Vowel Onset Point Detection for Low Bit Rate Coded Speech

Anil Kumar Vuppala, *Student Member, IEEE*, Jainath Yadav, Saswat Chakrabarti, *Member, IEEE*, and K. Sreenivasa Rao, *Member, IEEE*

***Abstract*—In this paper, we propose a method for detecting the vowel onset points (VOPs) for low bit rate coded speech. VOP is the instant at which the onset of the vowel takes place in the speech signal. VOP plays an important role for the applications, such as consonant–vowel (CV) unit recognition and speech rate modification. The proposed VOP detection method is based on the spectral energy present in the glottal closure region of the speech signal. Speech coders considered to carry out this study are Global System for Mobile Communications (GSM) full rate, code-excited linear prediction (CELP), and mixed-excitation linear prediction (MELP). TIMIT database and CV units collected from the broadcast news corpus are used for evaluation. Performance of the proposed method is compared with existing methods, which uses the combination of evidence from the excitation source, spectral peaks energy, and modulation spectrum. The proposed VOP detection method has shown significant improvement in the performance compared to the existing method under clean as well as coded cases. The effectiveness of the proposed VOP detection method is analyzed in CV recognition by using VOP as an anchor point.**

***Index Terms*—Glottal closure region, spectral energy, speech coders, vowel onset point (VOP).**

I. Introduction

ROBUST speech systems in wireless environments have gained a special interest in recent years in order to enable access to remote voice-activated services. In this context, three major challenges are: varying background conditions, speech coding, and transmission channel errors [1]. While the first one has already received a lot of attention and last two deserve further investigation. The main goal of this paper is to analyze the effect of speech coding on the detection of vowel onset point (VOP).

The instant at which the onset of vowel takes place in the speech signal is known as vowel onset point. The significance of VOP can be observed in speech applications like 1) consonant–vowel (CV) units recognition, 2) spotting CV segments in continuous speech, and 3) speech rate manipulation [2]–[7]. Phonemes and triphones are widely used as subword units of speech for speech recognition, but, recent studies reveal that syllables (combinations of phonemes) are the suitable subword units for speech recognition in Indian languages [4]. Among the syllables, CV units are the most frequently (around 90% in Indian languages) occurring units [4]. VOP plays an anchor role in the recognition of CV units in Indian languages. So present study is motivated by the speech recognition for Indian languages in mobile environment.

There are various methods available in literature for the detection of VOPs [2], [4], [8]–[14]. The method presented in [8] detects VOPs by identifying the points at which there is a rapid increase in the vowel strength. The vowel strength is calculated using the difference in the energy of each of the peaks and its corresponding valleys in the amplitude spectrum. This method requires unvoiced and voiced classification of the speech signal. VOP detection method presented in [9] defines a product function which is generated from the appropriate wavelet and scaling coefficients of input speech signal. The values of the product function during vowel segments are much larger than consonant segments. Therefore, the product function can be used to detect the VOP. The methods presented in [4], [10], and [11] use hierarchical neural network, multilayer feed-forward neural network (MLFFNN) and autoassociated neural network (AANN) models to detect the VOPs. They are trained by using the trends in the speech signal parameters at the VOPs. VOP detection using Hilbert envelope of excitation source information is presented in [12]. In [13], automatic voice onset time is detected using phone model based methods with forced alignment. Voice onset time detection using reassignment spectra is presented in [14]. In [15], voice onset time detection method is presented for unvoiced stops (/p/, /t/ and /k/) using the nonlinear energy tracking algorithm (Teager energy operator). In [2], a method has been presented by combining the evidence from excitation source, spectral peaks energy, and modulation spectrum for robust detection of VOP. Each of these evidence carries complementary information with respect to VOP. The performance of combined method is better compared to existing methods. Hence, combined VOP detection method is considered for comparing the performance of proposed VOP detection method.

Low-bit rate speech coders mainly preserve the spectral characteristics of the speech signal. Hence, spectral energy based methods may perform better compared to other existing methods for determining the VOPs from coded speech. An existing VOP detection method using spectral energy is modified in the proposed method to enhance the VOP detection performance. The proposed VOP detection method uses the evidence from the spectral energy of the speech segment in the

Manuscript received September 19, 2011; revised January 19, 2012; accepted March 12, 2012. Date of publication April 06, 2012; date of current version May 07, 2012. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Keiichi Tokuda.

A. K. Vuppala and S. Chakrabarti are with the G. S. Sanyal School of Telecommunications, Indian Institute of Technology, Kharagpur 721302, India (e-mail: anil.vuppala@gmail.com; saswat@ece.iitkgp.ernet.in).

K. S. Rao and J. Yadav are with the School of Information Technology, Indian Institute of Technology, Kharagpur 721302, India (e-mail: ksrao@iitkgp.ac.in; jaibhu38@gmail.com).

Digital Object Identifier 10.1109/TASL.2012.2191284

glottal closure region. Spectral energy is high in glottal closure region; hence, it is explored in the proposed VOP detection method by considering 30% glottal cycle (pitch period) starting from glottal closure instant instead of conventional 20-ms block processing for enhancing the spectral energy. Zero-frequency filter method is used to detect the glottal closure instants. In the proposed VOP detection method, spectral energy in 500–2500 Hz band is considered, where energy of vowel is much higher than consonant. Significance of the proposed VOP detection is demonstrated in this paper using recognition of CV units for south Indian language Telugu.

Speech coders used in this work are Global System for Mobile Communications (GSM) full rate (ETSI 06.10), code-excited linear prediction (CELP) (FS-1016), and mixed-excitation linear prediction (MELP) (TI 2.4 kbps). The rest of the paper is organized as follows. Section II describes the VOP detection methods based on excitation source, spectral peaks, modulation spectrum, and a combination of all these three evidences. Proposed VOP detection method is explained in detail in Section III. The experimental results and discussions related to detection of the VOPs under coding are presented in Section IV. Effectiveness of the proposed VOP detection method is analyzed in Section V by using VOP as an anchor point in two level CV recognition. Conclusions of the present work and scope for the future work are discussed in Section VI.

II. Vowel Onset Point Detection Using Excitation Source, Spectral Peaks, Modulation Spectrum, and Combination of These Three Evidence [2]

A. VOP Detection Using Excitation Source Information

VOP detection using excitation source information is carried out in following sequence of steps. Determine the Hilbert envelope (HE) of linear prediction (LP) residual (also known as excitation source) of speech signal. Smooth the HE of the LP residual by convolving with a Hamming window of size 50 ms. The change at the VOP present in the smoothed HE of the LP residual is further enhanced by computing its slope using first-order difference (FOD). These enhanced values are convolved with the first order Gaussian difference (FOGD) operator and the convolved output is the VOP evidence using excitation source. VOP evidence using excitation source for speech signal /"*Don't ask me to carry an*"/ is shown in Fig. 1(b).

B. VOP Detection Using Spectral Peaks Energy

VOP detection using the spectral peaks energy is carried out in following sequence of steps. The speech signal is processed in blocks of 20 ms with a shift of 10 ms. For each block, a 256-point DFT is computed, and ten largest peaks are selected from the first 128 points. The sum of these spectral peaks is plotted as a function of time. The change at the VOP available in the spectral peaks energy is further enhanced by computing its slope using FOD. These enhanced values are convolved with FOGD operator. The convolved output is the VOP evidence using spectral peaks energy. VOP evidence plot using spectral peaks energy for speech signal /"*Don't ask me to carry an*"/ is shown in Fig. 1(c).

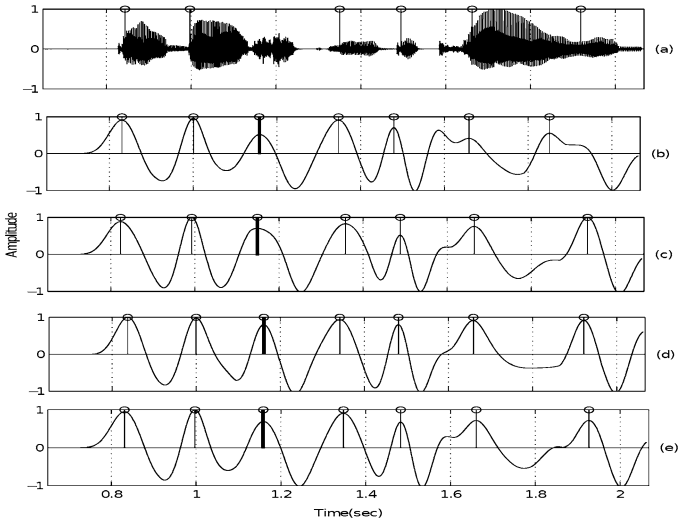

Fig. 1. VOP detection using combination of all three evidence for utterance /"*Don't ask me to carry an*"/. (a) Speech signal. VOP evidence plots for (b) excitation source. (c) spectral peaks, (d) modulation spectrum, and (e) combined method.

C. VOP Detection Using Modulation Spectrum Energy

Slowly varying temporal envelope of speech signal can be represented by using modulation spectrum. VOP detection using modulation spectrum energy is carried out in the following sequence of steps. The temporal envelope of speech is dominated by low-frequency components. The VOP evidence due to modulation spectrum is derived by passing the speech signal through a set of critical bandpass filters, and summing the components corresponding to 4–16 Hz. The change at the VOP available in the modulation spectrum energy is further enhanced by computing its slope using FOD. These enhanced values are convolved with FOGD operator and the convolved output is the VOP evidence using modulation spectrum energy. VOP evidence using modulation spectrum energy for speech signal /"*Don't ask me to carry an*"/ is shown in Fig. 1(d).

D. VOP Detection Using Combined Method

Each of the above three methods uses complementary information about the VOP, and hence they are combined for the enhancement of VOP detection performance. In the combined method, the evidence from excitation source, spectral peaks, and modulation spectrum energies are added sample by sample. VOP detection using individual and combination of all three evidences for speech signal /"*Don't ask me to carry an*"/ is shown in Fig. 1.

Fig. 1(a) shows the speech signal with manually marked VOPs for an utterance /"*Don't ask me to carry an*"/. Figs. 1(b), (c), and (d) show the VOP evidence correspond to excitation source, spectral peaks, and modulation spectrum, respectively. Fig. 1(e) shows the VOP evidence by combining the evidence. The peaks in the combined VOP evidence signal [Fig. 1(e)] are marked as the VOPs obtained from the combined method. From Fig. 1, it is observed that a spurious VOP is present in the third position in all VOP evidence plots.

III. Proposed VOP Detection Method

The major motivation for the proposed VOP detection method is the retention of spectral characteristics of speech signals by the low-bit rate speech coders. The proposed VOP detection method is based on spectral energies of speech segments present in the glottal closure region. In the existing spectral energy-based VOP detection method, the spectrum is derived using conventional block processing with a frame size of 20 ms and frame shift of 10 ms [2]. In general, we assume that speech signal in voiced region is stationary within 20–30 ms, but there exists a nonstationary behavior even in between two consecutive pitch cycles [7], [16], [17]. Therefore, spectrum estimated from the 20-ms frame corresponds to the average spectral characteristics of multiple pitch cycles present in the frame.

A pitch cycle (glottal cycle) is a combination of glottal closure and glottal open phases. During the glottal closure phase, the vocal tract is completely isolated from the trachea and lungs. Spectrum estimation during glottal closure phase will be more accurate, because true vocal tract (oral cavity) resonances will be present during that phase. Whereas in the glottal open phase, the spectrum refers to the combination of oral cavity, trachea, and lung cavity. This is due to coupling of the oral cavity with the trachea and lungs during the open phase of vocal folds. Therefore, the spectrum derived from the block processing consists of a mixture of vocal tract resonances, and resonances due to oral, trachea, and lung cavities together.

In the present work, for detecting the VOPs, spectral energy at the glottal closure region is used as evidence. In this study, the glottal closure instant (GCI) indicates starting of glottal closure phase. Therefore, the speech segment considered in this work for estimating the spectrum is 30% of glottal cycle (pitch period) starting from the glottal closure instant. The reason for choosing 30% of glottal cycle is to ensure that the speech segment under consideration should fall in the glottal closure phase. It is also known that speech signal during glottal closure phase has a high signal-to-noise ratio compared to other regions. Therefore, the spectral energy in the glottal closure region is high compared to glottal-open region [7], [16]. To identify the glottal closure region, we use glottal closure instant as the beginning of glottal closure phase. The glottal closure instants are also known as instants of significant excitation or epochs. In this work, epochs are estimated using zero-frequency filtering (ZFF) method [18]. ZFF method will give accurate GCI locations during voiced speech, and random locations for unvoiced speech. The sequence of steps for the extraction of glottal closure instants (epoch locations) used in the proposed VOP detection method is described in the following subsection.

A. Extraction of Glottal Closure Instants (Epochs) Using Zero-Frequency Filtering Method

Among the existing epoch extraction methods, the ZFF method determines the epoch locations with highest accuracy [18]. ZFF exploits the discontinuities due to impulse excitation reflected across all the frequencies including the zero frequency. The influence of the vocal tract system is negligible at zero frequency. Therefore, the zero frequency filtered speech signal carries excitation source information, which is used for extracting the epoch locations. The ZFF method consists of following sequence of steps:

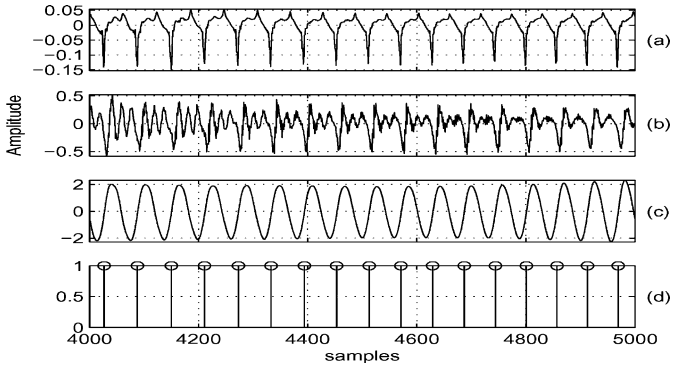

Fig. 2. Epoch (GCI) extraction using the zero-frequency filtering method. (a) Differenced electroglottogram signal. (b) Speech signal. (c) Zero frequency filtering signal. (d) Epochs (GCIs) derived from the zero-frequency filtered signal.

- Compute the difference of the input speech signal to remove any time-varying low frequency bias in the signal

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

- Compute the output of cascade of two ideal second-order digital resonators at 0 Hz, i.e.,

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

where $a_1 = +4$, $a_2 = -6$, $a_3 = +4$, $a_4 = -1$. Note that this is equivalent to passing the signal $x(n)$ through a digital filter given by

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

- Remove the trend, i.e.,

$$\hat{y}(n) = y(n) - \bar{y}(n) \tag{4}$$

where

$$\bar{y}(n) = \frac{1}{2N+1} \sum_{n=-N}^{N} y(n). \tag{5}$$

Here $2N+1$ corresponds to the size of the window used for computing the local mean, which is typically the average pitch period computed over a long segment of speech.

- The trend removed signal $\hat{y}(n)$ is termed as *zero-frequency filtered* (ZFF) signal. Positive zero-crossings in the ZFF signal correspond to epoch locations.

Epoch extraction for the segment of voiced speech using the ZFF method is shown in Fig. 2. Fig. 2(a) shows the differenced electroglottogram (EGG) signal of voiced speech segment shown in Fig. 2(b). The ZFF signal and the derived epoch locations are shown in Fig. 2(c) and (d), respectively. From the figure (see Fig. 2(a) and (d)) it is evident that the epochs extracted using ZFF method almost coincide with the negative peaks of differenced electroglottogram signal, which indicate the instants of glottal closure.

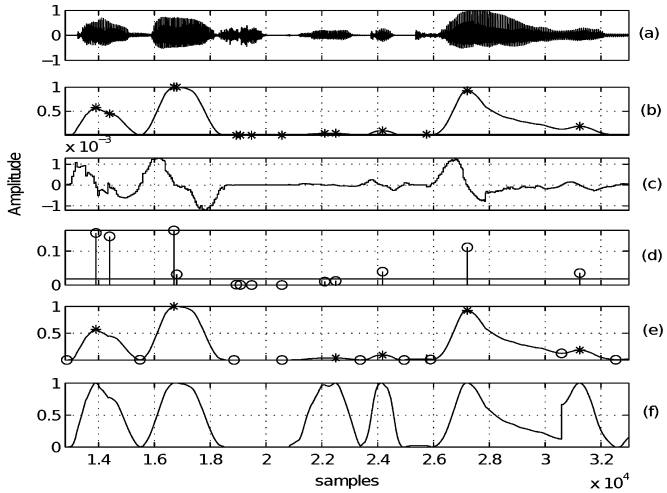

Fig. 3. Enhancement of VOP evidence for utterance /"*Don't ask me to carry an*"/. (a) Speech signal. (b) Smoothed spectral energy in 500–2500 Hz band around each epoch. (c) First-order difference (FOD) values. (d) Slope values computed at each peak locations. (e) Smoothed spectral energy plot with peak locations. (f) Enhanced values.

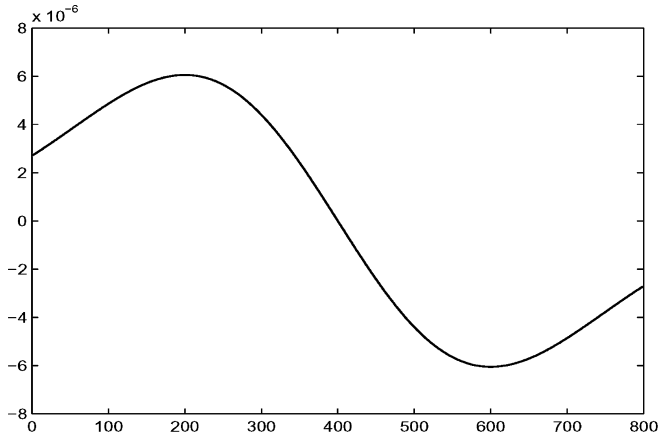

Fig. 4. FOGD operator with $L = 800$, and $\sigma = 200$.

B. Sequence of Steps in the Proposed VOP Detection Method

1) Determine the epoch locations (glottal closure instants) by using the ZFF method.
2) Compute the discrete Fourier transform (DFT) for the speech samples present in 30% of glottal cycle starting from the GCI.
3) Determine the spectral energy within the frequency band of 500–2500 Hz.
4) Spectral energy is plotted as a function of time. Fluctuations in the spectral energy contour are smoothed by using mean smoothing with 50-ms window.
5) The change at the VOP present in the smoothed spectral energy of the speech signal is enhanced by computing its slope using FOD. FOD of $x(n)$ is given by

$$x_d(n) = x(n) - x(n-1). \tag{6}$$

The finer details involved in the enhancement of VOP evidence are illustrated by using Fig. 3. Fig. 3(a) shows the speech utterance. Smoothed spectral energy in 500–2500 Hz band around each epoch is shown in Fig. 3(b). The FOD signal of smoothed spectral energy is shown in Fig. 3(c). Since FOD values corresponding to slopes, positive to negative zero crossings of slopes correspond to local peaks in the smoothed spectral energy signal. These local peaks are shown by the star (*) symbols in Fig. 3(b). The unwanted peaks in Fig. 3(b) are eliminated by using the sum of slope values within 10-ms window centered at each peak. Fig. 3(d) shows the sum of slope values within 10 ms around each peak. The peaks with the lower slope values are eliminated with a threshold set to 0.5 times the mean value of the slopes. This threshold has been considered after experimenting with huge data. Further, if two successive peaks present within 50 ms, then the lower peak among the two will be eliminated, based on the assumption that two VOPs were not present within the 50-ms interval. The desired peak locations are shown in Fig. 3(e) with the star (*) symbol after eliminating the unwanted peaks. At each local peak locations, the nearest negative to positive zero crossing points [see Fig. 3(c)] on either side are identified and marked by circles on Fig. 3(e). The regions bounded by negative to positive zero crossing points are enhanced by normalizing as shown in Fig. 3(f).

6) Significant changes in spectral characteristics present in the enhanced version of the smoothed spectral energy are detected by convolving with FOGD operator of length 100 ms. A Gaussian window $g(n)$ of length L is given by

$$g(n) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{n^2}{2\sigma^2}}, \quad n = 1, 2, \ldots, L \tag{7}$$

where σ is standard deviation. In this work, σ value of 200 is considered. The choice of length of Gaussian window (L) is based on assumption that the VOP occurs as gross level changes at intervals of about 100 ms [12], [2]. FOGD is given by $g_d(n)$ and it is shown in Fig. 4:

$$g_d(n) = g(n) - g(n-1). \tag{8}$$

The convolved output is the proposed VOP evidence plot.

7) Positive peaks in the proposed VOP evidence plot represent the VOP locations.

The flow diagram of the proposed VOP detection method is shown in Fig. 5. Output of each step in the proposed method is shown in Fig. 6 by using speech utterance /"*Don't ask me to carry an*"/. Fig. 6(a) shows the speech signal with manually marked VOPs. Spectral energy in 500–2500 Hz band and its smoothed signal are shown in Fig. 6(b) and (c), respectively. Fig. 6(d) shows the enhanced signal correspond to the signal present in Fig. 6(c). Fig. 6(e) shows the VOP evidence signal obtained by convolving the enhanced spectral energy signal with FOGD. We can observe that manual VOPs marked in Fig. 6(a) and detected VOPs marked in 6(e) are close to each other.

VOP detection using proposed and combined methods is shown in Fig. 7. Fig. 7(a) shows the speech segment for the utterance /"*Don't ask me to carry an*"/. VOP evidence plots for the speech signal shown in Fig. 7(a) using excitation source, spectral peaks, modulation spectrum, combined and proposed methods are shown in Fig. 7(b)–(f), respectively. In Fig. 7, it

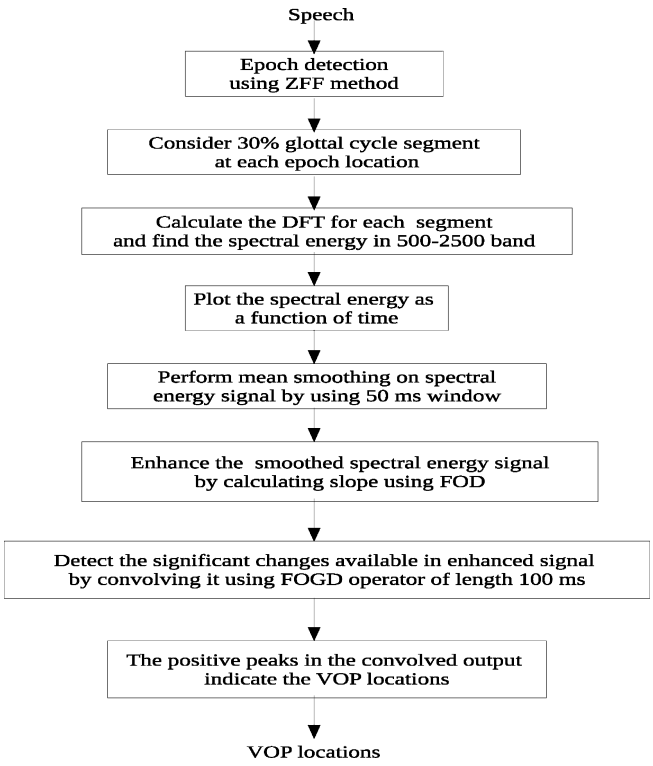

Fig. 5. Flow diagram of proposed VOP detection method.

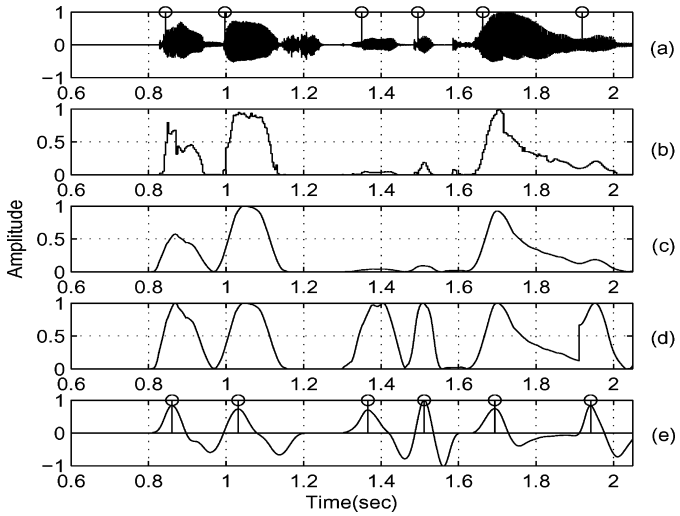

Fig. 6. VOP detection using proposed method for the utterance /"*Don't ask me to carry an*"/. (a) Speech signal with manually marked VOPs. (b) Spectral energy in 500–2500 Hz band around each epoch. (c) Mean smoothed spectral energy. (d) Enhanced spectral energy signal. (e) Proposed VOP evidence signal.

is observed that spurious VOP (third one) present in combined and individual methods is eliminated in the proposed method.

C. Choice of Frame Size

In the proposed VOP detection method, the size of speech frame to be considered at each epoch should fall within glottal closure interval, but determining the glottal closure region precisely within each glottal cycle is difficult. Therefore, we have analyzed various frame durations varying from 10% to 60% of

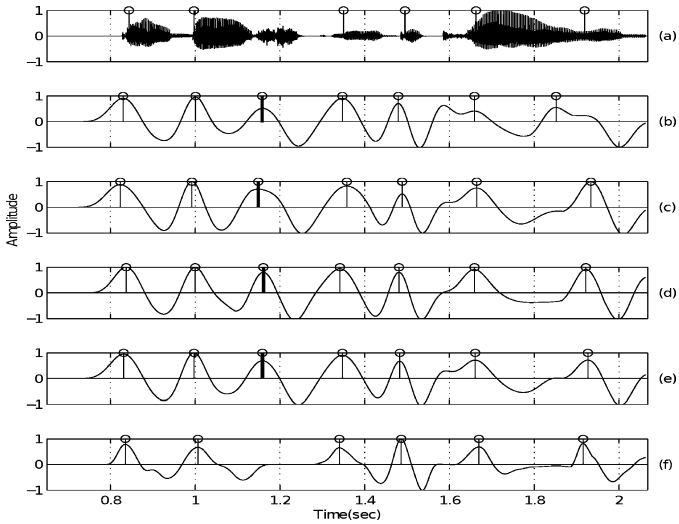

Fig. 7. VOP detection using existing and proposed VOP detection methods for utterance /"*Don't ask me to carry an*"/. (a) Speech signal. VOP evidence plots for (b) excitation source, (c) spectral peaks, (d) modulation spectrum, (e) combined method and (f) proposed method.

TABLE I
PERFORMANCE OF THE PROPOSED VOP DETECTION METHOD FOR DIFFERENT FRAME SIZES CONSIDERED AT GCIS

Percentage of pitch period	VOP detection (%) with 40 ms deviation	Missing rate
10	93.61	6.39
20	93.94	6.06
30	95.3	4.5
40	94.52	5.48
50	94.16	5.84
60	93.92	6.08

pitch period for choosing the appropriate frame size to represent the glottal closure region. For analyzing the effect of frame size on VOP detection, 110 sentences from the TIMIT database are considered. Table I shows the performance of VOP detection using the proposed method by considering different durations of speech frames in the glottal cycle. Column one indicates different duration of glottal closure speech segments considered for calculating the spectral energy. Column two indicates the percentage of VOPs detected within 40-ms deviation. In this study, 40 ms is the maximum deviation considered between reference and detected VOPs. Column three indicates the percentage of missed VOPs.

From the results presented in Table I, it is observed that accuracy in detection of VOPs is optimal by using 30% of glottal cycle as frame size in the proposed method compared to other segment durations. Hence, 30% of glottal cycle is considered in the proposed method for determining the VOPs. Robustness of proposed VOP detection method is analyzed by using speech utterances with different average pitch periods. For performing this study, we considered three sets of utterances with pitch periods varying from 2.5–5, 5–7, and 7–10 ms. Thirty utterances are recorded from children (in age group of 6 to 10 years) to cover 2.5–5 ms pitch periods. Utterances having 5–7 and 7–10 ms pitch periods are taken from the TIMIT database, and each set contains 30 utterances. In this study, 30% of pitch period is used for processing the speech segments in glottal closure region. Table II shows the performance of VOP

TABLE II
PERFORMANCE OF THE PROPOSED VOP DETECTION METHOD FOR SPEECH UTTERANCES WITH DIFFERENT AVERAGE PITCH PERIODS

Avg. pitch period (ms)	F0 values (Hz)	VOP detection (%) with 40 ms deviation	Missing (%) rate
2.5–5	200–400	95.21	4.79
5–7	142.86–200	95.44	4.56
7–10	100–142.86	95.31	4.69

detection for speech utterances having different average pitch periods. Column one indicates range of average pitch periods. Column two indicates the percentage of VOPs detected within the 40-ms deviation. Column three indicates the percentage of missed VOPs. From the results, it is observed that the proposed VOP detection method is robust to speech signals with different pitch periods.

D. Choice of Frequency Band

In this work, VOP evidence is derived by computing spectral energy of speech segment at each epoch. If the epoch is happen to be GCI in the vowel region, the spectral energy within 500–2500 Hz band is very high compared to consonant and non-speech regions. For voiced consonants or nasals most of the spectral energy is present below 500 Hz. For unvoiced consonants, fricatives and other sound units most of the spectral energy is present beyond 3000 Hz. Therefore, the spectral energy in frequency band 500–2500 Hz will provide accurate VOP evidence compared to sum of 10 spectral peaks used in the existing methods.

VOP detection using proposed method by considering spectral energy in different frequency bands, for an utterance /"*Don't ask me to carry an*"/ is shown in Fig. 8. Fig. 8(a) shows the speech signal. VOP evidence plots using proposed method by considering spectral energy in 0–4000 Hz, 750–2500 Hz, 500–2500 Hz, 1000–4000 Hz, and 2500–4000 Hz bands are shown in Fig. 8(b)–(f), respectively. In Fig. 8(c) we can observe that the third VOP is missed and in Figs. 8(e) and (f) spurious VOPs can be observed (VOPs indicated in bold). In Fig. 8(b) and (d), only genuine VOPs are observed. The choice of the 500–2500 Hz band in the proposed method is extensively studied by considering different frequency bands. Table III shows the performance of VOP detection for CV units from Telugu broadcast news speech corpus using the proposed method by considering spectral energy at different frequency bands in 0–4000 Hz range. For evaluation, 950 (10 utterances from each 95 most frequently occurred CV units) CV utterances are considered to cover various consonant and vowel combinations. Therefore, this evaluation also ensures the robustness of the chosen frequency band in the proposed VOP detection method for different classes of speech segments. Column one indicates different frequency bands considered for calculating the spectral energy. Column two indicates the percentage of VOPs detected within the 40-ms deviation. Column three indicates the percentage of missed VOPs. Column four indicates the average deviation (in ms) with respect to the manual marked VOPs. From Table III, it is observed that accuracy in detection of VOPs is better by using the 500–2500 Hz frequency band in the proposed method compared to other bands. Hence, in our

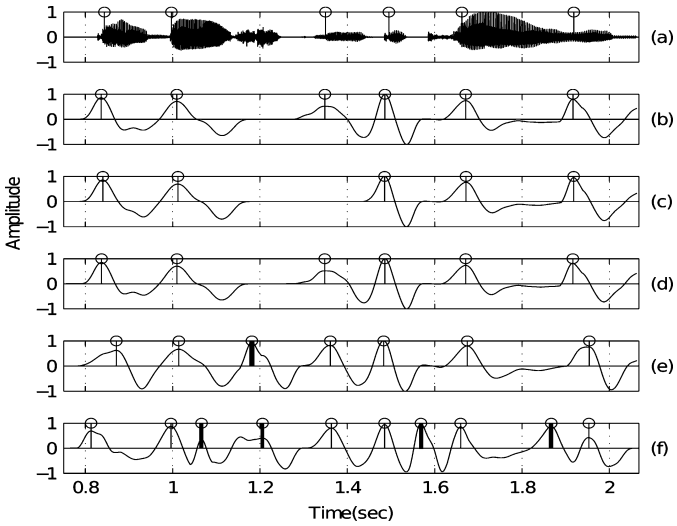

Fig. 8. VOP detection using proposed method for different frequency bands on the utterance /"*Don't ask me to carry an*"/. (a) Speech signal. VOP evidence plot by taking spectral energy at (b) 0–4000 Hz, (c) 750–2500 Hz, (d) 500–2500 Hz, (e) 1000–4000 Hz, and (f) 2500–4000 Hz.

TABLE III
PERFORMANCE OF THE PROPOSED VOP DETECTION METHOD FOR CV UNITS FROM THE TELUGU BROADCAST NEWS SPEECH CORPUS BY CONSIDERING SPECTRAL ENERGY AT DIFFERENT FREQUENCY BANDS

Frequency band	VOP detection (%) with 40 ms deviation	Missing rate (%)	Average deviation (ms)
0–4000	95.6	4.4	12.87
0–2500	96.5	3.5	13.23
0–2000	96.56	3.44	13.23
250–4000	96.56	3.44	12.67
250–3750	96.11	3.89	12.87
250–3250	96.56	3.44	12.69
250–2500	97.19	2.81	12.99
500–2750	96.88	3.12	12.25
500–2500	**97.5**	**2.5**	**11.22**
500–4000	96.88	3.12	12.88
750–2500	93.44	6.56	14.70

further studies the 500–2500 Hz frequency band is used in the proposed method for determining the VOPs.

IV. PERFORMANCE OF THE PROPOSED VOP DETECTION METHOD

In this work, performance of the proposed VOP detection method is compared with the combined method which uses the combination of evidences from the excitation source (EXC), spectral peaks (SPs), and modulation spectrum (MOD). VOP detection methods are evaluated on continuous speech and CV units from broadcast speech.

A. VOP Detection From Continuous Speech Under Coding

VOP detection studies are conducted on the TIMIT database [19] for analyzing the performance of the VOP detection methods. About 110 sentences (60 sentences spoken by female speakers and 50 sentences are spoken by male speakers) having 1197 manually marked VOPs are considered for analyzing the performance of the proposed VOP detection method. Among 1197 VOPs, 534 VOPs correspond to the utterances spoken by male speakers, and the rest 663 VOPs correspond to the utterances spoken by female speakers. GSM full rate (ETSI

TABLE IV
PERFORMANCE OF VOP DETECTION USING EXCITATION SOURCE (EXC), SPECTRAL PEAKS (SP), MODULATION SPECTRUM (MOD), COMBINED (COMB), AND PROPOSED METHODS ON TIMIT DATABASE (1197 REFERENCE VOPS)

VOP detection method	Hypothesized VOPs	VOPs detected within ms (%)				AVG dev.	MISS VOPs	SPU VOPs
		10	20	30	40	(≈ ms)	(%)	(%)
Clean								
EXC	1176	34	49	59	94	20	6	4
SP	1172	29	48	70	93	21	7	5
MOD	1126	33	50	73	92	18	8	2
COMB	1173	51	59	70	95	16	5	3
Proposed	1162	65	83	91	95	12	5	2
GSM								
EXC	1162	31	47	57	91	21	9	6
SP	1159	28	47	68	92	21	8	5
MOD	1066	27	45	65	85	20	15	4
COMB	1114	45	51	65	90	18	10	3
Proposed	1164	63	81	88	94	13	6	3
CELP								
EXC	1103	25	40	50	84	23	16	8
SP	1114	26	42	64	87	22	13	6
MOD	1089	24	41	62	85	21	15	6
COMB	1102	34	46	59	86	20	14	6
Proposed	1126	39	67	80	88	17	12	6
MELP								
EXC	1112	26	42	52	86	22	14	7
SP	1125	28	46	65	89	21	11	5
MOD	1054	26	44	64	84	20	16	4
COMB	1090	38	48	59	87	19	13	4
Proposed	1138	52	72	84	91	15	9	4

06.10), CELP (FS-1016), and MELP (TI 2.4 kbps) speech coders are considered in this study.

Performance of different VOP detection methods is compared using parameters like average deviation, missing rate, and spurious rate. VOPs detected within 40-ms deviation to the reference VOPs are considered as genuine VOPs. The ratio (in %) of number of genuine VOPs detected to the total number of reference VOPs are measured for different time resolutions (10–40 ms). Average deviation (in ms) is calculated from the deviations of genuine detected VOPs. The ratio (in %) of undetected VOPs to the total number of reference VOPs is termed as missing rate (MISS). VOPs detected other than genuine VOPs are termed as spurious VOPs. The ratio of spurious VOPs (in %) detected to the number of reference VOPs is termed as spurious rate (SPU).

Table IV shows the accuracy in detection of VOPs using different methods in the presence of various coders. Column one indicates different methods considered in the analysis for detecting the VOPs. Column two indicates the total number of VOPs detected using various methods. Columns three through six indicate the percentage of VOPs detected within the specified deviations. Column seven indicates the average deviation (in ms) with respect to the manual marked VOPs. Columns eight and nine indicate the percentage of missed and spurious VOPs, respectively. Accuracy in detection of VOPs is observed to be superior using the proposed method under both clean and coded cases, compared to existing methods (see Table IV). Average deviation has reduced significantly, in case of proposed method compared to other methods. Average deviation using proposed method is around 1, 5, and 3 ms higher compared to the clean case for GSM full rate, CELP, and MELP coders, respectively. Spurious VOPs are also observed to be increased due to coding (see Table IV).

VOP detection from CELP coded speech is observed to be less accurate compared to MELP coder even though the bit rate provided by MELP coder is less than CELP coder. The reason may be due to poor way of representation of excitation signal in CELP coding technique. CELP coder uses the code book to represent the excitation signal, which introduce more approximation compared to other coders. From the results, it is observed that performance of VOP detection methods based on the spectral energy are performing superior compared to the other methods in presence of coding. Among the methods based on spectral energy, performance of the proposed method is better in all aspects. The improved performance of the proposed method is due to exploiting the high SNR of the speech signal present in the glottal closure phase. In view of time complexity, proposed method is better compared to existing methods, since there is no need of picking spectral peaks or combination of evidences.

B. VOP Detection From Consonant-Vowel (CV) Units Under Coding

VOP plays a crucial role in CV unit recognition. VOP is an instant at which the consonant of CV utterance ends and onset of vowel takes place. CV units collected from the Telugu broadcast news corpus are used for evaluation [11]. In this study, 95 CV classes are considered. From each CV class 10 utterances are considered, with this total number of VOPs used in this study is 950. In this study analysis of spurious VOPs is not applicable, since each utterance contains only one VOP, and it is determined based on the peak in VOP evidence signal.

Table V shows the detection accuracy of VOPs for Telugu broadcast news data using the proposed and existing methods. Column one indicates different methods considered in the analysis for detecting the VOPs. Columns two through five indicate the percentage of VOPs detected within the specified deviations. Column six indicates the average deviation (in ms) with respect to the manual marked VOPs. Column seven indicates the percentage of missed VOPs. From Table V, it is observed that accuracy in detection of VOPs is better in case of proposed method compared to combined method. Average deviation using proposed method is around 5 ms and 1 ms higher compared to clean case for CELP and MELP coders, respectively, and for the GSM full-rate coder average deviation seems to be close to clean case (see Table V). In the presence of coding, the proposed method and method based on spectral peaks energy performing better compared to other methods.

V. RECOGNITION OF CONSONANT–VOWEL (CV) UNITS USING VOWEL ONSET POINTS

Significance of the accuracy of proposed VOP detection is demonstrated in this paper using consonant–vowel (CV) unit recognition in south Indian language Telugu. Syllables (combinations of phones) are found to be more appropriate subword units for speech recognition in Indian languages [20]–[22]. Syllables capture significant co-articulation effects and pronunciation variation compared to phones. In general, the syllables are of type C^mVC^n, where C refers to consonant, V refers to a vowel, and m, n refers to the number of consonants preceding

TABLE V
PERFORMANCE OF VOP DETECTION USING EXCITATION SOURCE (EXC), SPECTRAL PEAKS (SP), MODULATION SPECTRUM (MOD), AND COMBINED (COMB) METHODS ON CV UNITS FROM THE BROADCAST NEWS SPEECH CORPUS (950 VOPS ARE CONSIDERED)

VOP detection method	VOPs detected within ms (%)				AVG dev. ($\approx$ ms)	MISS VOPs (%)
	10	20	30	40		
Clean						
EXC	44	69	84	90	17	10
SP	48	78	92	95	13	5
MOD	36	50	74	92	20	8
COMB	50	74	89	95	15	5
Proposed	69	88	93	97	11	3
GSM						
EXC	42	67	82	89	18	11
SP	46	75	90	94	14	6
MOD	32	44	67	87	21	13
COMB	45	72	86	91	17	9
Proposed	67	86	93	96	11	4
CELP						
EXC	28	55	76	86	24	14
SP	37	68	85	93	19	7
MOD	27	45	67	84	22	16
COMB	29	62	82	91	22	9
Proposed	41	72	88	92	16	8
MELP						
EXC	29	53	73	85	22	15
SP	45	71	87	92	16	8
MOD	33	48	71	87	21	13
COMB	37	63	79	87	20	13
Proposed	54	75	84	92	12	8

and following the vowel. Among these units, the CV units are the most frequently (around 90% in Indian languages) occurring units [20]. Hence CV units are considered in this work for analyzing the accuracy of VOPs and recognition of CV units. The speech corpus considered for this study consists of broadcast news in Telugu [20], [21], [23]. In this work, CV recognition is carried out using the two-level approach proposed in our earlier work [24]. VOP plays a crucial role in two-level CV recognition system. Details of the database and CV recognition system are described in the following subsections.

A. Database

Duration of Telugu broadcast news database is about five hours, consists of 20 news bulletins read by 11 male speakers and 9 female speakers. Among 20 bulletins, 15 (8 male + 7 female) are used for training the CV recognition models, and 5 (3 male + 2 female) are used for testing the CV recognition system. Manually marked syllable boundaries available in the database are used for picking the CV units from continuous speech. In the context of Indian languages there are 145 (29 consonants and 5 vowels) CV units [20]. In this work, 95 CV classes whose frequency of occurrence is more than 50 in the database are considered for the analysis. Their contribution is found to be more than 95% of CV units present in the Telugu broadcast database. CV units considered in this work are listed in Table VI. Among 95 CV units, a, e, i, o, and u vowel groups contain 26, 16, 22, 10, and 21 consonant classes, respectively.

B. CV Recognition System

High similarity and large number of CV unit classes are the major issues involved in CV recognition. In our earlier work

TABLE VI
LIST OF 95 CV UNITS CONSIDERED FROM THE TELUGU BROADCAST NEWS CORPUS

Subgroup	CV units
/a/	ka, cha, Ta, ta, pa kha, Tha, tha, pha ga, ja, Da, da, ba gha, dha, bha, na, ma ya, ra, la, va, ha, sha, sa
/e/	ke, che, Te, te, pe phe, je, De, de, ne, me ye, re, le, ve, se
/i/	ki, chi, Ti, ti, pi thi, gi, ji, Di, di, bi dhi, bhi, ni, mi, yi, ri li, vi, hi, shi, si
/o/	ko, cho, to, po, do mo, yo, ro, lo, so
/u/	ku, chu, Tu, tu, pu thu, gu, ju, Du, du, bu dhu, bhu, nu, mu, yu ru, lu, vu, shu, su

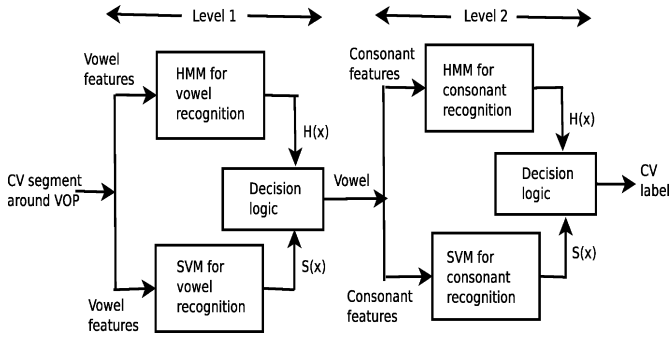

Fig. 9. Two-level CV recognition system using HMM and SVM.

[24], we have proposed the two-level approach for the recognition of CV units in Indian languages. In the first level, the vowel will be recognized, and in the second level, the consonant will be recognized. In both levels, complementary evidences from hidden Markov model (HMM) and support vector machine (SVM) models are combined with appropriate weights. Since, data is limited in the case of the Telugu broadcast news corpus, evidences of HMM and SVM models are combined to gain the advantages of sequential and distribution capturing nature of HMM, and the discriminative nature of SVM [24] to enhance the performance of the CV recognition system. The two-level CV recognition system used in this work is shown in Fig. 9.

In the two-level CV recognition system, separate features are used at each level by using VOP as an anchor point. At the first level, vowel recognition models are developed using features extracted from VOP to the end of CV segment (i.e., only vowel region), and at the second-level consonant models are developed using features extracted from consonant and transition regions. In this work, the 40-ms speech segment to the right of VOP is assumed as the transition region. Mel-frequency cepstral coefficients (MFCCs) [25] extracted from a frame of 20 ms with a frame shift of 5 ms are used for developing the acoustic

TABLE VII
CV RECOGNITION ACCURACY USING THE TWO-STAGE CV RECOGNITION SYSTEM FOR DIFFERENT VOP DETECTION METHODS UNDER CODING

VOP method	Recognition accuracy (%)			
	Clean	GSM	CELP	MELP
EXC	63.32	61.82	54.22	53.17
SP	63.81	61.16	54.16	53.42
MOD	62.14	60.11	52.22	52.27
COMB	64.36	62.18	54.66	54.37
Proposed	66.14	64.11	58.32	57.82

models. HMM models are developed using a maximum-likelihood approach using the HMM tool kit (htk) [26]. Feature vectors of size 39 dimensions (13 MFCC + delta + delta-delta coefficients) are used for developing of HMM models. HMM models are developed using three states and 64 mixtures per state. SVM models are developed using the one-against-the-rest approach using open-source SVMTorch [27]. Gaussian kernel with a width of 40 is used to build SVM models. SVM models are developed by extracting fixed-length feature vectors from consonant and vowel regions of CV units. The dimension of feature vector extracted from each utterance for developing the SVM models is 390 (10 frames $\times$ 39 MFCC per frame).

C. Performance of VOP Based CV Recognition System

The objective of this study is to analyze the effect of accuracy in VOP detection on CV recognition in different coding environments. In this study 38 729 CV utterances are used for training, and 13 974 are used for testing. CV recognition systems are developed separately for each of the coding environment. Table VII shows the recognition accuracy of CV units from Telugu broadcast news data using the different VOP detection methods under coding. In Table VII, column one indicates the VOP detection methods used in this study. Columns two through five indicate recognition accuracy of CV units for clean, GSM, CELP, and MELP coders under matched condition (training and testing performed in the same condition). From Table VII, it is observed that CV recognition accuracy is better using the proposed VOP detection method compared to existing methods. CV unit recognition based on the proposed VOP detection method has shown nearly 2%–4% improvement over the combined method. This improvement is very significant in view of the large number of CV classes. The improvement in recognition accuracy using the proposed VOP method, compared to the combined VOP method, is marginal in case of clean and GSM speech, but significant improvement is observed in case of CELP and MELP coders speech (see columns two through five in Table VII).

VI. SUMMARY AND CONCLUSION

In this paper, we have proposed a method for detecting the VOPs in low bit rate coded speech using spectral energies of the speech segments present in the glottal closure region. The reasons for choosing the speech segments at the glottal closure region for deriving the spectral energy are 1) speech during glottal closure phase has high signal-to-noise ratio and 2) vocal tract resonances during glottal closure phase are more accurate. These merits are exploited in the proposed VOP detection method by considering 30% of glottal cycle starting from glottal closure instant instead of conventional 20-ms frame with block processing. The zero-frequency filter method is used for the detection of glottal closure instants. The proposed VOP detection method uses the spectral energy in the 500–2500 Hz band, where the energy of the vowel is much higher than the consonant.

VOP detection performance is studied under GSM full rate, CELP, and MELP coding environments. From the results, it is observed that degradation in VOP detection performance is not significant in case of the GSM full rate coder, and significant in the case of CELP and MELP coders. It is observed that VOP detection from the CELP-coded speech is more degraded compared to the MELP coder, even though the bit rate of the MELP coder is less than the CELP coder. As speech coders preserve the characteristics of the vocal tract system, VOP detection methods based on spectral characteristics, performed better compared to other methods. In case of both clean and coded speech, the proposed VOP detection method gives improved performance compared to the excitation source, spectral peaks energy, modulation spectrum, and combined methods. Further, effectiveness of the proposed VOP detection method is evaluated by performing recognition of CV units under coding. The recognition accuracy of CV units based on the proposed VOP detection method has shown nearly 2% to 4% improvement over the combined VOP detection method. Future studies can be carried out by analyzing the robustness of the proposed VOP detection method in presence of noisy speech and noisy-coded speech corpus. Significance of the proposed VOP detection method can be further analyzed for applications like speech rate modification, spotting CV units from continuous speech, etc.

REFERENCES

[1] J. M. Huerta, "Speech recognition in mobile environments," Ph.D. dissertation, Carnegie Mellon Univ., Pittsburgh, PA, Apr. 2000.
[2] S. R. M. Prasanna, B. V. S. Reddy, and P. Krishnamoorthy, "Vowel onset point detection using source, spectral peaks, and modulation spectrum energies," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 4, pp. 556–565, May 2009.
[3] S. R. M. Prasanna, S. V. Gangashetty, and B. Yegnanarayana, "Significance of vowel onset point for speech analysis," in *Proc. Int. Conf. Signal Process. Commun.*, Bangalore, India, 2001, pp. 81–88.
[4] S. V. Gangashetty, C. C. Sekhar, and B. Yegnanarayana, "Detection of vowel onset points in continuous speech using autoassociative neural network models," in *Proc. Int. Conf. Spoken Lang. Process.*, Jeju Island, Korea, 2004, pp. 401–410.
[5] A. K. Vuppala, S. Chakrabarti, and K. S. Rao, "Effect of speech coding on recognition of consonant-vowel (CV) units," in *Proc. Int. Conf. Contemp. Comput. (Springer Commun. Comput. Inf. Sci. ISSN: 1865–0929)*, Noida, India, Aug. 2010, pp. 284–294.
[6] K. S. Rao and B. Yegnanarayana, "Duration modification using glottal closure instants and vowel onset points," *Speech Commun.*, vol. 51, pp. 1263–1269, 2009.
[7] K. S. Rao, "Voice conversion by mapping the speaker-specific features using pitch synchronous approach," *Comput. Speech Lang.*, vol. 24, pp. 474–494, Jul. 2010.
[8] D. J. Hermes, "Vowel onset detection," *J. Acoust. Soc. Amer.*, vol. 87, pp. 866–873, 1990.
[9] J.-H. Wang and S.-H. Chen, "A C/V segmentation algorithm for Mandarin speech using wavelet transforms," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Phoenix, AZ, 1999, pp. 1261–1264.

[10] J.-F. Wang, C. H. Wu, S. H. Chang, and J. Y. Lee, "A hierarchical neural network based C/V segmentation algorithm for Mandarin speech recognition," *IEEE Trans. Signal Process.*, vol. 39, no. 9, pp. 2141–2146, Sep. 1991.
[11] S. V. Gangashetty, C. C. Sekhar, and B. Yegnanarayana, "Extraction of fixed dimension patterns from varying duration segments of consonant-vowel utterances," in *Proc. IEEE ICISIP*, 2004, pp. 159–164.
[12] S. R. M. Prasanna and B. Yegnanarayana, "Detection of vowel onset point events using excitation source information," in *Proc. Interspeech*, Lisbon, Portugal, 2005, pp. 1133–1136.
[13] A. Kazemzadeh, J. Tepperman, J. Silva, H. You, S. Lee, A. Alwan, and S. Narayanan, "Automatic detection of voice onset time contrasts for use in pronunciation assessment," in *Proc. Int. Conf. Spoken Lang. Process.*, Pittsburgh, PA, 2006.
[14] V. Stouten and H. V. Hamme, "Automatic voice onset time estimation from reassignment spectra," *Speech Commun.*, vol. 51, pp. 1194–1205, 2009.
[15] J. H. L. Hansen, S. S. Gray, and W. Kim, "Automatic voice onset time detection for unvoiced stops (/p/, /t/, /k/) with application to accent classification," *Speech Commun.*, vol. 52, pp. 777–789, 2010.
[16] S. R. M. Kodukula, "Significance of excitation source information for speech analysis," Ph.D. dissertation, IIT Madras, Madras, Spain, Mar. 2009.
[17] S. Guruprasad, "Exploring features and scoring methods for speaker recognition," M.S. thesis, IIT Madras, Madras, Spain, 2004.
[18] K. S. R. Murty and B. Yegnanarayana, "Epoch extraction from speech signals," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 8, pp. 1602–1613, Nov. 2008.
[19] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue, "TIMIT acoustic-phonetic continuous speech corpus linguistic data consortium," in *Proc. IEEE ICISIP*, Philadelphia, PA, 1993.
[20] S. V. Gangashetty, "Neural network models for recognition of consonant-vowel units of speech in multiple languages," Ph.D. dissertation, IIT Madras, Madras, Spain, Oct. 2004.
[21] C. C. Sekhar, "Neural Network models for recognition of stop consonant-vowel (SCV) segments in continuous speech," Ph.D. dissertation, IIT Madras, Madras, Spain, 1996.
[22] R. Thangarajan, A. M. Natarajan, and M. Selvam, "Syllable modeling in continuous speech recognition for Tamil language," *Int. J. Speech Technol.*, vol. 12, pp. 47–57, 2009.
[23] S. V. Gangashetty, C. C. Sekhar, and B. Yegnanarayana, "Spotting multilingual consonant-vowel units of speech using neural networks," in *Proc. ISCA Tutorial Res. Workshop Non-Linear Speech Process.*, 2005, pp. 287–297.
[24] A. K. Vuppala, K. S. Rao, S. Chakrabarti, P. Krishnamoorthy, and S. R. M. Prasanna, "Recognition of consonant-vowel (CV) units under background noise using combined temporal and spectral preprocessing," *Int. J. Speech Technol.*, vol. 14, no. 1, 2011.
[25] J. W. Picone, "Signal modeling techniques in speech recognition," *Proc. IEEE*, vol. 81, no. 9, pp. 1215–1247, Sep. 1993.
[26] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book Version 3.0.*. Cambridge, U.K.: Cambridge Univ. Press, 2000.
[27] R. Collobert and S. Bengio, "SVMTorch: Support vector machines for large-scale regression problems," in *Proc. J. Mach. Learn. Res.*, 2001, pp. 143–160.

Anil Kumar Vuppala (S'09) received the B.Tech. degree in electronics and communications engineering from Jawaharlal Nehru Technological University, Hyderabad, India, in 2005, and the M.Tech. degree in electronics and communications engineering from National Institute of Technology, Kurukshetra, India, in 2007. He is currently pursuing the Ph.D. degree at the in G. S. Sanyal School of Telecommunications, Indian Institute of Technology, Kharagpur.

His research interests include speech processing and speech recognition.

Jainath Yadav received the M.Sc. degree in computer science from Banaras Hindu University, Varanasi, India, in 2009 and the M.Tech. degree in information technology from the Indian Institute of Technology Kharagpur, in 2011. He is currently pursuing the Ph.D. degree in the School of Information Technology, Indian Institute of Technology Kharagpur.

His research interests include speech and signal processing.

Saswat Chakrabarti (M'00) received the M.Tech. and Ph.D. degrees in electronics and electrical communications engineering from the Indian Institute of Technology (IIT), Kharagpur, in 1985 and 1992 respectively.

He has been with IIT, Kharagpur, since 1991, where he is currently a Professor in the G. S. Sanyal School of Telecommunications. His research focuses on wireless communications, error control coding, digital modulation schemes, wireless ad-hoc networks, and bio-telemetry. He has published more than 100 papers in international journals and conference proceedings. He has also been handling more than 20 consultancy and sponsored research and technology development projects.

K. Sreenivasa Rao (M'05) received the Ph.D. degree from the Department of Computer Science and Engineering, Indian Institute of Technology (IIT), Chennai, in 2005.

He is currently working as an Assistant Professor in the School of Information Technology, IIT, Kharagpur. His research interests are speech signal processing and neural networks. He has published over 100 papers in international journals and conference proceedings in these areas.