Vowel Onset Point Detection Using Source, Spectral Peaks, and Modulation Spectrum Energies

S. R. Mahadeva Prasanna, *Member, IEEE*, B. V. Sandeep Reddy, and P. Krishnamoorthy, *Student Member, IEEE*

***Abstract*—Vowel onset point (VOP) is the instant at which the onset of vowel takes place during speech production. There are significant changes occurring in the energies of excitation source, spectral peaks, and modulation spectrum at the VOP. This paper demonstrates the independent use of each of these three energies in detecting the VOPs. Since each of these energies represents a different aspect of speech production, it may be possible that they contain complementary information about the VOP. The individual evidences are therefore combined for detecting the VOPs. The error rates measured as the ratio of missing and spurious to the total number of VOPs evaluated on the sentences taken from the TIMIT database are 6.92%, 8.8%, 6.13%, and 4.0% for source, spectral peaks, modulation spectrum, and combined information, respectively. The performance of the combined method for VOP detection is improved by 2.13% compared to the best performing individual VOP detection method.**

***Index Terms*—Modulation spectrum and combining, source, spectral peaks, vowel onset point (VOP).**

I. Introduction

VOWELS are produced by keeping the vocal tract in an open position with minimum obstruction along the length and using glottal vibration as the excitation [1], [2]. The vowel type is determined by the shape of the vocal tract and is affected by the positioning of various articulators like the lips, jaw, and tongue. The beginning or onset of the vowel, more commonly termed as vowel onset point (VOP), is defined as the instant at which the onset of the vowel takes place [3]–[7]. VOP is an important event during speech production. This event marks the beginning of vowel and also the end of the consonant in case of consonant–vowel (CV) transitions [5]. Vowels are the major energy carriers. Thus, knowledge about the VOPs can be used in the detection of end-points of a speech utterance [8], [9]. The discriminatory information for speech recognition is present in a small region around the VOP in each of the CV units [10], [11]. Thus, VOPs can be used as anchor points to extract features selectively for speech recognition. There are significant changes occurring in the characteristics of the excitation source, vocal tract transfer function, and modulation components around the VOPs. Knowledge about the VOPs is useful in extracting information for tasks like speaker recognition, language identification, and expressive speech processing. As outlined above, knowledge about the VOPs is useful in many speech processing tasks and hence the motivation for the development of methods for the automatic detection of the VOPs [3]–[7].

There are several methods developed earlier for the detection of the VOPs. These include methods based on the appearance of rapidly increasing resonance peaks in the amplitude spectrum [3]; zero-crossing rate, energy, and pitch information [4]; wavelet transform [6]; energy derivative [5]; and neural network [4], [12]. More recently, a method using the excitation source information is proposed for the detection of the VOPs [13]. The performance of the existing methods depend on the amount of evidence available about the VOP in the selected speech feature. This in turn depends on the origin as well as method employed for the extraction of the feature. Hence, there may always be a limitation on the performance achieved using only one feature. For instance, in case of sound units like *ha* there may not be any change in the vocal tract shape at the VOP: the only change may be in terms of the excitation source [13]. Therefore, it may be beneficial to use two or more features, whose articulatory correlates are different. Accordingly the proposed work explores the energies in the excitation source, spectral peaks and modulation spectrum for the detection of the VOPs. The energy of the excitation source represents the changes in its energy levels, the energy of spectral peaks represents the changes in the vocal tract transfer function and the energy of modulation spectrum represents the changes in the slowly varying temporal envelope. Thus, these three energies are different and may carry independent evidence about the VOP. Initially each of these energies is independently explored to measure the amount of VOP information, then they are combined to detect the VOPs. Such a combination may further improve performance and may also provide robustness compared to any one energy [9]. This is the motivation for the present work.

As will be evident in the following sections, the novelty of the present work may be summarized as follows: Identifying the significance of different energies from speech for the detection of the VOP; method for the enhancement of VOP evidence in each speech feature; significantly revised version of our earlier method [13] for the VOP detection using the excitation source; VOP detection using energy of modulation spectrum; combination method for the VOP detection; and finally the different and extensive experimental studies conducted to validate the proposed VOP detection methods.

Manuscript received December 24, 2007; revised October 31, 2008. Current version published March 18, 2009. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Mark Hasegawa-Johnson.

S. R. Mahadeva Prasanna and P. Krishnamoorthy are with the Department of Electronics and Communication Engineering, Indian Institute of Technology Guwahati, Guwahati 781039, Assam, India (e-mail: prasanna@iitg.ernet.in; pkm@iitg.ernet.in; pkmkicha@yahoo.co.in).

B. V. Sandeep Reddy is with the Automatic Speech Recognition (ASR) Systems in Applications Technology, Inc., Noida 201301, India (e-mail: sandeep.reddy@glocaledge.com).

Digital Object Identifier 10.1109/TASL.2008.2010884

1558-7916/$25.00 © 2009 IEEE

The rest of the paper is organized as follows: The method for detection of the VOPs using the energy of the excitation source is described in Section II. Detection of the VOPs using the energy of the spectral peaks is explained in Section III. The method for detection of the VOPs using the energy of the modulation spectrum is described in Section IV. Section V describes the combination method for the detection of the VOPs. The experimental results and discussions related to the proposed methods are given in Section VI. Summary and conclusions of the present work and scope for the future work are mentioned in Section VII.

II. VOP Detection Using Excitation Source Energy

Speech is produced as a result of a time-varying excitation of a time-varying vocal tract system [1], [14]. The time-varying excitation characteristics include changes in the nature of excitation from unvoiced to voiced, level of voiced energy, and also associated periodicity. One or more of these changes may occur at the VOP [7], [13]. This aspect may be exploited for detection of the VOP using excitation source information. In this paper, the change in the energy level is exploited for detection of the VOPs.

The excitation source information from the speech signal is extracted using the linear prediction (LP) analysis [15]. The speech signal is processed in blocks of 20 ms with 10-ms shift. For each block of 20 ms, tenth-order LP analysis (speech is sampled at $F_s = 8$ kHz) is performed to estimate the LP coefficients (LPCs). The time-varying inverse filter is constructed using these LPCs. The speech signal is passed through the inverse filter to extract the LP residual signal. For the proper choice of LP order, the LP residual mostly contains the excitation source information.

The time-varying changes in the excitation characteristics are smeared in the LP residual due to its bipolar nature [16]. These changes are further enhanced by computing the Hilbert envelope of the LP residual [7], [16]. The Hilbert envelope (HE) $h_e(n)$ of the LP residual $e(n)$ is defined as [17]

$$h_e(n) = \sqrt{e^2(n) + e_h^2(n)} \tag{1}$$

where $e_h(n)$ is the Hilbert transform of $e(n)$, and is given by

$$e_h(n) = IDFT\left[E_h(k)\right] \tag{2}$$

where

$$E_h(k) = \begin{cases} -jE(k), & k = 0, 1, \ldots, \left(\frac{N}{2}\right) - 1 \\ jE(k), & k = \left(\frac{N}{2}\right), \left(\frac{N}{2}\right) + 1, \ldots, (N-1) \end{cases} \tag{3}$$

where IDFT denotes the inverse discrete Fourier transform and $E(k)$ is computed as the discrete Fourier transform (DFT) of $e(n)$ and N is the number of points used for computing the DFT.

The speech signal of nasal sound unit *na*, its LP residual and HE of the LP residual are shown in the Fig. 1(a)–(c), respectively. By comparing the plots in the Fig. 1(b) and (c), it can be observed that the change in the excitation source characteristics is relatively more prominent in the HE of the LP residual.

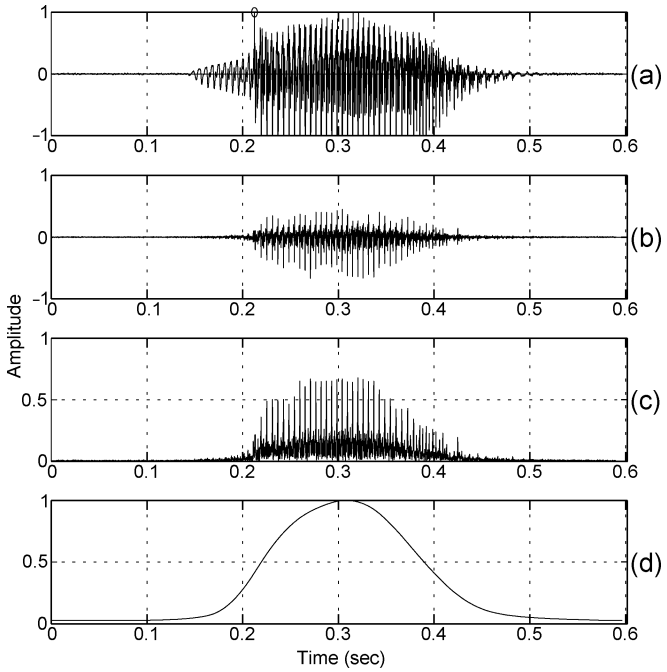


Fig. 1. Excitation source energy for the VOP detection. (a) Speech signal of *na*. (b) LP residual. (c) HE of the LP residual. (d) Smoothed HE of the LP residual.

Hence, HE of the LP residual is used as the excitation source information in the present work.

There are changes in the excitation characteristics both at the fine and gross levels during speech production. For instance, the fine level change may be from closed to open phase in a pitch period, and the gross level change may be in the energy level. The VOPs are the events associated with the changes at the gross level. In case of *na* shown in Fig. 1, even though there are many changes in the excitation characteristics as displayed in Fig. 1(c) by the HE of the LP residual, only the change at the onset of the vowel is of interest. The changes at the fine level therefore need to be smoothed by convolving the HE of the LP residual using a Hamming window of 50 ms (400 samples at $F_s = 8$ kHz). The smoothed HE of the LP residual is shown in Fig. 1(d). The change at the VOP is easily located in this plot, compared to the HE of the LP residual. In this paper, the smoothed HE of the LP residual is used as the representation of excitation source energy.

A. Enhancement of VOP Evidence

The change at the VOP available in the smoothed HE of the LP residual is further enhanced by computing its slope with the help of a first-order difference (FOD). The steps involved in the enhancement of change at the VOP are explained with the help of Fig. 2 for the speech of *Don't ask me to carry an oily rag like that*, taken from the TIMIT database [18]. Since FOD represents the slope, the positive to negative going zero transition in FOD locates the peaks in the smoothed HE of the LP residual. This is illustrated in Fig. 2(a)–(c). The smoothed HE of the LP residual and its FOD values are shown in Fig. 2(b) and (c), respectively. The positive to negative going zero transition points and the corresponding local peaks are represented by star (*) symbols in Fig. 2(b) and (c). The unwanted zero crossings detected at the

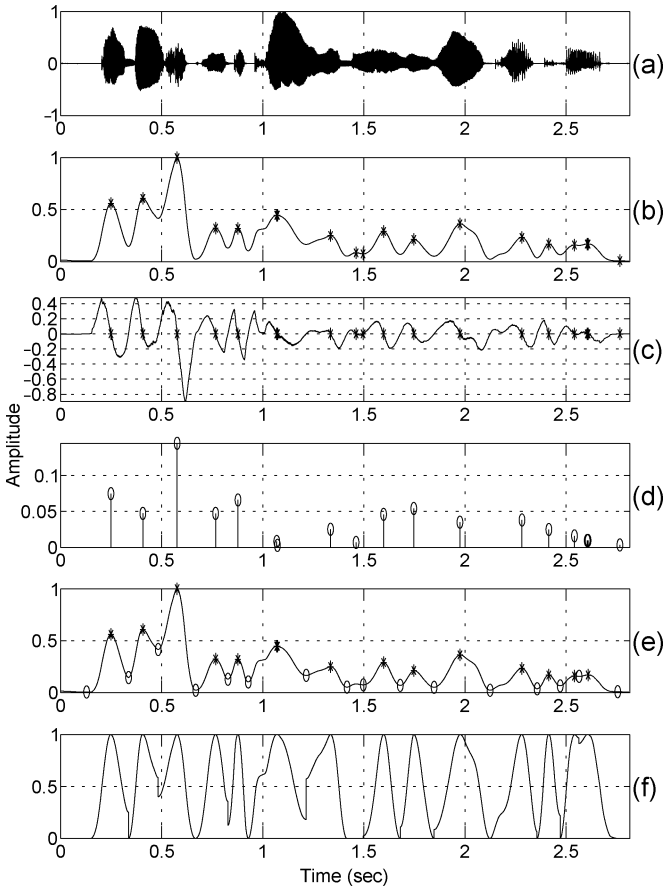


Fig. 2. Enhancement of high energy regions. (a) Time-domain waveform for the speech of *Don't ask me to carry an oily rag like that*. (b) Smoothed HE of the LP residual. (c) First-order difference (FOD) values. (d) Slope values computed at each peak location. (e) Smoothed HE of the LP residual with high SNR region locations. (f) Enhanced values.

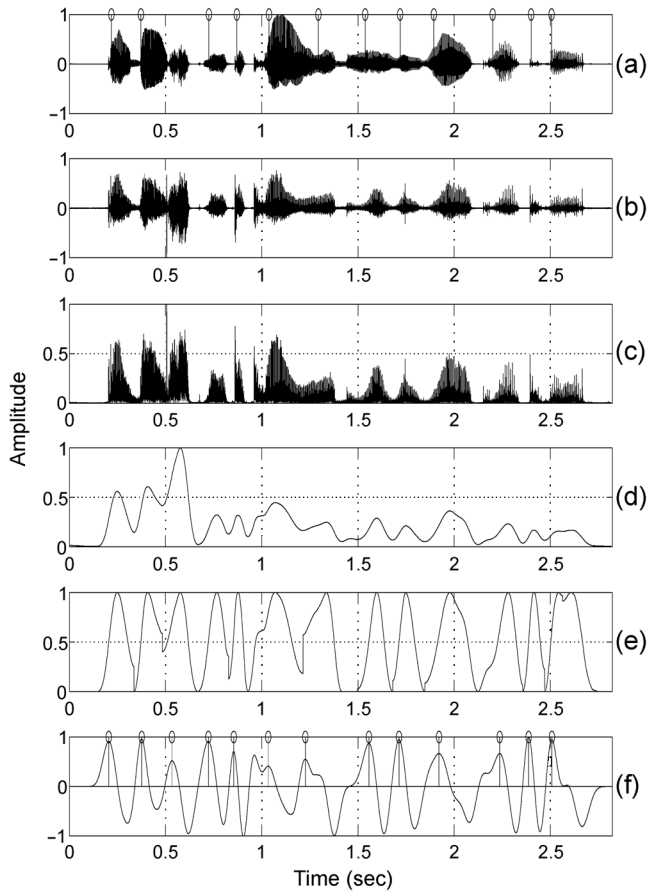


Fig. 3. VOP detection using energy of excitation source for the speech of *Don't ask me to carry an oily rag like that*. (a) Time-domain waveform. (b) LP residual. (c) HE of the LP residual. (d) Smoothed HE of the LP residual. (e) Enhanced version of the smoothed HE. (f) VOP evidence plot. Manually marked VOPs are shown in (a) and hypothesized VOPs are shown in (f).

low SNR regions are eliminated by finding the sum of slope values for a duration 10 ms (80 samples for $F_s = 8$ kHz) centered at each positive to negative going zero crossing point; these slope values are given in Fig. 2(d). The peaks with the lower slope values are eliminated with a threshold set equal to 0.5 times the mean value of the slope. This threshold has been tested only for the clean speech signals of length 2.5–5 s. In the next step, if two successive peaks happen to be within 50 ms, then the peak with lower value is eliminated, i.e., we assume that there will rarely be two VOPs within a 50-ms interval. The star $(^*)$ symbols in Fig. 2(e) show the peak locations after eliminating the undesirable peaks. With respect to each of these local peaks, the nearest negative to positive going zero transition points on either side are identified and marked by circles in Fig. 2(e). Segments bounded by negative to positive going zero transition points on either side are enhanced by normalizing so that the peak smoothed HE in each such region has an amplitude of 1.0, as shown in Fig. 2(f). As it can be observed, the change at the VOP is further enhanced by processing the smoothed HE of the LP residual.

The significant changes in the excitation characteristics present in the enhanced version of the smoothed HE of the LP residual are detected by convolving with a first-order Gaussian differentiator (FOGD) of length 100 ms [7], [13]. The choice of length of the FOGD operator is based on the assumption that the VOPs occur as gross level changes at intervals of about 100 ms [7]. The peaks in the convolved output represent the locations of the VOPs and are selected by using a simple peak-picking algorithm with small threshold to eliminate spurious peaks. The convolved output is termed as the **VOP Evidence Plot using Source**.

The robustness of this approach may be seen in the detection of VOPs in continuous speech, where changes at different VOPs are different due to the time-varying characteristics of excitation source. The speech for *Don't ask me to carry an oily rag like that*, taken from the TIMIT database is shown in Fig. 3. The LP residual, HE of the LP residual, smoothed version of the HE of the LP residual, enhanced version of smoothed HE of the LP residual and the VOP evidence plot computed are shown in Fig. 3. The VOP evidence plot shows peaks even under the time-varying characteristics of excitation source. This can be observed by comparing the manually marked VOPs in Fig. 3(a) with the peaks in the VOP Evidence Plot. There are 12 VOPs and the proposed excitation source energy based method hypothesizes 13 VOPs. Among the 13 VOPs hypothesized, 12 are true VOPs and one is a spurious VOP. The logic used in all the VOP detection methods is that for a peak to be valid as VOP, its value should be above certain threshold and also the peak should be

TABLE I
PERFORMANCE OF PROPOSED VOP DETECTION METHODS. THE DATABASE CONSISTS OF 750 VOPS. IN THE TABLE ABBREVIATIONS EXC, VT, MOD, COMB, AND HERMES REFER TO EXCITATION SOURCE, SPECTRAL PEAKS, MODULATION SPECTRUM, COMBINED VOP DETECTION AND HERMES VOWEL ONSET DETECTION METHODS, RESPECTIVELY. SIMILARLY, ABBREVIATIONS HYPO, DET, MISS, AND SPU REFER TO HYPOTHESIZED, DETECTED, MISSING AND SPURIOUS VOPS, RESPECTIVELY. TER REFERS TO THE TOTAL ERROR RATE WHICH IS THE SUM OF MISSING AND SPURIOUS VOPS. ALL ARE EXPRESSED AS % OF THE RATIO OF RESPECTIVE VOPS TO THE TOTAL NUMBER OF VOPS (750)

Method	HYPO	DET VOPs (within ms)				MISS	SPU	TER
	VOPs	± 10	± 20	± 30	± 40	VOPs	VOPs	
EXC	756	41.07	57.6	75.2	96.93	3.07	3.86	6.92
VT	782	35.3	58.5	76.3	97.73	2.27	6.53	8.80
MOD	728	44.6	63.7	77.6	95.47	4.53	1.60	6.13
COMB	742	49.8	64.3	78.8	97.46	2.54	1.46	4.0
HERMES	559	24.93	37.86	57.73	69.07	30.93	5.47	36.4

followed by a negative region. The negative region is a characteristic of following vowel region. For instance, in Fig. 3(f), there are two peaks around time instant $t = 1$ s. As per the peak value, the larger peak should be hypothesized as VOP, but since it does not have the negative region following, the other smaller peak is chosen.

For the initial assessment of this algorithm based on the excitation source energy, speech signals of 30 speakers are randomly selected from the test set of the TIMIT database. The VOPs present in them are noted from the vowel durations marked in the associated transcription files. The two sentences taken are *Don't ask me to carry an oily rag like that* and *she had your dark suit in greasy wash water all year*. There are 25 VOPs in these two sentences. Therefore, in total there are 750 VOPs for the 30 speakers. The speech signals are processed to extract the VOPs using excitation source energy as described earlier. The performance of this method is given in Table I.

The ratio (in %) of number of true VOPs detected to the total number of VOPs are measured initially for different time resolutions. The percentage of true VOPs that are matched by a detected VOP within $+/-T$ ms is called the detection rate (DET) with a resolution of T. The percentage of missing true VOPs after the maximum resolution of ± 40 ms is termed as missing rate (MISS). The percentage of spurious VOPs is termed as spurious rate (SPU). Both these are errors in the VOP detection and hence sum of these two error rates is termed as total error rate (TER).

Detection rate is 41.07% for ± 10-ms resolution and increases to 96.93% for ± 40-ms resolution. The performance at $+/-10$ ms may not be sufficient for applications that rely on accurate consonant-vowel transitions, e.g., neural networks trained to classify consonant-vowel transitions [5], [19]. The present method needs to be improved to increase the time resolution. Apart from this, detection rate of 96.93% at ± 40 ms may be sufficient in applications like end-point detection and identification of voiced/unvoiced regions.

III. VOP DETECTION USING SPECTRAL PEAKS ENERGY

Speech is produced as a sequence of sound units. These sound units are produced as a result of changes in the vocal tract shape. Distinct vocal tract shapes are associated with the production of vowels. The vocal tract shape is manifested in the spectrum of the speech signal. The spectrum of the speech signal will have amplitudes of the formants representing the vocal tract shape as spectral peaks and the pitch and its harmonics representing the excitation. The amplitudes of the formants may be estimated by picking some of the largest peaks in the spectrum. Since the objective is only to have gross information about the vocal tract shape, we have used only the ten largest peaks.

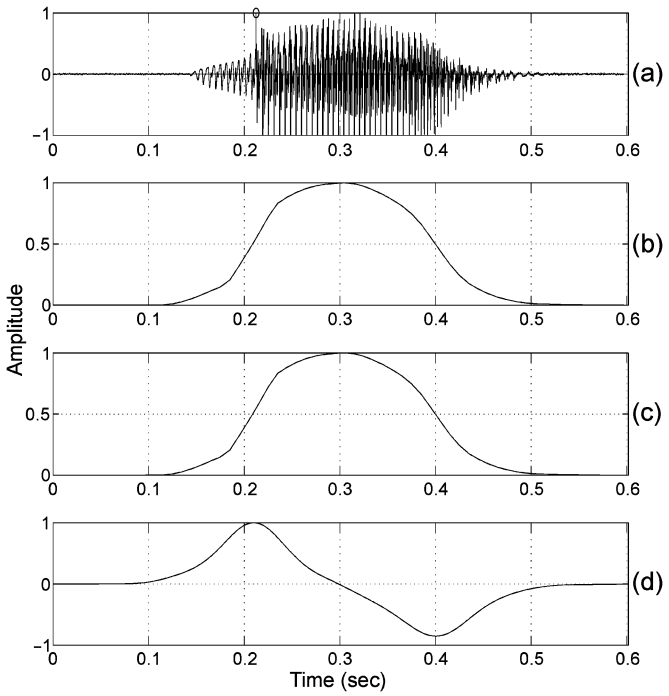


Fig. 4. VOP detection using energy of spectral peaks. (a) Speech signal of *na*. (b) Sum of ten largest peaks in the DFT spectrum. (c) Enhanced sum of ten largest peaks in the DFT spectrum. (d) VOP evidence plot.

The speech signal is processed in blocks of 20 ms with a shift of 10 ms. For each block of 20 ms, a 256-point DFT is computed, and the ten largest peaks are selected from the first 128 points. The sum of these amplitudes, plotted as a function of time, is used as the representation of energy of spectral peaks. The sound unit *na* and its sum of ten peaks in the DFT spectrum are shown in Fig. 4(a) and (b), respectively. The onset of the vowel can be observed as significant change in the sum of ten peaks in the DFT spectrum. The enhanced version of the same is shown in Fig. 4(c). These enhanced values are convolved with the FOGD operator and the convolved output is the **VOP Evidence Plot using Spectral Peaks**. The peaks in the VOP evidence plot indicate the possible VOP locations and are shown in Fig. 4(d).

The robustness of this approach may be seen in the detection of VOPs in continuous speech, where changes at the VOPs are different due to time varying nature of vocal tract shape. The speech of *Don't ask me to carry an oily rag like that*, taken from the TIMIT database is shown in Fig. 5(a). The sum of ten peaks in the DFT spectrum, its enhanced version and VOP evidence plot are shown in Fig. 5. The VOP evidence plot shows peaks at the VOPs, even under the time varying characteristics of the DFT spectrum. There are 12 VOPs and the proposed vocal

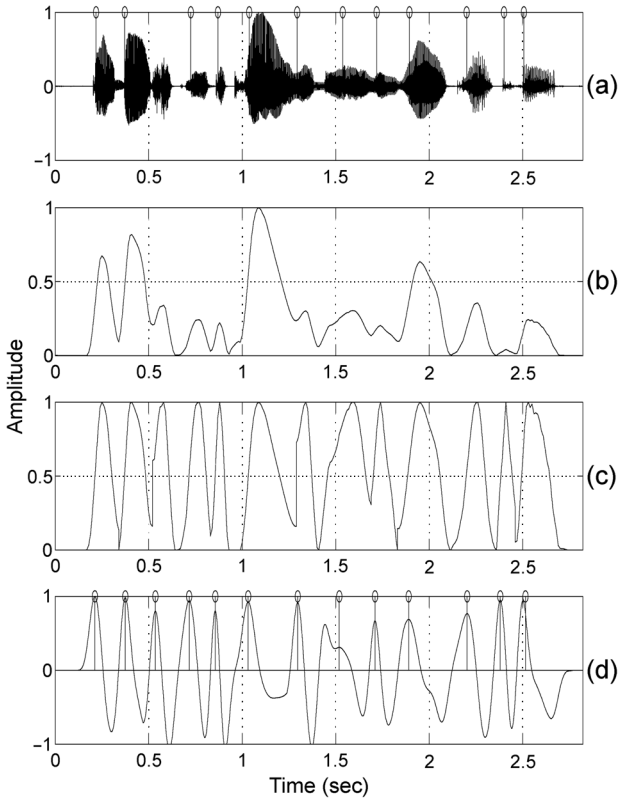


Fig. 5. VOP detection using energy of spectral peaks for the speech of *Don't ask me to carry an oily rag like that*. (a) Time-domain waveform. (b) Sum of ten largest peaks in the DFT spectrum. (c) Enhanced sum of ten largest peaks in the DFT spectrum. (d) VOP evidence plot.

tract feature based method hypothesizes 13 VOPs. Among the 13 VOPs hypothesized, 12 are the true VOPs, and one is the spurious VOP.

For the initial assessment of this algorithm based on vocal tract shape, the same database of 30 speakers considered in the earlier case is used. The speech signals are processed to extract the VOPs using energy of spectral peaks. The performance is given in Table I. This method also shows same performance trend as in the earlier case. The difference in the actual performance and error values can be attributed to the nature of VOP information present in the vocal tract shape, that is, sum of ten peaks in the DFT spectrum.

IV. VOP Detection Using Modulation Spectrum Energy

Modulation components refer to the slowly varying temporal envelope components in speech [20]. The temporal envelope of speech is dominated by low-frequency components of several Hz. A representation of this type has compelling parallels to the dynamics of speech production, in which the articulators move at rates of 2–12 Hz [21], and to the sensitivity of auditory cortical neurons to amplitude modulations at rates below 20 Hz. Several studies have been conducted earlier to demonstrate the significance of modulation spectrum in speech processing [22], [23]. A formal and rigorous focus on the modulation spectrum including the development of the modulation spectrogram has been demonstrated by Greenberg *et al.* [20], [24]. It is also demonstrated that the concentration of energy in the modulation spectrum corresponds to the syllable nuclei, that is, vowel. Thus, the onset of vowel may be manifested as a significant change in the modulation spectrum energy level in the 4–16 Hz band.

The modulation spectrum energy from the given speech signal is extracted as follows [20], [24]. The speech signal is analyzed into approximately 18 critical band filters between 0 and 4 kHz. The filters are trapezoidal in shape, and there is minimal overlap between adjacent bands. In each band, an amplitude envelope signal is computed by halfwave rectification and low pass filtering with cutoff frequency of 28 Hz. Each amplitude envelope signal is then downsampled to 80 samples/s and normalized by the average envelope level in that channel, measured over the entire utterance. The modulations of the normalized envelope signals are analyzed by computing the DFT over 250-ms Hamming windows with shift of 12.5 ms, in order to capture the dynamic properties of the signal. Finally, the 4–16 Hz components are added together, across all critical bands.

Mathematically the modulation transfer function energies are expressed as

$$m(i) = \sum_{p=1}^{18} \left[\sum_{k=k_1}^{k=k_2} \left| \hat{X}_p(k,i) \right|^2 \right] \tag{4}$$

where i is the frame index, p represents the critical band number, and k_1 and k_2 represent frequency index of 4 Hz and 16 Hz, respectively. $\hat{X}_p(k)$ is computed as

$$\hat{X}_p(k) = \sum_{n=0}^{N-1} \hat{x}_p(n) w(n) e^{-\frac{j2\pi nk}{N}}; \quad p = 1, 2, \ldots, 18. \tag{5}$$

where $\hat{x}_p(n)$ represents the normalized envelope of p^{th} filter output, $w(n)$ is a Hamming window, and N is the number of points used for computing the DFT. The modulation energy components computed for each frame are then upsampled to 8000 samples/s and plotted as function of time.

The sound unit *na* and its modulation spectrum energy are shown in Fig 6(a) and (b), respectively. The onset of vowel can be observed as significant change in the modulation spectrum energy. The change may be further enhanced by computing the slope and is shown in Fig. 6(c). The significant change may be detected by convolving the same using FOGD operator of length 100 ms and is shown in Fig. 6(d). The convolved output is the **VOP Evidence Plot using Modulation Spectrum**. The peak in the VOP evidence plot selected on a threshold basis indicates the location of the VOP.

The robustness may be seen in the detection of the VOPs in continuous speech, where changes at the VOPs are different due to time varying characteristics of modulation spectrum energy. The speech of *Don't ask me to carry an oily rag like that*, taken from the TIMIT database is shown in Fig. 7(a). The modulation spectrum energy, its enhanced version and VOP evidence plot are shown in Fig. 7. The VOP evidence plot shows peaks at the VOPs, even under time-varying characteristics of the modulation spectrum energy. There are 12 VOPs and the proposed method hypothesizes 14 VOPs. Among the 14 VOPs hypothesized, 12 are true VOPs and 2 are spurious VOPs. For the initial

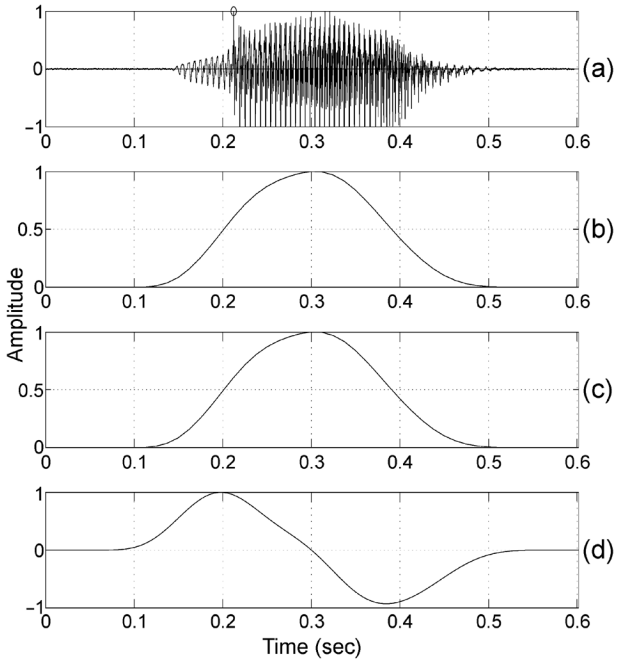


Fig. 6. VOP detection using energy of modulation spectrum. (a) Speech signal of *na*. (b) Modulation spectrum energy. (c) Enhanced modulation spectrum energy value. (d) VOP evidence plot.

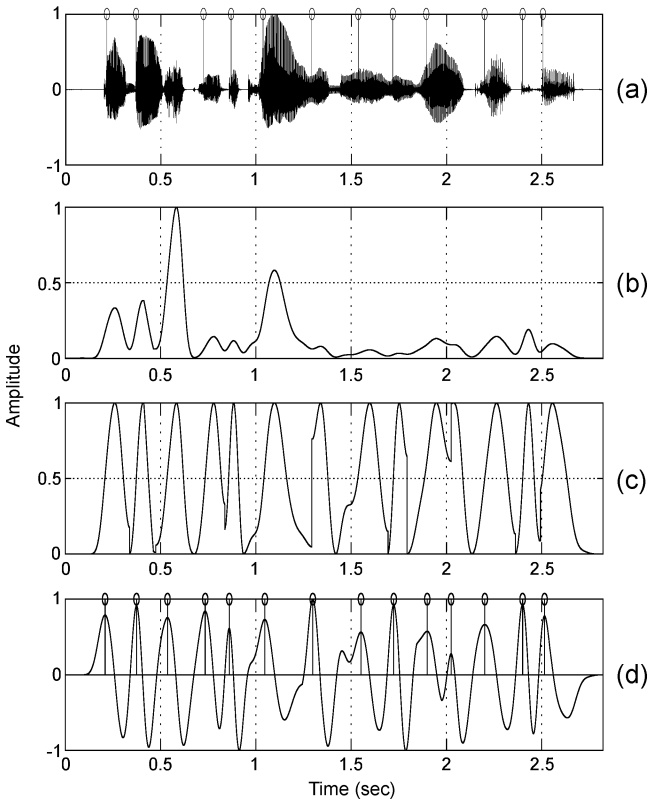


Fig. 7. VOP detection using energy of modulation spectrum for the speech of *Don't ask me to carry an oily rag like that*. (a) Time-domain speech waveform. (b) Modulation spectrum energy. (c) Enhanced modulation spectrum energy values. (d) VOP evidence plot.

assessment and comparison, the same database of 30 speakers used earlier is considered. The speech signals are processed to

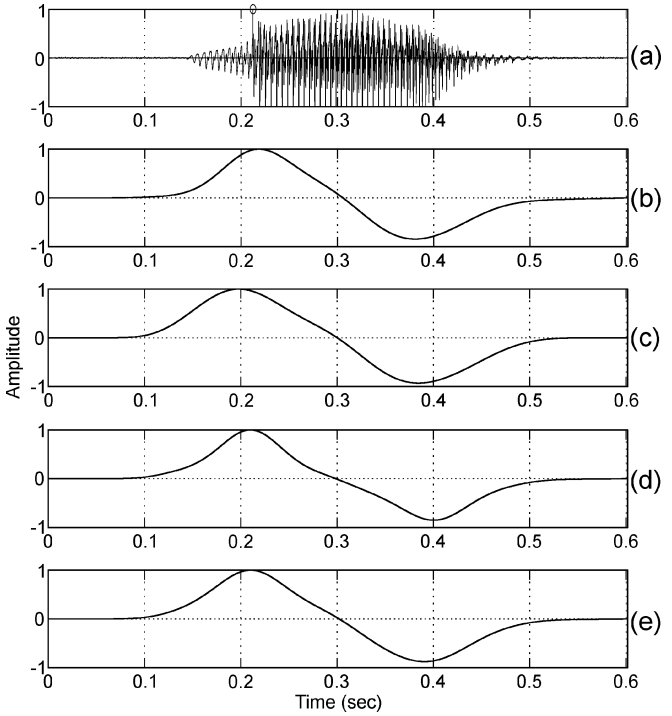


Fig. 8. VOP detection using combination of all three evidences. (a) Speech signal of *na*. VOP evidence plot for (b) excitation source. (c) Modulation spectrum. (d) Spectral peaks. (e) Combined VOP evidence plot.

extract the VOPs using modulation spectrum energy. The performance of this method is given in Table I. This method shows the same performance trend as in the earlier cases. The difference in the actual performance and error values can be attributed to the nature of VOP information present in the modulation spectrum energy.

V. VOP Detection by Combining Evidences From Energies of Source, Spectral Peaks, and Modulation Spectrum

The HE of the LP residual mostly represents the excitation source information. The sum of ten largest peaks in the DFT spectrum represents the vocal tract shape. The modulation spectrum represents the slowly varying temporal envelope. Thus, the origin of each of these features is different. However, each of them carry information about the VOPs, as demonstrated by their nearly equal performance. As analyzed earlier, the distribution pattern of detected VOPs at different resolutions, number of spurious and missing VOPs are different in each case. This implies that the evidence about the VOPs available in each of them may be different and hence can be combined. The combined evidence may have stronger and robust information about the VOPs. This may also lead to improved performance.

The speech signal for nasal sound *na* and the VOP evidence plots for excitation source, modulation spectrum and spectral peaks are shown in Fig. 8(a)–(d), respectively. The change in the nature of VOP evidence plots indicate different VOP evidences available in each case. This can be further seen better by considering continuous speech signal for *Don't ask me to carry an oily rag like that*, taken from the TIMIT database. The speech signal and the VOP evidence obtained in each case are shown in

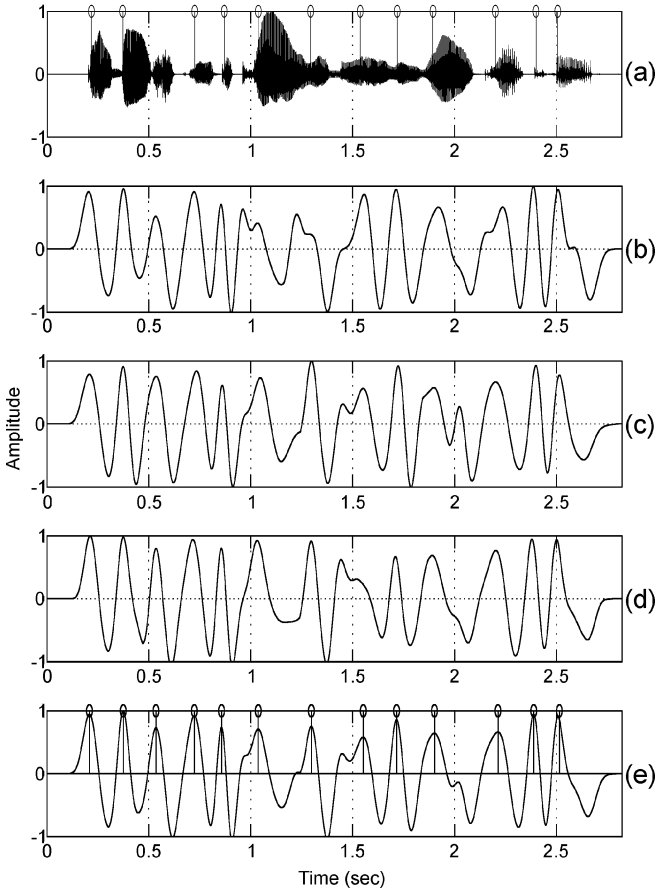


Fig. 9. VOP detection using combination of all three evidences for the speech of *Don't ask me to carry an oily rag like that*. (a) Time-domain waveform. VOP evidence plot for (b) excitation source. (c) Modulation spectrum. (d) Spectral peaks. (e) Combined VOP evidence plot.

Fig. 9. The evidences at each of the VOPs are different in each case.

The VOP evidence from all the three cases may have strong or weak evidence at the true VOPs, no or weak evidence at the spurious VOPs. This leads to correct detection of most true VOPs, missing few true VOPs, and also hypothesizing a smaller number of spurious VOPs. Since the evidence about the VOP is strong in each case, all the true VOPs may have strong evidence, whereas spurious VOPs may not have the same level of evidence due to differences among the evidence plots. This aspect may be exploited in combining the evidences. The combined VOP evidence plot is obtained by adding the three evidence plots and normalizing so that the maximum is 1.0. The **Combined VOP Evidence Plot** thus obtained is shown in Fig. 9(e). There are 12 VOPs and the proposed combination method hypothesizes 13 VOPs. Among the 13 VOPs hypothesized, 12 are true VOPs and one is a spurious VOP.

For comparison of the combination method with the individual methods, the same database of 30 speakers considered in the earlier cases is used. The speech signals are processed to extract the VOP evidences from each of the features and are added and normalized to obtain the combined VOP evidence. The performance of this method is given in Table I. Vocal tract energy has the lowest missing rate, modulation spectral energy has the lowest spurious detection rate, and excitation energy balances the two. Alternatively, the combination method provides nearly minimum errors, both in terms of spurious as well as missing, and hence the total error is minimum. This result demonstrates the complementary and independent evidence available in each of the speech features chosen for the study.

VI. Experimental Results and Discussions

The initial analysis of different VOP detection methods is carried out in the previous section using a database of 30 speakers having two sentences per speaker and containing 750 VOPs. The speech signals are processed to extract the VOPs using energy of excitation source, vocal tract shape and modulation spectrum. The performance of each of the methods are shown in Table I. Among the three features, the energy of spectral peaks provides relatively better performance in terms detecting the true VOPs. However, the number of spurious VOPs hypothesized are also high. Ideally, there should not be any missing VOPs and also no spurious VOPs. Hence, the total error of missing as well as spurious VOPs may be considered as the parameter for evaluating the performance. The Total Error Rate (TER) for different VOP detection methods are shown in Table I. According to the TER criterion, the energy of modulation spectrum provides relatively better performance. The TER of the combined VOP evidence is also given in Table I. TER shows an improvement of 2.13% compared to the best performing individual system. This is mainly due to the significant reduction of the spurious VOPs and demonstrates the significance of combining evidences for the detection of the VOPs.

In this section, we further analyze and evaluate the ability of the proposed VOP detection methods by conducting some more experiments. In the first experiment, the proposed methods are compared with one of the existing methods in the literature. For this, we have implemented the Hermes vowel onset detection method based on the measurement of vowel strength in terms of rapidly increasing resonance peaks in the amplitude spectrum [3]. The performance of this method for the 30-speaker database having 750 VOPs is given in Table I. The results indicate that the proposed methods provide significantly higher performance than the one reported in [3].

In the second experiment, the database size is increased from 30 to 168 speakers, which constitutes all test speakers in the TIMIT database. The same two sentences taken earlier are considered in this case also. Therefore, there are 25 VOPs per speaker and hence a total of 4200 VOPs. The performance of the proposed VOP detection methods are given in Table II. The same trend as observed in 30 speakers is also observed in this case. All the methods provide nearly the same performance. The combined method provides an improvement by 2% compared to the best performing individual system. This shows the robustness of the proposed features for VOP detection with reference to the increase in database size.

To know about the gender-dependent performance of the proposed methods, the database of 168 speakers is segregated into male and female speakers and the VOPs are computed separately. In this database, there are 112 male speakers and 56 female speakers. The performance of proposed VOP detection methods for 112 male speakers having a total of 2800 VOPs is given in Table III. The performance for 56 female speakers

TABLE II
PERFORMANCE OF PROPOSED VOP DETECTION METHODS FOR INCREASED DATABASE SIZE. THE DATABASE CONSISTS OF 4200 VOPS. IN THE TABLE ABBREVIATIONS EXC, VT, MOD, AND COMB REFER TO EXCITATION SOURCE, SPECTRAL PEAKS, MODULATION SPECTRUM, AND COMBINED VOP DETECTION METHODS, RESPECTIVELY. SIMILARLY, ABBREVIATIONS HYPO, DET, MISS, AND SPU REFER TO HYPOTHESIZED, DETECTED, MISSING, AND SPURIOUS VOPS, RESPECTIVELY. TER REFERS TO THE TOTAL ERROR RATE WHICH IS THE SUM OF MISSING AND SPURIOUS VOPS. ALL ARE EXPRESSED AS % OF THE RATIO OF RESPECTIVE VOPS TO THE TOTAL NUMBER OF VOPS (4200)

Method	HYPO	DET VOPs (within ms)				MISS	SPU	TER
	VOPs	$\pm$ 10	$\pm$ 20	$\pm$ 30	$\pm$ 40	VOPs	VOPs	
EXC	4219	37.4	51.2	63.2	95.60	4.40	4.86	9.26
VT	4256	29.3	57.2	76.6	95.43	4.57	5.90	10.48
MOD	3977	38.3	47.7	69.3	92.62	7.38	2.07	9.45
COMB	4170	52.1	63.1	73.1	96.02	3.26	3.98	7.27

TABLE III
PERFORMANCE OF PROPOSED VOP DETECTION METHODS FOR 112 MALE SPEAKERS. THE DATABASE CONSISTS OF 2800 VOPS. IN THE TABLE ABBREVIATIONS EXC, VT, MOD, AND COMB REFER TO EXCITATION SOURCE, SPECTRAL PEAKS, MODULATION SPECTRUM, AND COMBINED VOP DETECTION METHODS, RESPECTIVELY. SIMILARLY, ABBREVIATIONS HYPO, DET, MISS, AND SPU REFER TO HYPOTHESIZED, DETECTED, MISSING, AND SPURIOUS VOPS, RESPECTIVELY. TER REFERS TO THE TOTAL ERROR RATE WHICH IS THE SUM OF MISSING AND SPURIOUS VOPS. ALL ARE EXPRESSED AS % OF THE RATIO OF RESPECTIVE VOPS TO THE TOTAL NUMBER OF VOPS (2800)

Method	HYPO	DET VOPs (within ms)				MISS	SPU	TER
	VOPs	$\pm$ 10	$\pm$ 20	$\pm$ 30	$\pm$ 40	VOPs	VOPs	
EXC	2791	37.8	48.4	64.6	95.43	4.57	4.25	8.82
VT	2770	31.4	64.0	78.6	94.82	5.18	4.11	9.29
MOD	2652	39.5	43.1	61.1	93.07	6.93	1.64	8.57
COMB	2723	50.7	65.7	76.4	95.75	1.50	4.25	5.75

TABLE IV
PERFORMANCE OF PROPOSED VOP DETECTION METHODS FOR 56 FEMALE SPEAKERS. THE DATABASE CONSISTS OF 1400 VOPS. IN THE TABLE ABBREVIATIONS EXC, VT, MOD, AND COMB REFER TO EXCITATION SOURCE, SPECTRAL PEAKS, MODULATION SPECTRUM, AND COMBINED VOP DETECTION METHODS, RESPECTIVELY. SIMILARLY, ABBREVIATIONS HYPO, DET, MISS, AND SPU REFER TO HYPOTHESIZED, DETECTED, MISSING, AND SPURIOUS VOPS, RESPECTIVELY. TER REFERS TO THE TOTAL ERROR RATE WHICH IS THE SUM OF MISSING AND SPURIOUS VOPS. ALL ARE EXPRESSED AS % OF THE RATIO OF RESPECTIVE VOPS TO THE TOTAL NUMBER OF VOPS (1400)

Method	HYPO	DET VOPs (within ms)				MISS	SPU	TER
	VOPs	$\pm$ 10	$\pm$ 20	$\pm$ 30	$\pm$ 40	VOPs	VOPs	
EXC	1428	36.6	56.7	60.4	95.93	4.07	6.07	10.14
VT	1486	25.1	43.8	72.4	96.64	3.36	9.50	12.86
MOD	1325	35.9	57.0	85.6	91.71	8.29	2.93	11.21
COMB	1447	54.9	58.0	66.4	97.57	5.78	2.42	8.20

having a total of 1400 VOPs is given in Table IV. In both the cases the different VOP detection methods show the same trend as observed in the earlier cases. That is, all individual methods provide nearly the same performance and the combined method shows an improvement by about 2%. Slight poor performance in the case of female speakers may be attributed to the relative poor performance of individual VOP detection methods.

The next experiment conducted is to analyze the performance of proposed VOP detection methods in different categories of sound units. The different categories considered are stop, fricative, affricate, nasal, and semivowel CV units. The 30-speaker database used earlier in the initial evaluation is considered for this study. In this database, there are 750 VOPs. These VOPs are segregated according to the different sound unit categories mentioned above. Out of 750 VOPs we have 148 stops, 150 fricatives, 34 affricates, 87 nasals, and 331 semivowels. The same trend is observed across different VOP detection methods. Due to the space constraint, the performance only for the combination method, for each of these categories is given in Table V. The sound class categories, in order of decreasing detection rate at $+/-40$-ms temporal resolution, is: fricative, stop, nasal, affricate, semivowel. In case of fricatives, stops, and nasals, there is a significant difference in the signal characteristics of the region preceding VOP, compared to the following region. Therefore, the changes at VOP are prominent and the performance is relatively better. The poor performance in case of semivowels and affricates may be attributed to the similarity of the signal characteristics preceding and following VOP. Future work should focus on exploring methods to improve performance in the semivowel and affricate categories of sound units, so that the overall performance improves. Since the VOP detection of each sound unit is done in an isolated fashion, there are no spurious VOPs and hence TER is not computed.

TABLE V
DETECTION RATE OF COMBINED VOP DETECTION METHOD FOR DIFFERENT CATEGORIES OF SOUND UNITS. THE DATABASE CONSISTS OF 750 VOPS. AMONG THESES WE HAVE 148 STOP, 150 FRICATIVES, 34 AFFRICATES, 87 NASALS, AND 331 SEMIVOWELS. IN THE TABLE, ABBREVIATIONS STP, FRI, AFF, NAS, AND SEM REFER TO STOPS, FRICATIVES, AFFRICATES, NASALS, AND SEMIVOWELS, RESPECTIVELY. ALL ARE EXPRESSED AS % OF THE RATIO OF RESPECTIVE VOPS TO THE TOTAL NUMBER OF VOPS OF THAT CATEGORY

Category	DET VOPs (within ms)			
	$\pm$ 10	$\pm$ 20	$\pm$ 30	$\pm$ 40
STP	48.6	68.9	81.1	98.64
FRI	67.3	84.0	99.3	99.33
AFF	23.5	35.3	79.4	97.05
NAS	65.5	70.1	78.2	97.70
SEM	40.5	53.4	67.4	96.07

The next experiment conducted is to analyze the performance of VOP detection methods in difficult cases of continuous speech from the TIMIT database. The sentences considered are *Where were you while we were away* and *The triumphant warrior exhibited naive heroism*. These sentences are spoken only by seven speakers in the whole database, and in total we have 173 VOPs. The performance of different VOP detection methods is given in Table VI. The poor performance compared to the earlier cases may be attributed to the highly similar signal characteristics both preceding and following regions of the VOP and also highly confusable and semivowel sound units present in both the sentences. However, the combined VOP detection method shows an improved performance by about 2.75%.

Finally, to observe the performance of the VOP detection methods with respect to accent, speech signals are collected for another 30 speakers from the laboratory environment. The sentences are *write a letter to krishna* and *he wrote an opinion essay to rama*. These sentences have Indian English accent. The manual VOPs are marked by carefully observing different signal characteristics like energy, pitch, zero crossing rate, and

TABLE VI
PERFORMANCE OF VOP DETECTION METHODS FOR DIFFICULT CASES OF CONTINUOUS SPEECH. THE DATABASE CONSISTS OF 173 VOPS. IN THE TABLE ABBREVIATIONS EXC, VT, MOD, AND COMB REFER TO EXCITATION SOURCE, SPECTRAL PEAKS, MODULATION SPECTRUM, AND COMBINED VOP DETECTION METHODS, RESPECTIVELY. SIMILARLY, ABBREVIATIONS HYPO, DET, MISS, AND SPU REFER TO HYPOTHESIZED, DETECTED, MISSING, AND SPURIOUS VOPS, RESPECTIVELY. TER REFERS TO THE TOTAL ERROR RATE WHICH IS THE SUM OF MISSING AND SPURIOUS VOPS. ALL ARE EXPRESSED AS % OF THE RATIO OF RESPECTIVE VOPS TO THE TOTAL NUMBER OF VOPS (173)

Method	HYPO	DET VOPs (within ms)				MISS	SPU	TER
	VOPs	$\pm$ 10	$\pm$ 20	$\pm$ 30	$\pm$ 40	VOPs	VOPs	
EXC	146	32.9	46.8	61.8	81.5	18.5	2.89	21.39
VT	148	30.6	41.6	60.1	79.8	20.2	5.78	25.80
MOD	135	35.8	50.3	67.1	76.3	23.7	1.73	25.43
COMB	149	43.3	53.2	67.6	83.8	16.2	2.31	18.51

TABLE VII
PERFORMANCE OF VOP DETECTION METHODS FOR DIFFICULT ACCENT. THE DATABASE CONSISTS OF 600 VOPS. IN THE TABLE, ABBREVIATIONS EXC, VT, MOD, AND COMB REFER TO EXCITATION SOURCE, SPECTRAL PEAKS, MODULATION SPECTRUM, AND COMBINED VOP DETECTION METHODS, RESPECTIVELY. SIMILARLY, ABBREVIATIONS HYPO, DET, MISS, AND SPU REFER TO HYPOTHESIZED, DETECTED, MISSING, AND SPURIOUS VOPS, RESPECTIVELY. TER REFERS TO THE TOTAL ERROR RATE WHICH IS THE SUM OF MISSING AND SPURIOUS VOPS. ALL ARE EXPRESSED AS % OF THE RATIO OF RESPECTIVE VOPS TO THE TOTAL NUMBER OF VOPS (600)

Method	HYPO	DET VOPs (within ms)				MISS	SPU	TER
	VOPs	$\pm$ 10	$\pm$ 20	$\pm$ 30	$\pm$ 40	VOPs	VOPs	
EXC	596	68.2	77.8	86.5	96.0	4.0	3.33	7.33
VT	623	62.0	74.8	82.7	96.5	3.5	7.33	10.83
MOD	585	69.0	79.0	89.3	96.3	3.7	1.20	4.90
COMB	589	71.8	83.0	89.8	97.3	2.7	0.83	3.53

the speech signal itself [7]. The performance of the proposed VOP detection methods are given in Table VII. The individual methods show the same trend as in the previous cases. Further, the combined method provides an improvement of 1.5%. Thus, the effectiveness of the proposed method is nearly independent of the nature of the speech signal and also accent of the speakers.

VII. SUMMARY AND CONCLUSION

In this paper, the significance of different types of energies from the speech signal for the detection of the VOPs is demonstrated. A method for the VOP detection using HE of the LP residual, representing the energy of excitation source is developed. The VOP detection using the energy of spectral peaks in the DFT spectrum, representing the vocal tract shape is also developed. A method is then developed for the VOP detection using the energy of modulation spectrum, representing the slowly varying temporal envelope components in speech. The performance of each of these methods is evaluated. Finally, the VOP evidences from each of these features are combined to get a combined VOP detection method. The combined method shows about 2.13% improvement in the performance demonstrating the different evidences available in each of the features for the VOP detection. The proposed VOP detection methods are subjected to different experimental studies like increased database size, gender, sound units categories, difficult cases of continuous speech, and accent. In all the cases, the proposed method provides nearly same performance and/or graceful degradation. This demonstrates the robustness of the proposed methods.

In this paper, the objective was only to demonstrate the significance of different speech features for the detection of the VOPs. The limitation of the proposed methods is the poor performance at $<$40 ms. This limits the use of the proposed methods in practical applications like speech recognition. The future work should focus on improving the performance at these temporal resolutions. Once we have VOP detection methods with sufficient accuracy, the methods may then be evaluated in the presence of different degradations like noise, reverberation, and speech from other speakers. After this, the usefulness of the methods may be explored in different speech processing tasks.

REFERENCES

[1] D. O'Shaughnessy, *Speech Communications: Human and Machine*, 2nd ed. New York: IEEE Press, 1999.

[2] J. R. Deller, J. G. Proakis, and J. H. Hansen, *Discrete Time Processing of Speech Signals*. Upper Saddle River, NJ: Prentice-Hall, 1993.

[3] D. J. Hermes, "Vowel onset detection," *J. Acoust. Soc. Amer.*, vol. 87, pp. 866–873, 1990.

[4] J.-F. Wang, C.-H. Hu, Shin-Hung, and J.-Y. Lee, "A heirarchical neural network based C/V segmentation algorithm for Mandarin speech recognition," *IEEE Trans. Signal Process.*, vol. 39, no. 9, pp. 2141–2146, Sep. 1991.

[5] C. Chandra Sekhar, "Neural Network models for recognition of stop consonant-vowel (SCV) segments in continuous speech," Ph.D. dissertation, Dept. of Comput. Sci. and Eng., Indian Inst. of Technol. Madras, Chennai, India, 1996.

[6] J.-H. Wang and S.-H. Chen, "A C/V segmentation algorithm for Mandarin speech using wavelet transforms," in *Proc. Int. Conf. Acoust. Speech, Signal Process.*, Sep. 1999, vol. 1, pp. 1261–1264.

[7] S. R. M. Prasanna, "Event based analysis of speech," Ph.D. dissertation, Dept. of Comput. Sci. and Eng., Indian Inst. of Technol. Madras, Chennai, India, 2004.

[8] S. R. M. Prasanna, J. M. Zachariah, and B. Yegnanarayana, "Begin-end detection using vowel onset points," in *Proc. Workshop Spoken Lang. Process.*, Jan. 2003, pp. 33–39.

[9] B. Yegnanarayana, S. R. M. Prasanna, J. M. Zachariah, and C. S. Gupta, "Combining evidence source, suprasegmental and spectral features for a fixed-text speaker verification system," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 4, pp. 575–582, Jul. 2005.

[10] S. Furui, "On the role of spectral transition for speech perception," *J. Acoust. Soc. Amer.*, vol. 80, no. 4, pp. 1016–1025, 1983.

[11] C. Chandra Sekhar and B. Yegnanarayana, "A constraint satisfaction model for recognition of stop consonant-vowel (SCV) utterances," *IEEE Trans. Speech Audio Process.*, vol. 10, no. 7, pp. 472–480, Oct. 2002.

[12] J. Y. Siva Rama Krishna Rao, C. Chandra Sekhar, and B. Yegnanarayana, "Neural network based appraoch for detection of vowel onset points," in *Proc. Int. Conf. Adv. Pattern Recognition Digital Tech.*, Dec. 1999, vol. 1, pp. 316–320.

[13] S. R. M. Prasanna and B. Yegnanarayana, "Detection of vowel onset point events using excitation source information," in *Proc. INTERSPEECH'05*, Sep. 2005, pp. 1133–1136.

[14] T. F. Quatieri, *Discrete-Time Speech Signal Processing: Principles and Practice*, 1st ed. N. Delhi, India: Pearson Education, 2001.

[15] J. Makhoul, "Linear prediction: A tutorial review," *Proc. IEEE*, vol. 63, no. 4, pp. 561–580, April 1975.

[16] T. Ananthapadmanabha and B. Yegnanarayana, "Epoch extraction from linear prediction residual for identification of closed glottis interval," *IEEE Tans. Acoust., Speech, Signal Process.*, vol. ASSP-27, no. 4, pp. 309–319, Aug. 1979.

[17] A. V. Oppenheim and R. W. Schafer, *Digital Signal Processing*. N. Delhi, India: Prentice-Hall India, 1975.

[18] *TIMIT Acoustic-Phonetic Continuous Speech Corpus*, NTIS Order PB91-505065, NIST, Gaithersburg, MD, 1990, Speech Disc 1-1.1.

[19] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. J. Lang, "Phoneme recognition using time-delay neural networks," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 37, no. 3, pp. 328–339, Mar. 1989.

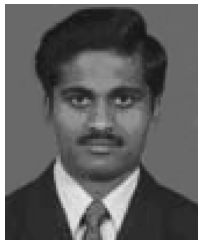
[20] S. Greenberg and B. E. D. Kingsbury, "The modulation spectrogram: In pursuit of an invariant representation of speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Apr. 1997, pp. 1647–1650.
[21] C. L. Smith, C. P. Browman, R. S. McGowan, and B. Kay, "Extracting dynamic parameters from speech movement data," *J. Acoust. Soc. Amer.*, vol. 93, no. 3, pp. 1580–1588, Mar. 1993.
[22] H. Dudley, "Remaking speech," *J. Acoust. Soc. Amer.*, vol. 11, no. 2, pp. 169–177, Oct. 1939.
[23] R. Drullman, J. M. Festen, and R. Plomp, "Effect of temporally envelope smearing on speech reception," *J. Acoust. Soc. Amer.*, vol. 95, no. 2, pp. 1053–1064, Feb. 1994.
[24] B. E. D. Kingsbury, N. Morgan, and S. Greenberg, "Robust speech recognition using the modulation spectrogram," *Speech Commun.*, vol. 25, pp. 117–132, 1998.

S. R. Mahadeva Prasanna (M'05) was born in India in 1971. He received the B.E. degree in electronics engineering from Sri Siddartha Institute of Technology, Bangalore University, India, in 1994, the M.Tech. degree in industrial electronics from the National Institute of Technology, Surathkal, India, in 1997, and the Ph.D. degree in computer science and engineering from the Indian Institute of Technology Madras, Chennai, in 2004.

He is currently an Associate Professor in the Department of Electronics and Communication Engineering, Indian Institute of Technology Guwahati, Guwahati. His research interests are in speech and signal processing, application of AI tools for pattern recognition tasks in speech and signal processing.

B. V. Sandeep Reddy was born in Khammam, India, in 1986. He received the B.Tech. degree in electronics and communication engineering from Jawaharlal Nehru Technological University, Hyderabad, India, in 2006 and the M.Tech. degree in signal processing from the Indian Institute of Technology, Guwahati, in 2008.

He is currently working in Automatic Speech Recognition (ASR) Systems in Applications Technology (AppTek), Inc., Noida, India, where he is a Member of the Technical Staff. His current research interests include speech recognition and signal processing.



P. Krishnamoorthy (S'06) was born in Tamil Nadu, India, in 1980. He received the B.E. degree in electrical and electronics engineering from Thiagarajar College of Engineering, Madurai, India, in 2001 and the M.Tech. degree in applied electronics from P.S.G. College of Technology, Coimbatore, India, in 2003. He is currently pursuing the Ph.D. degree in electronics and communication engineering at the Indian Institute of Technology Guwahati, India.

His research interests include speech and signal processing.