Multimed Tools Appl (2018) 77:4753–4767
DOI 10.1007/s11042-017-5044-8

Improved vowel region detection from a continuous speech using post processing of vowel onset points and vowel end-points

Ramakrishna Thirumuru[1] · **Suryakanth V. Gangashetty**[1] · **Anil Kumar Vuppala**[1]

Received: 25 January 2017 / Revised: 19 June 2017 / Accepted: 18 July 2017 /
Published online: 1 August 2017

Abstract Vowels are produced with an open configuration of the vocal tract, without any audible friction. The acoustic signal is relatively loud with varying strength of impulse-like excitation. Vowels possess significant energy content in the low-frequency bands of the speech signal. Acoustic events such as vowel onset point (VOP) and vowel end-point (VEP) can be used as landmarks to detect vowel regions in a speech signal. In this paper, a two-stage algorithm is proposed to detect precise vowel regions. In the first level, the speech signal is processed using zero frequency filtering to emphasize energy content in low-frequency bands of speech. Zero frequency filtered signal predominantly contains low-frequency content of the speech signal as it is filtered around 0 Hz. This process is followed by the extraction of dominant spectral peaks from the magnitude spectrum around glottal closure regions of the speech signal. The vowel onset points and vowel end-points are obtained by convolving the enhanced spectral contour of zero frequency filtered signal with first order Gaussian differentiator. In the next level, a post-processing is carried out in the regions around VOP and VEP to remove spurious vowel regions based on uniformity of epoch intervals. In addition, the positions of VOPs and VEPs are also corrected using the strength of the excitation of the speech signal. The performance of the proposed vowel region detection method is compared with the existing state of art methods on TIMIT acoustic-phonetic speech corpus. It is reported that this method produced significant improvement in vowel region detection in clean and noisy environments.

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

Suryakanth V. Gangashetty
svg@iiit.ac.in

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

[1] Language Technology Research Center, International Institute of Information Technology Hyderabad, Hyderabad, India

Keywords Vowel onset point (VOP) · Vowel end-point (VEP) · Zero frequency filtering · Magnitude spectrum · Epoch intervals · Strength of the excitation

1 Introduction

Vowels are speech sounds produced with an open configuration of the vocal tract, with vocal cords vibration. The air stream is not severely obstructed in this process. These are primary units of the sound system of a language [1]. Vowel region detection is a task of identifying a vowel occurrences in a continuous speech with precise boundary markings. Vowel onset point (VOP) is the time instant at which vowel region begins and vowel end-point (VEP) can be considered as the time instant at which vowel region ends in a continuous speech. These two landmarks constitute vowel boundaries. The speech signal is a non-stationary signal and its generation can be perceived as a sequence of changes and these changes are known as events. VOP and VEP are two such acoustic events.

The importance of precise vowel region detection from a continuous speech can be felt in many applications [5]. The detection of accurate vowel regions from a continuous speech is expected in automatic speech recognition, smart audio filtering, recognition of CV units in Indian languages, determining the duration of vowels, speech rate manipulation and multimedia synchronization [3, 9, 11, 12]. The knowledge of phonetic classes can be integrated into the statistical based automatic speech recognition (ASR) systems to improve the overall performance of the system. This mechanism can provide a platform for an automatic speech recognition systems to utilize segment-based approach instead of dominant frame based HMM models [4, 13]. The motivation for the proposed work comes from the fact that, the performance of the above mentioned applications depends on the accurate estimation of vowel boundaries.

In Indian languages most of the syllables are of the consonant-vowel (CV) type. Vowel region starts with vowel onset point and ends with vowel end-point. In the literature, vowel region is considered as 100 ms duration from the vowel onset point. Next, the VOP and VEP detection methods are prone to speech production uncertainties. Therefore a robust vowel region detection method is required by considering the fixed duration and speech production uncertainties. In this paper, an improved method is presented for the detection of vowel regions using acoustic features. The proposed method is carried out on two levels. In the first level, the landmarks such as vowel onset point and vowel end-point are detected with an assumption that vowels exhibit high sonority and loudness. A combined approach is proposed from the previous works, using the features of zero frequency filtered signal for vowel region detection. These landmarks are detected based on spectral content intensity variation around glottal closure instants of zero frequency filtered speech signal. Zero frequency filtered speech signal around glottal closure regions is processed to bring out high information pertaining to the vowel regions. In the second level, two clues such as uniformity of epochs and strength of the excitation of the speech signal are used to eliminate spurious vowel regions along with the correction of onset point and end-point of vowels.

Rest of the paper is organized as follows: Related works for the vowel region detection are briefly discussed in Section 2. Section 3 describes baseline methods for detecting vowel regions using vowel onset points and end-points. The zero frequency filtering technique and extraction of the strength of the excitation of speech signal are also briefly described in this section. The proposed method for vowel region detection is described in Section 4.

The performance of the proposed method is studied using TIMIT acoustic-phonetic speech corpus. and, results are discussed in Section 5. Section 6 provides summary and conclusions of this work.

2 Related works

Acoustic-phonetic feature based landmark detection has received considerable attention in recent years. In the literature, the problem of detecting vowel onset points and vowel end-points is studied using the extraction of distinctive features of the speech signal. Several vowel onset point detection methods have been proposed using the acoustic features of the speech signal and few of them are referred here [8, 10, 14, 16]. In [8] VOPs are detected from the excitation information of speech signal. The speech signal is processed using the Linear Prediction (LP) analysis technique to extract the LP residual. The Hilbert envelope of the LP residual is used as an evidence to detect VOPs as this signal contain excitation information of the speech signal. Vowel onset point detection methods are proposed based on rising slope of spectral amplitude in the magnitude spectrum of the speech signal. A method is proposed by combining evidence from the excitation source, spectral peaks and modulation spectrum for the detection of vowel onset point [10]. Each of these evidence possess complementary information pertaining to the VOPs and it can produce enhanced vowel detection performance. vowel onset point detection for vowel region extraction is proposed based on spectral energy present in glottal closure regions of speech signal [14, 16]. In this method, formants are extracted using the group delay function, and glottal closure instants are extracted by using zero frequency filtering technique. The spectral energy computed around the glottal closure instants is high and robust. Therefore it is used as an evidence for detecting the VOPs. The locations of the vowel onset points were corrected based on the uniformity of epoch intervals in the vowel region [15]. In another attempt, two methods were proposed to detect vowel end-points from the speech signal [18]. The first method uses the combined evidence from excitation source, spectral peaks and modulation spectrum for determining the vowel end-point. In the second method, spectral energy within the glottal closure region is used for determining the vowel end-point. In both the methods, enhanced evidence is convolved with FOGD operator to detect the locations of the VEPS. The region between vowel onset point and vowel end-point is considered as a vowel region. These methods report that most vowel onset and end-points fall within 40 ms deviation. Recently a technique was proposed [6] on improvements in detection of vowel onset point and vowel end-point using three-class classifier with front-end feature extraction technique that exploits spectral and temporal characteristics of the excitation source information of the speech signal.

3 Baseline methods for VOP and VEP detection techniques, Zero frequency filtering and strength of the excitation of the speech signal

Vowel boundaries start with vowel onset point and end with vowel end-point. In this Section, state of the art methods for vowel detection using VOP and VEP in a continuous speech is discussed. Two vowel region detection methods are fromulated using the detection of VOPs and VEPs. These methods are based on previous works mentioned in [10, 14, 18]. The experimental results obtained for the proposed method are compared with these methods.

In addition, we briefly discuss speech signal processing technique known as Zero frequency filtering and strength of the excitation of the speech signal.

3.1 VOP region detection by combining the evidence from excitation source, spectral peaks and modulation spectrum

The combined evidence for the boundary markings of a vowel region is derived from three parameters namely excitation source, spectral peaks and modulation spectrum [10]. LP residual corresponds to the excitation source information and it is a bipolar signal. Hilbert envelope of linear prediction (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 is enhanced using first order difference (FOD) to improve the detection of VOP. The enhanced signal is convolved with first order Gaussian difference (FOGD) operator, and convolved output is the evidence for VOP. This acoustic event is based on the nature of the gradient of the output 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 are selected from first 128 points and plotted as a function of a time, and this represents the energy of the spectral peaks. The VOP can be observed as a significant change in a complimentary manner in this time varying signal. The changes in VOP is enhanced in the spectral energy signal using FOD. The variations that represent VOP is extracted using FOGD operator.

The information in speech signal can also be represented by changes in the time domain envelope and this is referred as amplitude-modulation spectrum or simply modulation spectrum. The modulation spectrum of speech is dominated by the low-frequency components. The speech signal is analyzed using 18 trapezoidal critical band filters between 0 and 4 kHz. An amplitude envelope signal is computed using half wave rectification and low pass filtering on all bands. Amplitude envelope signals are down-sampled to 80 samples/s 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 evidence for VOP using FOD and FOGD respectively. This method uses three independent and complementary evidence to derive a single combined evidence. This is obtained by adding three shreds of evidence sample by sample and the performance of the detection of VOP is better than individual evidence. The positive peaks of combined evidence are marked as VOPs in the speech signal.

3.2 VEP detection by combining the evidence from excitation source, spectral peaks and modulation spectrum

The time instant at which the offset of vowel take place in the speech signal is known as vowel end-point [18]. The individual vowel evidence are derived from the excitation source, spectral peaks and modulation spectrum respectively. They are enhanced and convolved with a the FOD of Gaussian window. After convolving each of the evidence with FOGD operator, the three individual evidence are combined by adding sample by sample to obtain a combined evidence. The negative peaks in the combined vowel evidence signal are hypothesized as the locations of vowel end-points

Vowel regions are detected by locating the VOPs and and VEPs using the evidence from excitation source, spectral peaks and modulation spectrum. This vowel region detection method is termed as COMB method in this paper. Figure 1 demonstrates the experimental result for the COMB method for a speech utterance /"personal predispositions tend to blunt"/. Figure 1a refers to a continuous speech utterance with the ground truth. Combined evidence is shown in Fig. 1b. Figure 1c, d, e, f refer to mean smoothed evidence contour, enhanced evidence using first order difference, VOP and VEP marking for the speech signal after convolving with FOGD operator and prediction respectively. This method produced high false alarm due to spurious peaks.

3.3 VOP detection using spectral energy around glottal closure instants

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. In this method, spectral energy at the glottal closure instants is used as an evidence to detect VOP in a continuous speech [14]. Firstly, the glottal closure instants are extracted using zero frequency filtering. Around the glottal closure instants, formants are computed for 30% of speech samples using group delay function. The spectral energy of a speech signal is derived from the first three formants and the same is plotted as a function of time. This contour is smoothed using mean smoothing window of 50 ms and enhanced using FOD operator. Significant changes in the spectral characteristic present in the enhanced signal are detected by convolving the same with FOGD operator having 100 ms length and 25 ms standard deviation. After eliminating the spurious peaks, positive peaks of this signal represent locations of VOPs.

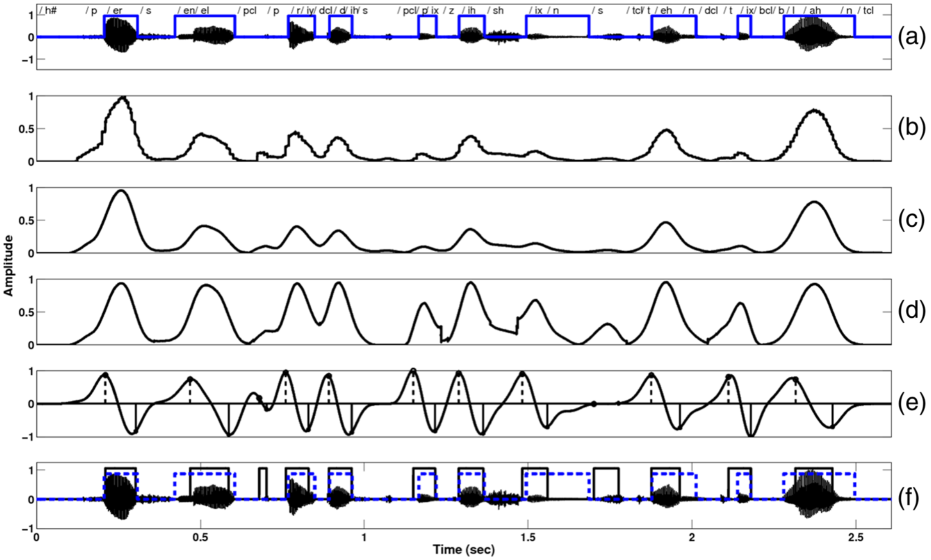

Fig. 1 Detection of vowel region using COMB method for a speech utterance /"personal predispositions tend to blunt"/. **a** Speech signal with phonemic marking of vowel regions given in TIMIT acoustic-phonetic speech corpus. **b** Combined evidence. **c** Mean smoothed evidence contour. **d** Enhanced evidence using FOD. **e** Hypothesized VOPs and VEPs for the speech signal. **f** Predicted vowel regions

3.4 VEP detection using spectral energy around glottal closure instants

In [18], another method of VEP detection is also proposed. The spectral energy contour of a speech signal is derived from the first three formats around the glottal closure region. This is enhanced using first order difference function and convolved with a FOGD operator having 100 ms length and 25 ms standard deviation. The convolved output obtained represents vowel evidence. The spurious negative peaks of the this evidence signal are removed using distance criteria and the remaining negative peaks represent the locations of VEPs.

The regions between VOPs and VEPs obtained using spectral energy are marked as vowel regions. This method is referred as FGCI method in this paper. Figure 2 depicts the result for a speech utterance /“personal predispositions tend to blunt”/ using FGCI method. Figure 2a, b, c, d, e and f refers to a continuous speech utterance, Sum of first 3 formant peaks, mean smoothed evidence contour, enhanced evidence using first order difference operator, VOP and VEP marking after convolving with FOGD operator and prediction respectively.

3.5 Zero frequency filtering and extraction of strength of the excitation information

Zero frequency filtering of a speech signal highlights the high information in lower frequency bands [7, 20]. 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.

$$x(n) = s(n) - s(n-1) \tag{1}$$

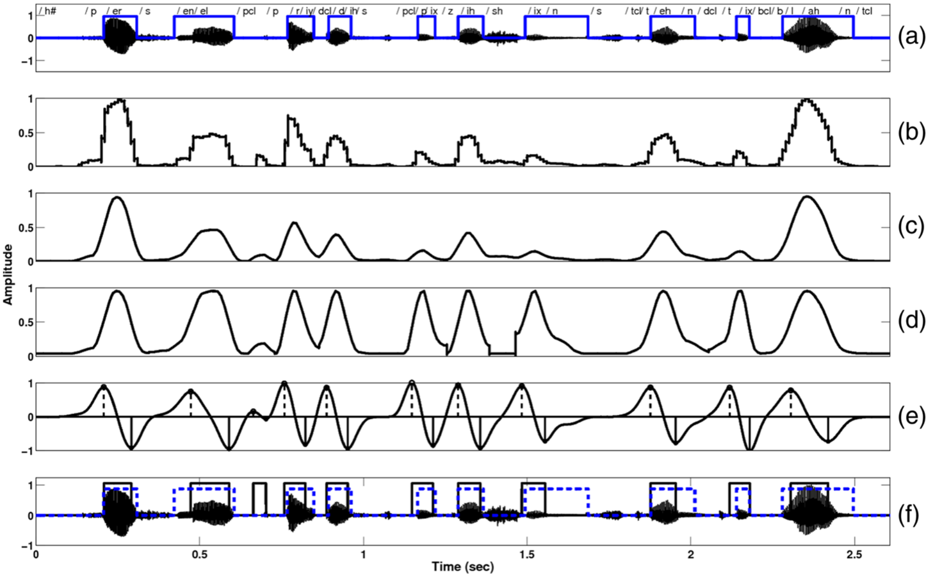

Fig. 2 Detection of vowel region using FGCI method for a speech utterance/“personal predispositions tend to blunt”/. **a** Speech signal with phonemic marking of vowel regions given in TIMIT acoustic-phonetic speech corpus. **b** Combined evidence. **c** Mean smoothed evidence contour. **d** Enhanced evidence using FOD. **e** Hypothesized VOPs and VEPs for the speech signal. **f** Predicted vowel regions

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

$$y\left(n\right)=\sum_{k=1}^{4}\alpha_{k}y\left(n-k\right)+x\left(n\right) \tag{2}$$

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

$$H\left(z\right)=\frac{1}{\left(1-z^{-1}\right)^{4}} \tag{3}$$

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

$$\hat{y}\left(n\right)=y\left(n\right)-\tilde{y}\left(n\right) \tag{4}$$

where $\tilde{y}\left(n\right)=\frac{1}{2N+1}\sum_{n=-N}^{N}\mathrm{y}(n)$. Here 2N+1 is the size of window used for computing local mean, which is typically average pitch period. The resulting output signal is called zero frequency filtered (ZFF) signal. The strength of the excitation (SoE) can be computed from the zero frequency filtered signal [19]. The negative to positive zero crossings of zero frequency filtered signal corresponds to glottal closure instants (GCIs). The gradient of zero frequency filtered signal at each GCI is termed as SoE [2, 17, 19]. The SoE contour can be used as a complimentary evidence for finding the positions of VOPs and VEPs.

4 Proposed vowel region detection method

This section describes an improved method to detect vowel regions in a continuous speech. The zero frequency filtering serves as a tool, which emphasizes the low-frequency content of speech signal. The sharp rise and fall of energies around glottal closure instants can be visualized as VOPs and VEPs. The proposed method for vowel region detection is implemented in two stages. In the first stage, vowel onset and end-points are detected from the zero frequency filtered speech signal. The positions of VOPs and VEPs are corrected along with the removal of spurious vowel regions. The steps involved in the proposed method are illustrated in the Fig. 3.

Let $x(n)$, $n = 0, 1, 2, ..., N - 1$ be a zero frequency filtered speech signal. This signal is perceived in terms of segments, possibly overlapping, of length L with the starting points of the segments D units distance apart. In this case, a window $w(n)$ of size L with overlap is used to specify the speech segment for the processing. It can be written as follows

$$x_p(n) \equiv x(n + (p - 1)D).w(n),\ p = 1, 2, ...P \tag{5}$$

where P denote the number of segments of speech signal under consideration. The entire speech signal in terms of overlapping segments can be represented as

$$x(n) \equiv \{x_p(n)\},\ p = 0, 1, 2.........., P - 1 \tag{6}$$

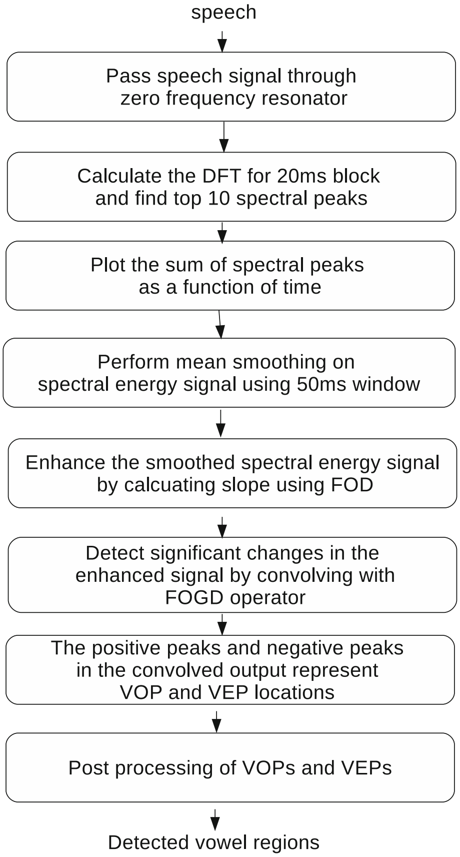
speech
Pass speech signal through zero frequency resonator
Calculate the DFT for 20ms block and find top 10 spectral peaks
Plot the sum of spectral peaks as a function of time
Perform mean smoothing on spectral energy signal using 50ms window
Enhance the smoothed spectral energy signal by calcuating slope using FOD
Detect significant changes in the enhanced signal by convolving with FOGD operator
The positive peaks and negative peaks in the convolved output represent VOP and VEP locations
Post processing of VOPs and VEPs
Detected vowel regions

Fig. 3 Flow chart describing the sequence of steps in the proposed for detecting vowel regions

In this work, the speech signal is processed in blocks of 20 ms with a shift of 10 ms. For each segment of length L, a 256 point discrete Fourier transform (DFT) is computed in the following manner

$$X_p(k) = \frac{1}{L}\sum_{i=0}^{L-1} x_p(n)\exp\left(-j\frac{2\pi nk}{L}\right), k = 0, 1, 2....., L-1 \tag{7}$$

The largest ten peaks from each block are selected from the magnitude spectrum and sum of these spectral peaks are plotted as a function of time. Thus spectral energy contour is obtained for the speech signal. The VOPs and VEPs are characterized by the spectral peaks and valleys of this contour. This energy contour is enhanced by computing its slope using first order difference function. A peak finding method using first order Gaussian differentiator (FOGD) is used for automatic detection of the peaks and valleys of the spectral energy contour of the speech signal. A M-point Gaussian window function is defined as

$$w(n) = e^{-\frac{1}{2}\frac{(n-M/2)^2}{\sigma^2}}, n = 0, 1, 2, ..., M \tag{8}$$

$M/2$ is mean and σ^2 is its variance. The derivative of a Gaussian which denotes the slope of a series at each sample is computed as

$$w_d(n) = w(n+1) - w(n) \tag{9}$$

The Gaussian window and its first order derivative are shown in the Fig. 4. The convolution of FOGD function with spectral contour highlights positive slopes and negative slopes in the spectral contour. It can be noted that spectral positive slope in the contour gets manifested as peak and negative slope as a valley in the output. These peaks and valleys correspond to the VOPs and VEPs in the speech signal. These landmarks are susceptible to the uncertainties associated with speech production and noise. This procedure is termed as ZFF method in this paper. Experimentation results for a speech utterance is shown in Fig. 5. Zero frequency filtered signal, smoothed contour of spectral peaks, enhanced FOD signal and FOGD operated signal are shown in Fig. 5b, c, d, and e respectively. The vowel regions detected by the ZFF method (first stage) are not in-line with the phonemic marked vowel regions as per TIMIT acoustic-phonetic speech corpus. The prediction signal possesses spurious and misaligned vowel regions. This is depicted in in Fig. 5f.

In the second stage, spurious vowel regions are removed and positions of vowel landmarks are corrected. The spurious vowel regions are eliminated based on the uniformity of the epochs with an assumption that the successive pitch cycles will be same in the vowel region [15]. The positions of VOPs and VEPs are corrected based on strength of the excitation as it exhibits positive trend from a local minimum at VOP and a negative trend from a local minimum at VEP respectively. ZFF method is used to find epoch intervals and the

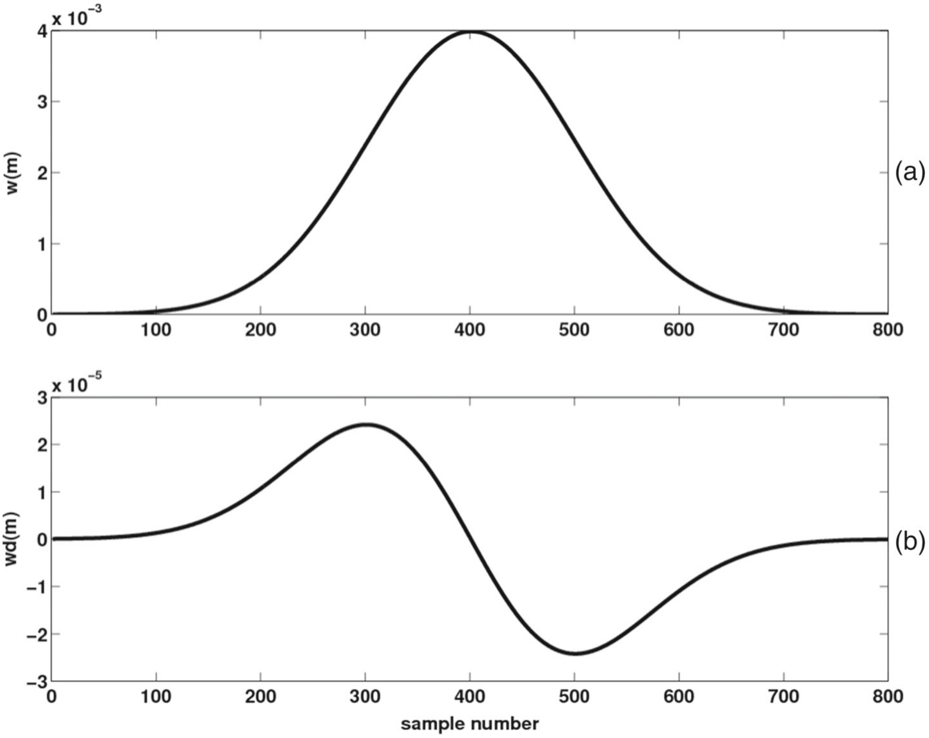

Fig. 4 **a** 801 point Gaussian window with spread = 100. **b** The FOGD operator

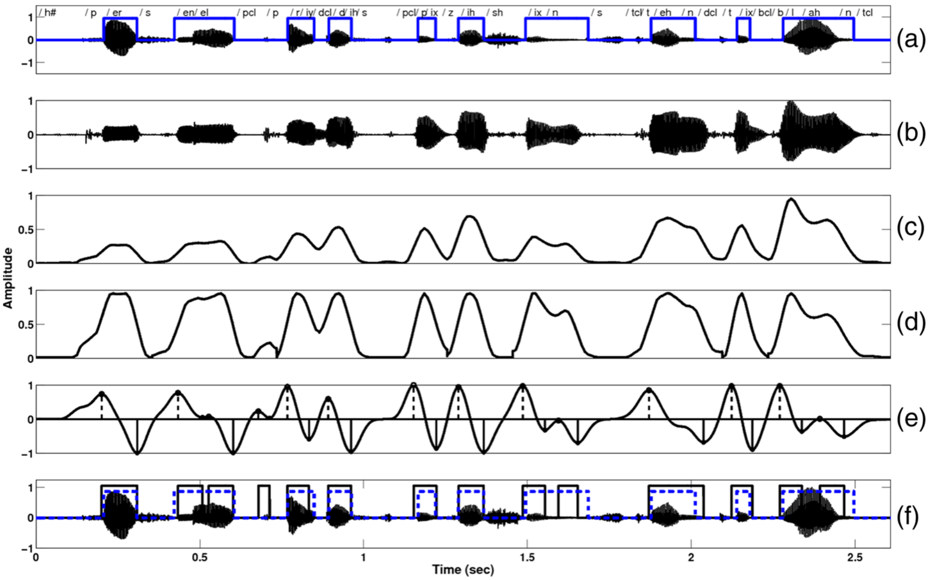

Fig. 5 Detection of vowel region using ZFF method (stage I) for a speech utterance/“personal predispositions tend to blunt”/. **a** Speech signal with phonemic marking of vowel regions given in TIMIT acoustic-phonetic speech corpus. **b** ZFF signal. **c** Sum of 10 spectral peaks in ZFF signal. **d** Enhanced spectral energy signal. **e** Hypothesized VOPs and VEPs for the speech signal. **f** Predicted vowel regions

strength of the excitation signal. This mechanism is demonstrated in Fig. 6 using a continuous speech utterance. Figure 6a shows the continuous speech signal with ground truth. Figure 6b, c, d, e and f correspond to prediction in the first level, epoch intervals, spurious removed vowel regions, strength of the excitation of the continuous speech and the hypothesized prediction obtained via post processing of VOPs and VEPs. The results produced by the proposed method is relatively better than COMB method and FGCI methods in terms of detection rate and false alarms.

5 Performance of the proposed method

In this section, experimental results are reported and comparisons between the proposed method and existing methods are provided. The proposed two-stage vowel region detection method is evaluated by considering a subset of TIMIT acoustic-phonetic speech corpus. 1000 test utterances from TIMIT acoustic-phonetic speech corpus, spoken by 100 speakers (50 male and 50 female) are used for evaluating vowel detection method from a continuous speech. The detected vowel regions are compared with phonemic marked vowel boundaries given in TIMIT acoustic-phonetic speech corpus. These boundaries are considered as ground truths for vowel region marking. The region between VOP and VEP in a continuous speech is considered as a vowel region. The performance of this method is evaluated based on metrics such as detection rate and false alarm for different amounts of overlap of vowel region with ground truth. Detection rate refers to number of vowel regions identified correctly in-line with ground truth and false alarm corresponds to the number of non-vowels

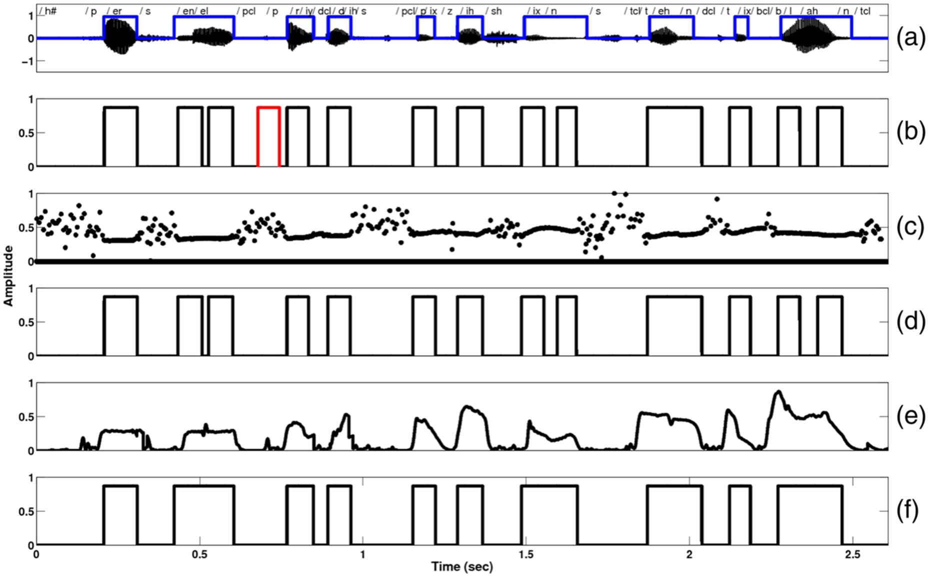

Fig. 6 Detection of vowel region using ZFF method with post processing (stage II) for a speech utterance/"personal predispositions tend to blunt"/. **a** Speech signal with phonemic marking of vowel regions given in TIMIT acoustic-phonetic speech corpus. **b** Prediction. **c** Epoch intervals of speech 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

interpreted as vowels. Tables 1 and 2 demonstrate the performance comparison of proposed method in the first stage (ZFF method) using TIMIT acoustic-phonetic speech corpus with combined method and formant energy based method for clean and noisy speech respectively. The first column indicates different vowel region detection techniques. Columns 2–6 indicate the different percentage of overlap with the ground-truth. The last column indicates the percentage of false alarm. From these Tables, it is observed that the first level processing produced competitive results when compared with state of art techniques. However, false alarm rate is found to be high and detection rate for 90% overlap with the ground-truth is

Table 1 Performance analysis of vowel region detection using combined method (COMB), group delay based formants around GCI (FGCI) and **ZFF method** for a **clean** speech on TIMIT database (Det = Detection; FA = False alarm)

Vowel region detection method	Overlap with ground-truth (%)					FA in (%)
	25	50	60	80	90	
	Det	Det	Det	Det	Det	
COMB	78	68	61	41	26	24
FGCI	77	66	40	35	25	13
ZFF	88	74	65	52	48	12

Table 2 Performance analysis of vowel region detection using combined method (COMB), group delay based formants around GCI (FGCI) and **ZFF method** for a **noisy speech** on TIMIT database (Det = Detection; FA = False alarm)

Vowel region detection method	Overlap with ground-truth (%)					FA in (%)
	25	50	60	80	90	
	Det	Det	Det	Det	Det	
SNR 20 dB						
COMB	75	68	61	40	26	24
FGCI	77	66	41	35	25	13
ZFF	86	74	65	52	46	12
SNR 10dB						
COMB	75	68	61	40	26	24
FGCI	77	65	40	33	24	13
ZFF	86	74	65	52	46	12
SNR 5dB						
COMB	75	68	61	38	24	24
FGCI	77	66	41	33	21	13
ZFF	86	74	65	52	46	12

low(47%). This is attributed to different kinds of uncertainties associated with the speech signal.

Tables 3 and 4 illustrates the results obtained after the post processing of VOPs and VEPs. From these results, it is evident that the performance of the proposed vowel detection method is significantly better than the other state of art methods in clean and noisy conditions. It is observed that about 30% more vowel regions are detected for 90% overlap with the ground-truth. In addition, false alarm rate (6% and 8%) also reduced in clean and noisy conditions using the proposed method. This method produced significant improvement in vowel region detection through the post processing of VOPs and VEPs. The performance of the proposed method is more significant at a higher amount of overlap with the ground truth.

Table 3 Performance analysis of vowel region detection using combined method (COMB), group delay based formants around GCI (FGCI) and **ZFF method with post processing** for a **clean speech** on TIMIT database with landmark correction (Det = Detection; FA = False alarm)

Vowel region detection method	Overlap with ground-truth (%)					FA in (%)
	25	50	60	80	90	
	Det	Det	Det	Det	Det	
COMB	90	82	75	62	56	16
FGCI	90	82	72	60	55	10
ZFF + Post processing	94	94	92	85	78	6

Table 4 Performance analysis of vowel region detection using combined method (COMB), group delay based formants around GCI (FGCI) and **ZFF method with post processing** for a **noisy speech** on TIMIT database with landmark correction (Det = Detection; FA = False alarm)

Vowel region detection method	Overlap with ground-truth (%)					FA in (%)
	25	50	60	80	90	
	Det	Det	Det	Det	Det	
SNR 20dB						
COMB	90	82	75	62	56	16
FGCI	90	82	72	60	57	10
ZFF + Post processing	94	94	92	85	76	6
SNR 10dB						
COMB	90	82	75	62	52	17
FGCI	90	82	72	60	57	12
ZFF + Post processing	93	93	92	84	75	8
SNR 5dB						
COMB	90	82	75	60	50	17
FGCI	90	81	73	60	54	12
ZFF + Post processing	93	90	90	82	75	8

6 Summary and conclusions

In this paper, an improved method for the vowel region detection through the post-processing of VOP and VEP were discussed. Although energy transition detection is used for vowel region detection in a continuous speech because of it’s simplicity and accuracy, it is difficult to spot VOP and VEP accurately due to speech production and noise uncertainties. A post processing methodology was introduced to circumvent this difficulty. We used the combination of the speech signal processing techniques to detect vowel regions and performed post processing of the same. Experimentation was carried out on zero frequency filtered speech signal to highlight the low-frequency content of speech. To enhance the detection performance, the proposed scheme jointly utilized uniformity of the epochs and the strength of the excitation to eliminate spurious vowel regions and to correct the positions of VOPs and VEPs respectively. The performance of the proposed method was compared with recent state of art methods for both clean and noisy speech. 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 existing methods. The robustness of the proposed can be evaluated by detecting the vowel regions on different forms of speech corpus for different applications.

References

1. Deller JR, Proakis JG, Hansen JH (1993) Discrete time processing of speech signals. Prentice Hall PTR, Englewood Cliffs
2. Gangamohan P, Kadiri SR, Gangashetty SV, Yegnanarayana B (2014) Excitation source features for discrimination of anger and happy emotions. In: INTERSPEECH, pp 1253–1257

3. Gangashetty SV, Sekhar CC, Yegnanarayana B (2004) Detection of vowel onset points in continuous speech using autoassociative neural network models. In: Proceedings of the international conference on spoken language processing, pp 401–410
4. Glass JR (2003) A probabilistic framework for segment-based speech recognition. Comput Speech Lang 17(2):137–152
5. Hermes DJ (1990) Vowel-onset detection. J Acoust Soc Am 87(2):866–873
6. Kumar A, Shahnawazuddin S, Pradhan G (2016) Improvements in the detection of vowel onset and offset points in a speech sequence. In: Circuits, systems, and signal processing, pp 1–26
7. Murty KSR, Yegnanarayana B (2008) Epoch extraction from speech signals. IEEE Trans Audio Speech Lang Process 16(8):1602–1613
8. Prasanna SM, Yegnanarayana B (2005) Detection of vowel onset point events using excitation information. In: INTERSPEECH, pp 1133–1136
9. Prasanna SM, Gangashetty SV, Yegnanarayana B (2001) Significance of vowel onset point for speech analysis. In: Proceedings of international conference on signal processing and communications. Citeseer, pp 81–88
10. Prasanna SM, Reddy BS, Krishnamoorthy P (2009) Vowel onset point detection using source, spectral peaks, and modulation spectrum energies. IEEE Trans Audio Speech Lang Process 17(4):556–565
11. Rao KS, Vuppala AK (2013) Non-uniform time scale modification using instants of significant excitation and vowel onset points. Speech Commun 55(6):745–756
12. Rao KS, Yegnanarayana B (2009) Duration modification using glottal closure instants and vowel onset points. Speech Commun 51(12):1263–1269
13. Schutte K, Glass JR (2005) Robust detection of sonorant landmarks. In: INTERSPEECH, pp 1005–1008
14. Vuppala AK, Rao KS (2013) Vowel onset point detection for noisy speech using spectral energy at formant frequencies. Int J Speech Technol 16(2):229–235
15. Vuppala AK, Rao KS, Chakrabarti S (2012) Improved vowel onset point detection using epoch intervals. AEU-IntJ Electron Commun 66(8):697–700
16. Vuppala AK, Yadav J, Chakrabarti S, Rao KS (2012) Vowel onset point detection for low bit rate coded speech. IEEE Trans Audio Speech Lang Process 20(6):1894–1903
17. Vydana HK, Vikash P, Vamsi T, Kumar KP, Vuppala AK (2015) Detection of emotionally significant regions of speech for emotion recognition. In: Annual IEEE India conference (INDICON), pp 1–6
18. Yadav J, Rao KS (2013) Detection of vowel offset point from speech signal. IEEE Signal Process Lett 20(4):299–302
19. Yegnanarayana B, Murty KSR (2009) Event-based instantaneous fundamental frequency estimation from speech signals. IEEE Trans Audio Speech Lang Process 17(4):614–624
20. Yegnanarayana B, Prasanna SM, Guruprasad S (2011) Study of robustness of zero frequency resonator method for extraction of fundamental frequency. In: IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 5392–5395

Ramakrishna Thirumuru received his B.Tech. in Electronics and Instrumentation Engineering from Kakatiya University, Telangana, India, in 1991, M.Tech. in Electronics and Communications Engineering from JNTU, Hyderabad, in 2003. Presently he is a research scholar in speech processing lab, LTRC at IIITH, Hyderabad, India. His area of interest is speech signal processing and its appications.

Dr. Suryakanth V. Gangashetty has joined to IIIT Hyderabad on 23rd August 2006 as an Assistant Professor. He has done his B.E (Computer Science and Engineering) from Govt.(BDT) College of Engineering Davangere (Now it is UBDTCE) in 1991, M.Tech (Systems Analysis and Computer Applications) from Karnataka Regional Engineering College Surathkal (Now it is NITK) in 1998 and Ph.D (Neural Network Models for Recognition of Consonant-Vowel Units of Speech in Multiple Languages) from Indian Institute of Technology Madras in 2005. Before joining to IIIT Hyderabad, he has worked as a Senior Project Officer at Speech and Vision Laboratory, IIT Madras. He has worked as a member of faculty at BIET Davangere karnataka, from 1991 to 1999. He has also worked as a visiting research scholar at OGI Portland (USA) for three months during the summer of 2001. He has done his post-doctoral studies (PDF) at Carnegie Mellon University (CMU) Pittsburgh (PA, USA) during April 2007 to July 2008. He is a author of about 115 papers published in national as well as international conferences and journals. He has co-authored four book chapters in edited volumes published by Springer and World Scientific publishing company. He is a life member of the CSI, IE, IUPRAI, ASI, IETE, ORSI, and ISTE. He has reviewed papers for reputed journals and conferences. His research interests include Speech Processing, Neural Networks, Multimedia Signal Processing, Pattern Recognition, Soft Computing, Machine Learning, Image Processing, Natural Language Processing, Artificial Intelligence, Fuzzy Logic.

Anil Kumar Vuppala received his B.Tech. in Electronics and Communications Engineering from JNTU, Hyderabad, India, in 2005, M.Tech. in Electronics and Communications Engineering from NIT, Kurukshetra, in 2007, and PhD in signal processing from IIT Kharagpur, in 2012. From March 2012 he is working as Assistant Professor at IIITHyderabad. His research interests lie primarily in speech processing in mobile and practical environments. He has published 2 book chapters, 13 reputed journals, and 34 international conference papers. He is currently handling 5 sponsored projects.