflects in the instantaneous energy variation of the speech signal. The energy in a speech signal is the average of the sum of the squares of the magnitude of the speech signal either in a time domain or frequency domain. It is given by $|x(t)|^2$ or the envelope of the signal as:

$$S[x(t)] = |x(t) + jH[x(t)]|^2 \tag{1}$$

where $S[.]$ is the envelope operator and $H[.]$ is the Hilbert transform operator. It produces energy as a function of amplitude of the signal. From the physics perspective, it has understood that the system requires more energy to generate a high-frequency signal than a low-frequency signal with the same amplitude. In this context, Kaiser proposed an energy measure based on Teagers work which includes not only amplitude but also the frequency of the signal (Teager and Teager 1990; Kaiser 1993). This non-linear energy measure has been referred as Teager–Kaiser energy operator. It is an instantaneous energy measure that differs from the signal processing perspective. The Teager–Kaiser energy measure for the continuous signal $x(t)$ is defined as a second order differential equation:

$$\psi[x(t)] = \dot{x}^2(t) - x(t)\ddot{x}(t) \tag{2}$$

where $\dot{x}(t) = dx(t)/dt$ and $\ddot{x}(t) = d^2x(t)/dt^2$. The discrete counter part of Teager–Kaiser energy operator is given by:

$$\psi[x[n]] = x^2[n] - x[n-1]x[n+1] \tag{3}$$

It is noted that this instantaneous energy measure is estimated from the three current samples and depends on both amplitude and frequency of the speech signal. Typical value

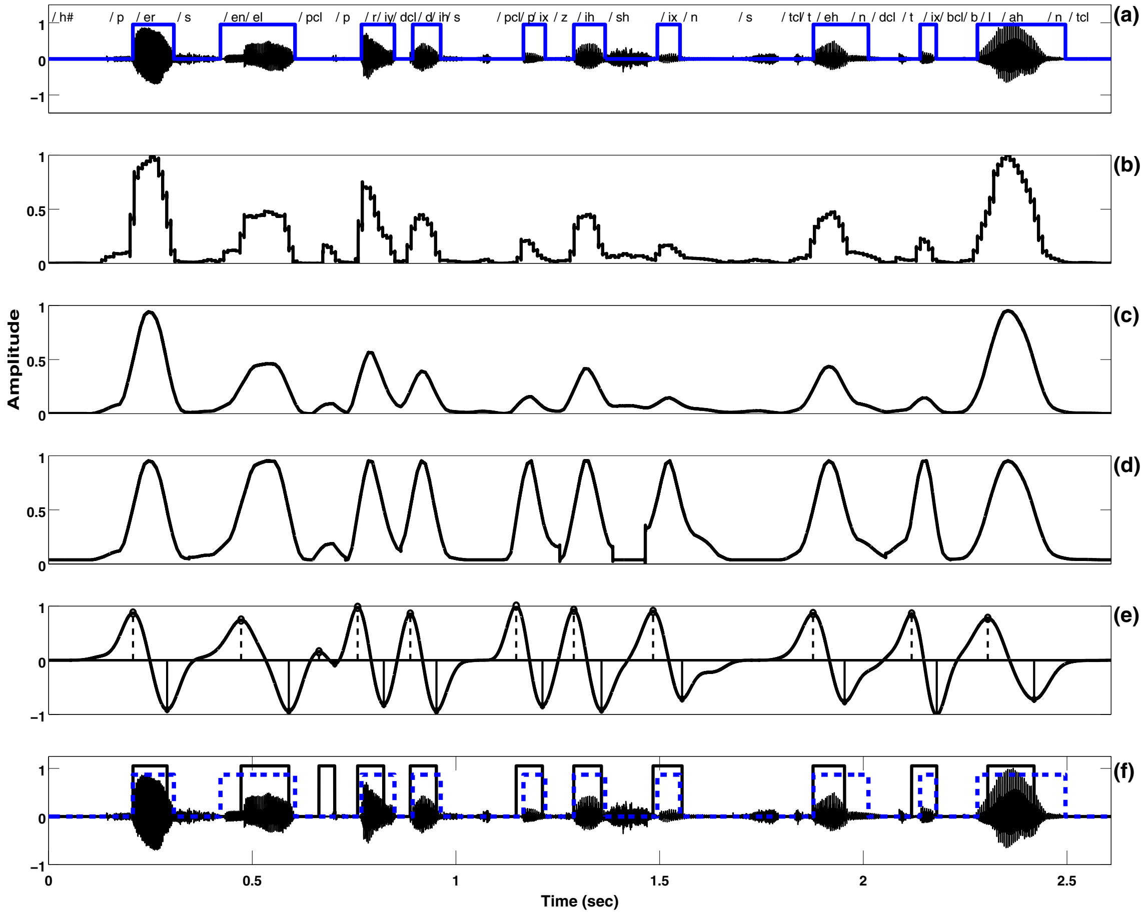

Fig. 2 Detection of vowel region using FGCI method for a speech utterance /"Personal predispositions tend to blunt"/. **a** Speech signal with vowel boundaries marked as per TIMIT acoustic-phonetic speech corpus. **b** Formant energy contour. **c** Mean smoothed formant energy contour. **d** Enhanced evidence using first order difference operator. **e** Hypothesized VOPs and VEPs for the speech signal. **f** Prediction and the speech signal with ground truth (dashed line)

of $S\left[Acos\left(\omega_0 t\right)\right]$ is equal to A^2. A non-negative frequency-weighted energy operator was used to assess instantaneous energy in EEG signal (O'Toole et al. 2014). Inspired by the success of non-negative, frequency-weighted energy operator to compute instantaneous energy in biomedical signal processing, and a similar idea is used in the context of speech signal processing (Kaiser 1990; Palmu et al. 2010). This energy measure is proposed based on the derivative of envelope of the signal, which includes frequency information. It is formulated by applying a weighting filter with frequency response $|H(\omega)|^2 = \omega^2$ to the signal. The derivative function is selected as a filter to maintain similarity with Teager–Kaiser operator and this operator is defined as

$$\Gamma[x(t)] = |\dot{x}(t) + jH[\dot{x}(t)]|^2 = \dot{x}^2(t) + H[\dot{x}(t)]^2 \tag{4}$$

This energy operator is termed as an envelope of the derivative of a signal (O'Toole et al. 2014). It satisfies all important properties of Teager–Kaiser operator and it differs to Teager–Kaiser operator, which includes additional modulation term. In addition, this energy measure does not create negative values in multi-component signals like speech signal and exhibits the non-negative property. For comparison, the instantaneous energy contours of Teager–Kaiser operator and an envelope-derivative operator are shown in the Fig. 3.

4 Proposed method for the detection of vowel regions

An alternative method has been proposed to detect vowel regions from the speech signal using a non-negative, frequency-weighted energy operator. It is based on the energy transitions in the instantaneous energy contour of the speech signal. The motivation for the proposed vowel region detection is that the levels of energy in speech signal is distributed across a range of frequencies and change with time, A signal processing tool to track the dynamic energy transitions can

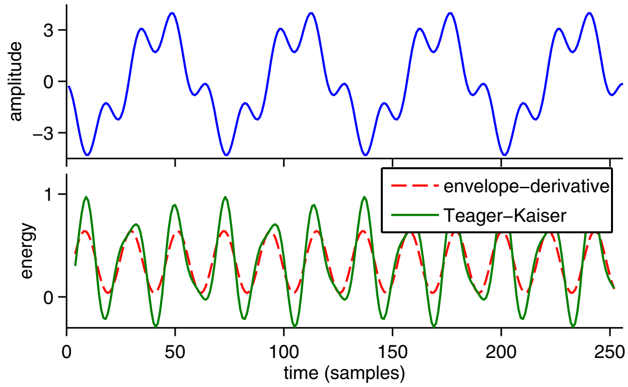

Fig. 3 Comparative analysis of Teager–Kaiser energy measure and energy measure using derivative of envelope of the signal

be used as cues to detect landmarks using frequency dependent non-negative energy operator. In spectral energy based landmark detection methods, a spectrum for the speech signal is estimated using a frame size of 20 ms with a frame shift of 10 ms. The spectral energy around the regions of GCIs is computed and used as an evidence in other methods. In these GCI synchronous methods, the speech segment considered for estimating spectrum is 30 % of the pitch period with an assumption that speech signal during glottal closure phase has a high signal to noise ratio compared to the other regions. The non-negative, frequency-weighted energy operator serves as a tool, which produces instantaneous energy contour of the speech signal eliminating the block processing of the speech signal. It produces better time localization pertaining to the energy contour. The sharp rise and fall of energies around GCIs can be visualized as VOPs and VEPs. This vowel region detection method has been implemented in two stages. In the first stage, onset and end-points of the vowel are detected using the instantaneous energy contour of the speech signal. In the second stage, the positions of VOPs and VEPs have been corrected along with the removal of spurious vowel regions. This is carried out based on the uniformity of epochs and the SoE profile. These regions can be considered as linguistically relevant information possessing regions, that can be used in the front-end of the automatic speaker recognition system.

4.1 Stage 1 : VOP and VEP detection using non-negative frequency dependent energy measure

As a part of the first step in vowel region detection, the VOPs and VEPs are detected from the continuous speech signal in the following manner: an envelope of the derivative of the signal, which is non-negative frequency dependent energy operator is applied to the speech signal $x[n]$, to produce instantaneous energy contour. It is computed from the discrete counterpart of the Eq. 4 and given by:

$$y[n] = \Gamma[x[n]] = \frac{1}{4}\left[x^2[n+1] + x^2[n-1] + h^2[n+1] + h^2[n-1]\right] + \frac{1}{2}[x[n+1]x[n-1] + h[n+1]h[n-1]] \tag{5}$$

The Hilbert transform of the signal $x[n]$ is denoted as $h[n] = H[x[n]]$. It is defined as $IDFT\{-jsgn[N/2-k]sgn[k]X[k]\}$, and $X[k] = DFT\{x[n]\}$, where N is length of the speech signal, DFT is discrete Fourier transform and $IDFT$ is the inverse DFT. This energy operator is nearly instantaneous in discrete time as energy computation of a speech signal is done with three samples at each time instant. It provides good time resolution to capture energy fluctuations of a speech signal within a glottal cycle. The fluctuations produced in the energy contour are smoothed by using mean smoothing with 50 ms window. The change at the acoustic landmarks present in the smoothed instantaneous energy contour of the speech signal is enhanced by computing it’s slope using the first order difference of the resulting signal is given by:

$$y_d[n] = y'[n] - y'[n-1] \tag{6}$$

where $y'[n]$ is the mean smoothed instantaneous energy contour. The regions associated with zero crossing points from both positive to negative and negative to positive are enhanced by normalization process given by:

$$y_N[n] = \frac{y_d[n] - min}{max - min} \tag{7}$$

where $y_N[n]$ is the normalized value of the $y_d[n]$. The min and max are the local minimum and maximum respectively. The predominant energy changes present in the enhanced instantaneous energy contour associated to the vowel landmarks are detected by convolving with the first order Gaussian differentiator operator of 100 ms. A Gaussian window $g[n]$ of length L is given by:

$$g[n] = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{n^2}{2\sigma^2}},\ n = 1, 2, 3 \ldots L \tag{8}$$

where σ is the standard deviation and L corresponds to the length of the Gaussian window. The first order Gaussian window is termed as $g_d[n]$ and given by:

$$g_d[n] = g[n] - g[n-1] \tag{9}$$

First order Gaussian differentiator provides a mechanism to compute slope at each sample. Considering the enhanced energy contour $y_N[n]$ as weighted sum of unit sample sequence and given by:

$$y_N[n] = \sum_{k=1}^{N} a_k \delta[n-k] \tag{10}$$

Here a_k corresponds to the weights and $\delta[n]$ is the unit sample sequence. The convolved output of the enhanced energy contour $y_N[n]$ and FOGD operator $g_d[n]$ is $p[n]$ given by:

$$p[n] = y_N[n] * g_d[n] = \sum_{k=1}^{N} a_k \cdot g_d[n] \cdot \delta[n-k] \tag{11}$$

It produces zero at the output for the constant slope region, a positive peak to the energy transition on the rising note, a negative peak to the energy transition on the falling note respectively. The positive peaks and negative peaks from the evidence plot correspond to the VOP and VEP locations. The region between a VOP and VEP is considered as a vowel region. The result for vowel region detection using envelope of the derivative of a speech signal shown in Fig. 4. A speech utterance /"Personal predispositions tend to blunt"/ with vowel boundaries marked as per TIMIT acoustic-phonetic speech corpus is shown in Fig. 4a. In this work, ground truth containing a cluster of vowels that appear in a sequence are treated as a single unit. The envelope of the derivative of the speech signal, mean smoothed energy contour and the enhanced energy contour are shown in Fig. 4b, c and d respectively. The output signal obtained by convolving energy contour with the FOGD operator and the detected vowel boundaries are given in Fig. 4e. The predicted vowel regions using the proposed method and the ground truth (dashed line) are shown in Fig. 4f. The vowel regions detected by the proposed method (first stage) are not completely in-line with the vowel regions marked in the TIMIT acoustic-phonetic speech corpus. From the Fig. 4f, it is noted that the proposed method is superior to the baseline methods in terms of spurious vowel detection.

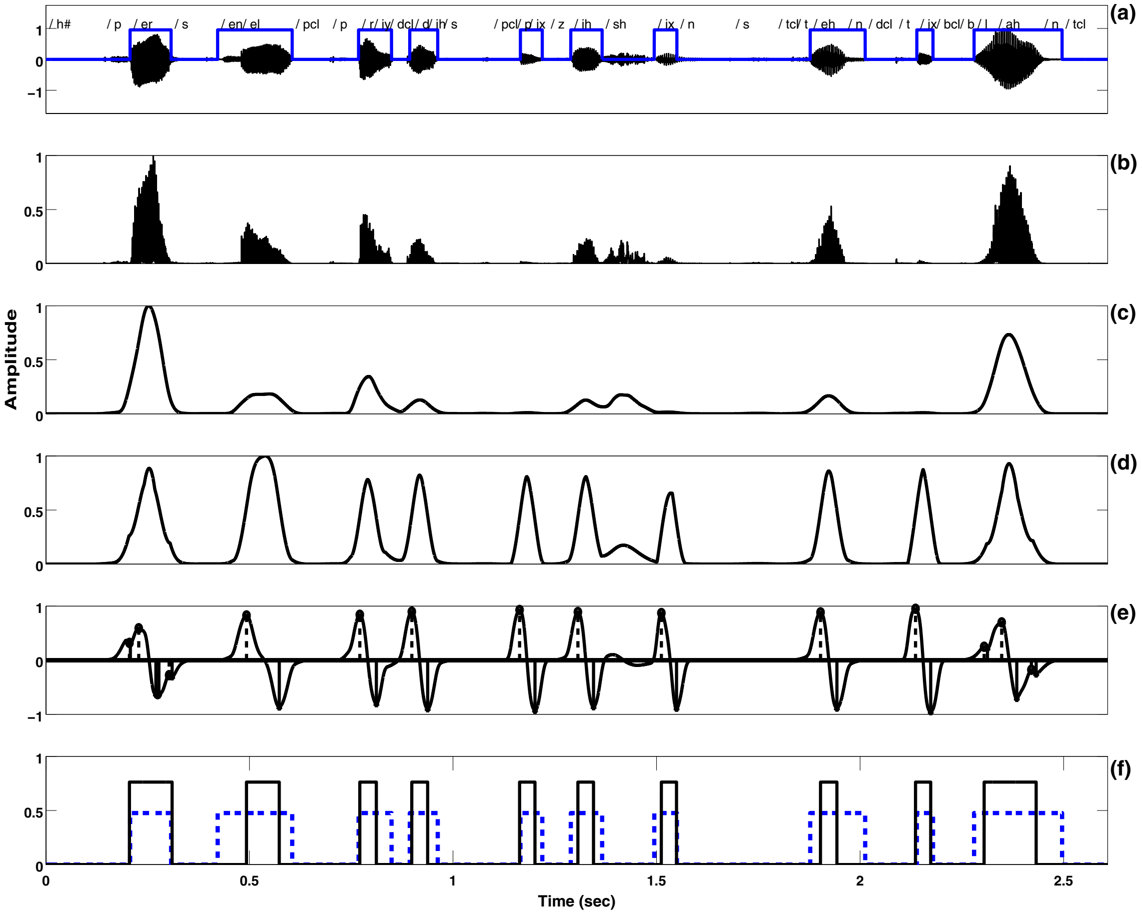

Fig. 4 Detection of vowel region using the proposed method (stage I) for a speech utterance /"Personal predispositions tend to blunt"/. **a** Speech signal with vowel boundaries marked as per TIMIT acoustic-phonetic speech corpus. **b** Energy contour of a speech signal. **c** Mean smoothed energy contour. **d** Enhanced Energy signal contour. **e** VOP and VEP marking after convolving with FOGD operator. **f** Prediction and the ground truth (dashed line)

4.2 Stage 2 : Post processing of VOP and VEP locations using uniformity of the epochs and strength of the excitation

The resulting prediction is further improved in the second stage by removing spurious vowel regions and correcting the positions of VOP and VEP locations using the uniformity of epoch intervals and the SoE of the speech. It is understood that speech is comprised of source and system related information. The energy changes in the speech signal during the vowel production is reflected in the excitation source information. Therefore, these changes are characterized by the SoE. The fundamental frequency of vibration of vocal folds during the vowel production is near-uniform in the speech signal and it is measured as an inverse time period between two epoch locations where, the epoch location is the instant at which significant excitation takes place during the speech production. The SoE and uniformity of the epochs are computed from the zero frequency filtered (ZFF) signal (Yegnanarayana and Murty 2009) as it highlights the high information in lower frequency bands (Murty and Yegnanarayana 2008; Yegnanarayana et al. 2011). The procedure of computing these parameters is carried out in the following manner: Consider a speech signal $s[n]$ and perform high frequency boosting as it is noted that higher frequencies are more important for signal disambiguation than lower frequencies.

$$x[n] = s[n] - s[n-1] \tag{12}$$

The speech signal is fed to a resonator centered at 0 Hz. The resonator is realized using the following transfer function. The output of cascade of two ideal second order digital resonators at zero frequency is computed as:

$$y[n] = \sum_{k=1}^{4} \alpha_k y[n-k] + x[n] \tag{13}$$

where $a_1 = 4$, $a_2 = -6$, $a_3 = 4$ and $a_4 = -1$. The transfer function of the system is given by:

$$H(z) = \frac{1}{\left(1 - z^{-1}\right)^4} \tag{14}$$

The progression can be removed from the output signal using progression removal operation, which involves subtracting the local mean of the original signal at every instant of time. This is represented using the following expression:

$$\hat{y}[n] = y[n] - \tilde{y}[n] \tag{15}$$

where $\tilde{y}[n] = 1/2N + 1 \sum_{n=-N}^{N} y[n]$. Here 2N+1 is the size of the window used for computing local mean, which is typically average pitch period. The resulting output signal is called ZFF signal. The negative to positive zero crossings of ZFF signal corresponds GCIs. The gradient of ZFF signal at each GCI is termed as SoE (Yegnanarayana and Murty 2009; Gangamohan et al. 2014; Vydana et al. 2015).

The SoE is high and successive pitch cycles will be similar in the vowel regions. The spurious vowel regions are removed based on the uniformity of the epochs and the SoE. The positions of VOPs and VEPs are corrected based on combined cues from the SoE and uniformity of epoch intervals. The SoE exhibits positive trend from a local minimum at VOP and a negative trend from a local minimum at VEP respectively. The uniformity transition points on the pitch contour are also corresponds to the vowel boundaries. Therefore, the SoE contour and uniformity of epoch intervals can be used as an evidence for correcting the positions of VOPs and VEPs. The post-processing mechanism is demonstrated in Figs. 5, 6 using two different speech utterances. Figure 5a shows the speech signal used in the stage I with ground truth. Figure 5b–f correspond to the prediction in the first level, epoch intervals contour, spurious removed vowel regions, SoE of the continuous speech and the hypothesized prediction obtained via post processing of VOPs and VEPs. In this case, spurious vowels are not detected. However, the positions of vowel boundaries are corrected based on the SoE profile and uniformity of epochs. To demonstrate the significance of removing spurious vowel regions, a different speech utterance /"The ear and in turn the voice as well"/ is considered. The post processing mechanism for this signal is demonstrated in Fig. 6. A spurious vowel region (marked red in color) is shown in Fig. 6b. This region is masked as the epoch intervals are found to be non-uniform in that region and the corrected prediction is shown in Fig. 6c. Thus obtained prediction is further enhanced by correcting the locations of VOP and VEP based on the SoE. The VOP and VEP locations are aligned towards the positive and negative slopes of the SOE respectively. The starting point of positive trend and ending point of the negative trend associated with SoE were used to correct the locations through thresholding. The results produced after the post processing in the proposed method are significantly better than the baseline methods in terms of detection rate (DR) and FA. The combined sequence of steps from stage 1 and stage 2, for the proposed vowel region detection method are listed in below:

Sequence of steps	Vowel region detection
1	Compute instantaneous energy measure of speech signal using envelope of the derivative of the speech signal
2	Mean smooth of energy contour using 50 ms window
3	Enhance the smoothed energy contour using first order difference operator
4	Convolve the resulting signal using FOGD operator to detect VOPs and VEPs

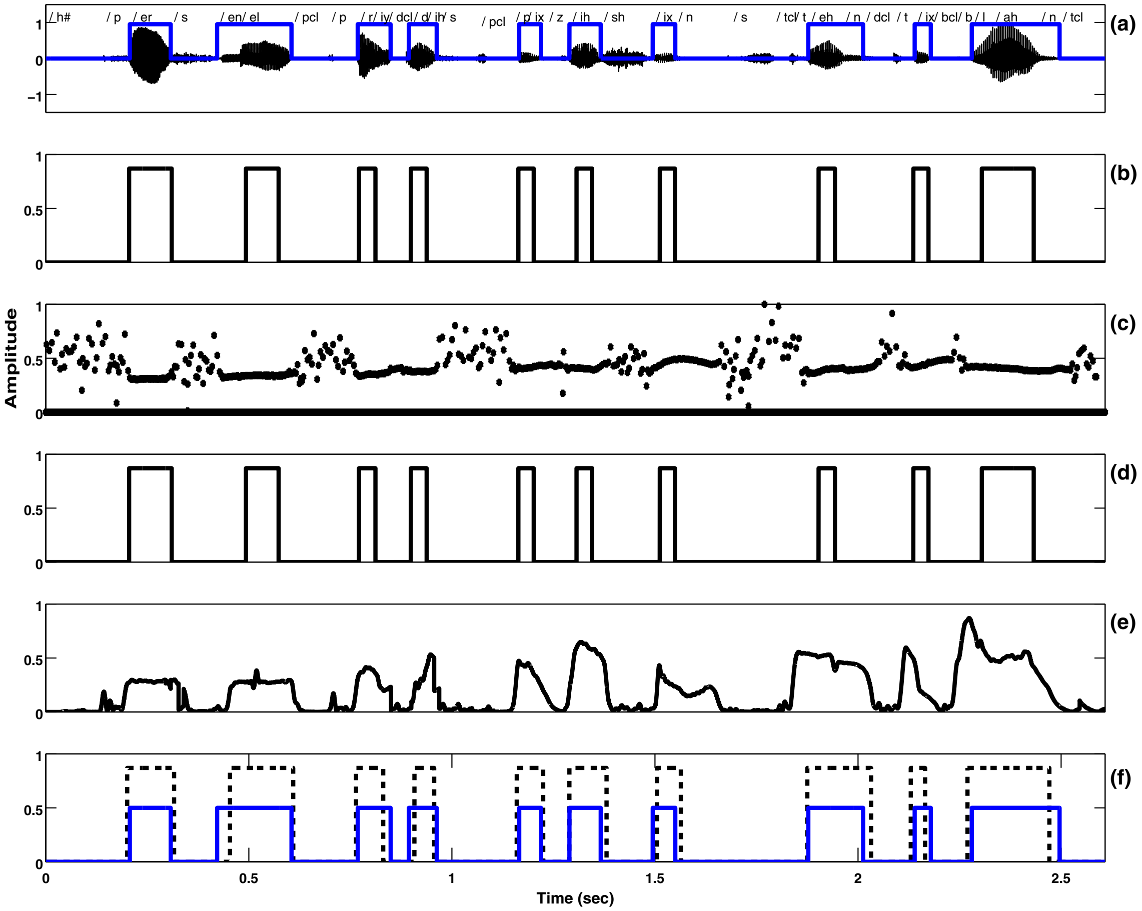

Fig. 5 Detection of vowel region using the proposed method with post processing (stage II). **a** Speech signal utterance /"Personal predispositions tend to blunt"/ with marked vowel boundaries. **b** Prediction. **c** Epoch intervals of speech signal. **d** Prediction without spurious vowel regions. **e** Strength of the excitation of speech signal. **f** Corrected vowel regions by modifying the positions of VOPs and VEPs of the speech signal

Sequence of steps	Vowel region detection
5	Remove spurious consonant regions in between a vowel region based on the distance criteria (20 ms)
6	Remove spurious vowel regions based on uniformity of epochs
7	Correct the locations of VOPs and VEPs based on the trend of the SoE

5 Performance evaluation of the proposed method

The experimental results are reported and comparisons between the proposed method and other state of the art methods is discussed in this section. The proposed two-stage vowel region detection method is evaluated by considering a subset of TIMIT acoustic-phonetic speech corpus. 500 test utterances from TIMIT acoustic-phonetic speech corpus, spoken by 50 speakers (25 male and 25 female) are used for evaluating the proposed vowel region detection method. For scoring, the detected vowel regions are compared with labeled boundaries of vowel regions given in TIMIT acoustic-phonetic speech corpus. The phones are mapped into two manner classes namely vowels and non-vowels. Vowel class includes vowels, semi-vowels, and diphthong sound units and, the remaining phones are treated as non-vowels. These classes are considered as ground truths for vowel region marking.

The performance of the proposed method is compared using metrics such as DR, missing rate (MR), total error (TE) and FA rate for different amounts of overlap of vowel region with ground truths.

- *DR refers to the ratio of a number of genuine vowel regions detected to the total number of reference ground truths.*

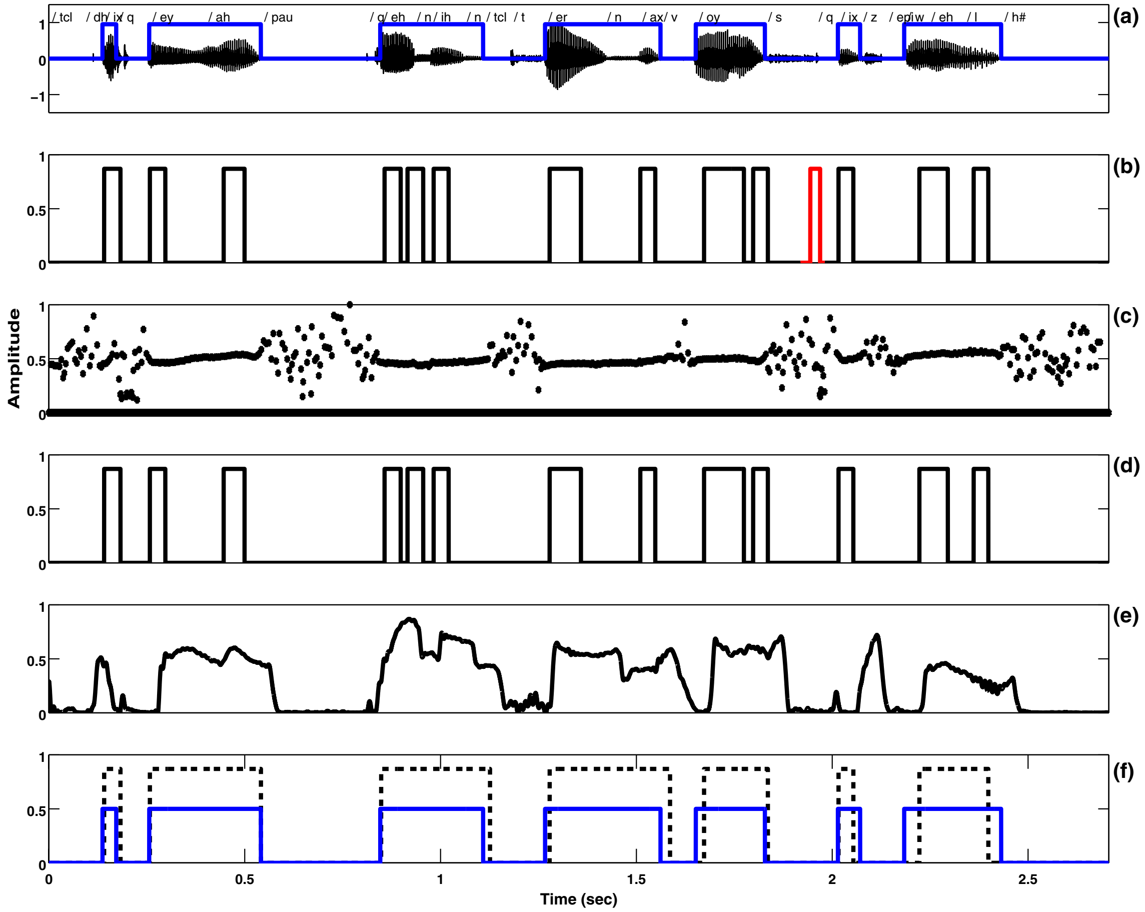

Fig. 6 Detection of vowel region using the proposed method with post processing (stage II). **a** Speech signal utterance "The ear and in turn the voice as well" with boundaries marked vowel regions. **b** Prediction. **c** Epoch intervals of speech signal. **d** Prediction without spurious vowel regions. **e** Strength of the excitation of speech signal. **f** Corrected vowel regions by modifying the positions of VOPs and VEPs of the speech signal

– *The ratio of undetected vowel regions to the total number of reference ground truth vowel regions is termed as MR.*
– *The ratio of vowel regions detected other than the genuine vowel regions to the total number of reference non-vowel regions is termed as FA rate or spurious rate.*
– *The sum of MR and the FA corresponds to the TE in the vowel detection.*

Table 1 demonstrate the performance comparison of proposed method using TIMIT acoustic-phonetic speech corpus with combined method and formant energy based method for clean speech. The first column indicates different vowel

Table 1 Performance analysis of vowel region detection using combined method (COMB), group delay based formants around GCI (FGCI) and proposed method for a clean speech on TIMIT acoustic-phonetic speech corpus

Vowel region detection method	Overlap with ground truth (%)					FA (%)
	25	50	60	80	90	
	DR	DR	DR	DR	DR	
COMB	78.24	68.18	60.33	41.95	26.42	24.56
FGCI	77.16	66.04	40.18	35.81	25.92	13.16
Proposed	92.82	81.73	71.29	60.84	53.26	7.87
Proposed with post-processing	96.67	96.16	95.47	89.52	82.37	3.85

DR detection rate, *FA* false alarm

region detection techniques. Columns two to six indicate DRs for the different percentage of overlap with the ground truth. The last column indicates the percentage of FA. From this table, it is observed that the first stage processing produced competitive results when compared with the state of the art techniques. However, FA rate is found to be slightly high (7.87%). The DR for 90% overlap criteria with the ground truth is low (53.26%). This is attributed to different kinds of production uncertainties associated with the speech signal production. Thus, instantaneous energy contour produced using envelope of the derivative of a speech signal is a quite robust feature that can be considered as an evidence to detect vowel regions using VOP and VEP as an anchor points. The performance of the proposed method has been significantly improved after the post-processing of the prediction. It produced a DR of 96.67 and 82.37 for 25% overlap and 90% overlap respectively.

Results for the proposed vowel region detection method in noisy environments are tabulated separately for different overlapping criteria with the ground truth in Table 2. The proposed vowel region detection method has been tested on speech utterances in presence of additive white Gaussian noise at signal to noise ratios ranging from 20 to 5 dB. From these results, it is evident that the DR of the proposed vowel detection method is significantly better than the state-of-the-art methods in noisy conditions. It is observed that the DR is 50.42 for 90% overlap with the ground truth at 5 dB and whereas DRs of other state-of-the-art methods are found to be 24.13 and 21.86%. After the post-processing, the DR has been increased and observed to be 94.14 and 77.06% for 25 and 90% respectively. Moreover, the FA rate (3.88%) is reduced in noisy conditions after the post-processing. It is noted that few nasal consonants are detected as vowels as they may have formant pattern that resemble those of vowels and possess side oral resonance cavity which is blocked by tongue or lips. Thus the proposed method produced significant improvement in vowel region detection at a higher amount of overlap with the ground truth through the post processing of VOPs and VEPs of the prediction. Therefore, it also proves that the energy contour generated by the envelope of the derivative of a speech signal possess desirable discriminative ability of vowels from the non-vowels in noisy environment.

Additional analysis is presented by comparing the proposed method with the recently proposed vowel detection techniques (Kumar et al. 2017; Kashani et al. 2017) in terms of DR, MR and FA in clean and noisy environments. During this evaluation, a vowel region overlap of

Table 3 Performance comparison of vowel region detection methods for clean speech in terms of detection rate (DR), missing rate (MR), and false alarm (FA)

Method	DR (%)	MR (%)	FA (%)
COMB	96.53	3.47	24.56
FGCI	96.65	3.35	13.16
Method I (Kumar et al. 2017)	85.12	14.88	14.39
Method II (Kashani et al. 2017)	91.60	8.40	10.50
Proposed with post processing	98.45	1.55	3.85

Table 2 Performance analysis of vowel region detection using combined method (COMB), group delay based formants around GCI (FGCI) and proposed method for a noisy speech on TIMIT acoustic-phonetic speech corpus

Vowel region detection method	Overlap with ground truth (%)					FA (%)
	25	50	60	80	90	
	DR	DR	DR	DR	DR	
SNR 20 dB						
COMB	75.26	68.16	60.27	40.79	26.40	24.87
FGCI	77.89	66.28	41.93	35.18	25.72	13.62
Proposed	90.51	81.63	71.11	58.42	53.26	7.87
Proposed with post-processing	95.78	95.37	93.68	85.47	78.87	4.24
SNR 10 dB						
COMB	75.18	68.16	60.01	40.36	26.50	25.97
FGCI	77.89	66.26	41.64	33.31	24.73	14.82
Proposed	90.51	81.54	70.93	58.32	51.34	8.80
Proposed with post-processing	94.45	94.23	92.84	84.86	77.33	4.78
SNR 5 dB						
COMB	75.18	68.03	60.01	38.23	24.13	27.04
FGCI	77.24	66.03	41.02	33.06	21.86	15.91
Proposed	90.26	81.43	70.13	58.11	50.42	10.12
Proposed with post-processing	94.15	94.17	92.31	84.44	77.06	5.95

DR detection rate, *FA* false alarm

Table 4 Performance comparison of vowel region detection methods for a noisy speech in terms of total error (TE = MR + FA)

Method	Total error (%)		
	Clean	Noise (20 dB)	Noise (10 dB)
COMB	28.03	28.34	29.44
FGCI	16.51	16.97	18.17
Method II (Kashani et al. 2017)	18.90	19.70	24.20
Proposed with post processing	5.40	5.79	6.33

5% with the ground is considered for the proposed method. From Table 3, it can be observed that the proposed method exhibits better performance than the method I, method II and formulated baseline methods in terms of MR and spurious rate for clean speech. The DR, missed rate and FA rate for the proposed method are found to be 98.45, 1.55 and 3.85% respectively. Table 4 demonstrates the performance comparison of the proposed method with different baseline methods in terms of TE in noisy environment. TE is defined as sum of FA rate and missed rate. The TE resulted for the proposed method is 5.40% for the clean speech, 5.79% at 20 dB SNR and 6.33% at 10 dB SNR respectively. It is noted that the proposed method outperformed the baseline methods in terms TE. Method I is not included as a part of this analysis, as results for the same are not reported in noisy environment. However, it is noted that the method I is based on the statistical models, which is expected to under-perform in mismatched conditions. VOPs and VEPs are considered to be the instant properties of a speech signal. The manifestation of these properties can be attributed to the source and the system information in the speech signal. In this context, instantaneous energy measure using the envelope derivative of the speech signal is found to be a better tool used to bring out the fine evidence to detect vowel boundaries in continuous speech signal.

6 Summary and conclusion

This work has shown that use of temporal energy transition measure using non-negative, frequency-weighted energy operator for the vowel region detection, which is feasible through the detection of VOPs and VEPs. Although previous studies have investigated the use of energy transition measure in temporal and spectral domains of speech signal to detect acoustic landmarks, it is noted that this work has highlighted in terms of less spurious and missed landmarks detection in both clean and noisy environment. This energy operator could produce high temporal resolution energy contour for a non-stationary speech signal to spot acoustic events in a precise manner eliminating block processing. The instantaneous energy of the signal can be calculated using three samples including the current sample. Additional advantage of this method is that no thresholding mechanism is used to spot spurious landmarks from the evidence. As noted by the use of additional post-processing operation for the removal of spurious events and position correction of landmark locations, the performance of this method is enhanced. The proposed scheme jointly utilized uniformity of the epochs and the SoE to eliminate spurious vowel regions and to correct the positions of VOPs and VEPs respectively. The performance of this method was evaluated using TIMIT acoustic-phonetic speech corpus and a significant improvement in vowel region detection was observed using proposed method compared to the state of the art methods. The robustness of the proposed can be evaluated by detecting the vowel regions on different forms of speech corpus for different applications. In the next level of work, VOP and VEP locations can be detected more precisely using combined temporal and spectral cues using advanced time–frequency analysis techniques in very high noisy environments.

References

Ananthapadmanabha, T., & Yegnanarayana, B. (1979). Epoch extraction from linear prediction residual for identification of closed glottis interval. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *27*(4), 309–319.

Deller, J. R, Jr., Proakis, J. G., & Hansen, J. H. (1993). *Discrete time processing of speech signals*. Englewood Cliffs: Prentice Hall PTR.

Donaldson, G. S., Rogers, C. L., Cardenas, E. S., Russell, B. A., & Hanna, N. H. (2013). Vowel identification by cochlear implant users: Contributions of static and dynamic spectral cues. *The Journal of the Acoustical Society of America*, *134*(4), 3021–3028.

Dumpala, S. H., Nellore, B. T., Nevali, R. R., Gangashetty, S. V., & Yegnanarayana, B. (2016). Robust vowel landmark detection using epoch-based features. In *INTERSPEECH* (pp. 160–164).

Fant, G. (1971). *Acoustic theory of speech production: With calculations based on X-ray studies of Russian articulations*. Berlin: Walter de Gruyter.

Gangamohan, P., Kadiri, S. R., Gangashetty, S. V., & Yegnanarayana, B. (2014). Excitation source features for discrimination of anger and happy emotions. In *Fifteenth annual conference of the International Speech Communication Association*.

Glass, J. R. (2003). A probabilistic framework for segment-based speech recognition. *Computer Speech & Language*, *17*(2), 137–152.

Hansen, J. H., Gray, S. S., & Kim, W. (2010). Automatic voice onset time detection for unvoiced stops (/p/,/t/,/k/) with application to accent classification. *Speech Communication*, *52*(10), 777–789.

Hermes, D. J. (1990). Vowel-onset detection. *The Journal of the Acoustical Society of America*, *87*(2), 866–873.

Johnson, K. (2004). Acoustic and auditory phonetics. *Phonetica*, *61*(1), 56–58.

Juneja, A., & Espy-Wilson, C. (2008). A probabilistic framework for landmark detection based on phonetic features for automatic speech recognition. *The Journal of the Acoustical Society of America*, *123*(2), 1154–1168.

Kaiser, J. F. (1990). On a simple algorithm to calculate the 'energy' of a signal. In *Proceedings of the 1990 international conference on acoustics, speech, and signal processing (ICASSP-90)*, pp. 381–384.

Kaiser, J. F. (1993). Some useful properties of Teager's energy operators. In *Proceedings of the 18th IEEE international conference on acoustics, speech, and signal processing (ICASSP '93)*, vol. 3, pp. 149–152.

Kashani, H. B., Sayadiyan, A., & Sheikhzadeh, H. (2017). Vowel detection using a perceptually-enhanced spectrum matching conditioned to phonetic context and speaker identity. *Speech Communication*, *91*, 28–48.

Kumar, A., Shahnawazuddin, S., & Pradhan, G. (2017). Improvements in the detection of vowel onset and offset points in a speech sequence. *Circuits, Systems, and Signal Processing*, *36*(6), 2315–2340.

Liu, S. A. (1996). Landmark detection for distinctive feature-based speech recognition. *The Journal of the Acoustical Society of America*, *100*(5), 3417–3430.

Makhoul, J. (1975). Linear prediction: A tutorial review. *Proceedings of the IEEE*, *63*(4), 561–580.

Murty, K. S. R., & Yegnanarayana, B. (2008). Epoch extraction from speech signals. *IEEE Transactions on Audio, Speech, and Language Processing*, *16*(8), 1602–1613.

O'Toole, J. M., Temko, A., & Stevenson, N. (2014). Assessing instantaneous energy in the EEG: A non-negative, frequency-weighted energy operator. In *Engineering in Medicine and Biology Society (EMBC), 2014 36th annual international conference of the IEEE*, pp. 3288–3291.

Palmu, K., Stevenson, N., Wikström, S., Hellström-Westas, L., Vanhatalo, S., & Palva, J. M. (2010). Optimization of an nleo-based algorithm for automated detection of spontaneous activity transients in early preterm EEG. *Physiological Measurement*, *31*(11), N85.

Pradhan, G., & Prasanna, S. M. (2013). Speaker verification by vowel and nonvowel like segmentation. *IEEE Transactions on Audio, Speech, and Language Processing*, *21*(4), 854–867.

Prasanna, S. M. & Yegnanarayana, B. (2005). Detection of vowel onset point events using excitation information. In *Ninth European conference on speech communication and technology*.

Prasanna, S. M., & Pradhan, G. (2011). Significance of vowel-like regions for speaker verification under degraded conditions. *IEEE Transactions on Audio, Speech, and Language Processing*, *19*(8), 2552–2565.

Prasanna, S. M., Reddy, B. S., & Krishnamoorthy, P. (2009). Vowel onset point detection using source, spectral peaks, and modulation spectrum energies. *IEEE Transactions on Audio, Speech, and Language Processing*, *17*(4), 556–565.

Rao, K. S., & Yegnanarayana, B. (2009). Duration modification using glottal closure instants and vowel onset points. *Speech Communication*, *51*(12), 1263–1269.

Rose, P. (2003). *Forensic speaker identification*. Boca Raton: CRC Press.

Saha, P., Laskar, R. H., & Laskar, A. (2016). A pre-processing method for improvement of vowel onset point detection under noisy conditions. *Speech Communication*, *80*, 71–83.

Salomon, A., Espy-Wilson, C. Y., & Deshmukh, O. (2004). Detection of speech landmarks: Use of temporal information. *The Journal of the Acoustical Society of America*, *115*(3), 1296–1305.

Schutte, K., & Glass, J., (2005). Robust detection of sonorant landmarks. In *Ninth European conference on speech communication and technology*.

Stevens, K. N. (2000). *Acoustic phonetics*. Cambridge: MIT Press.

Teager, H., & Teager, S. (1990). Evidence for nonlinear sound production mechanisms in the vocal tract. *Speech Production and Speech Modelling*, *55*, 241–261.

Vuppala, A. K., & Rao, K. S. (2013). Vowel onset point detection for noisy speech using spectral energy at formant frequencies. *International Journal of Speech Technology*, *16*(2), 229–235.

Vuppala, A. K., Rao, K. S., & Chakrabarti, S. (2012). Improved vowel onset point detection using epoch intervals. *AEU-International Journal of Electronics and Communications*, *66*(8), 697–700.

Vuppala, A. K., Yadav, J., Chakrabarti, S., & Rao, K. S. (2012). Vowel onset point detection for low bit rate coded speech. *IEEE Transactions on Audio, Speech, and Language Processing*, *20*(6), 1894–1903.

Vydana, H. K., Vikash, P., Vamsi, T., Kumar, K. P., & Vuppala, A. K. (2015). Detection of emotionally significant regions of speech for emotion recognition. In *India conference (INDICON), 2015 Annual IEEE*, pp. 1–6.

Vydana, H. K., & Vuppala, A. K. (2016). Detection of fricatives using s-transform. *The Journal of the Acoustical Society of America*, *140*(5), 3896–3907.

Yadav, J., & Rao, K. S. (2013). Detection of vowel offset point from speech signal. *IEEE Signal Processing Letters*, *20*(4), 299–302.

Yegnanarayana, B., Prasanna, S. M. & Guruprasad, S. (2011). Study of robustness of zero frequency resonator method for extraction of fundamental frequency. In *2011 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 5392–5395.

Yegnanarayana, B., & Murty, K. S. R. (2009). Event-based instantaneous fundamental frequency estimation from speech signals. *IEEE Transactions on Audio, Speech, and Language Processing*, *17*(4), 614–624.