Int J Speech Technol (2013) 16:229–235
DOI 10.1007/s10772-012-9179-8

Vowel onset point detection for noisy speech using spectral energy at formant frequencies

Anil Kumar Vuppala · K. Sreenivasa Rao

Received: 24 June 2012 / Accepted: 11 October 2012 / Published online: 23 October 2012

Abstract In this paper, we propose a method for robust detection of the vowel onset points (VOPs) from noisy speech. The proposed VOP detection method exploits the spectral energy at formant frequencies of the speech segments present in glottal closure region. In this work, formants are extracted by using group delay function, and glottal closure instants are extracted by using zero frequency filter based method. Performance of the proposed VOP detection method is compared with the existing method, which uses the combination of evidence from excitation source, spectral peaks energy and modulation spectrum. Speech data from TIMIT database and noise samples from NOISEX database are used for analyzing the performance of the VOP detection methods. Significant improvement in the performance of VOP detection is observed by using proposed method compared to existing method.

Keywords Vowel onset point (VOP) · Formant frequencies · Glottal closure region · Excitation source · Spectral peaks · Modulation spectrum

1 Introduction

The instant at which the onset of vowel takes place in the speech signal is known as vowel onset point (VOP). VOP plays an anchor role in applications such as consonant-vowel (CV) unit recognition, speech segmentation and speech rate modification (Prasanna et al. 2009, 2001; Gangashetty et al. 2004a; Rao and Yegnanarayana 2009; Vuppala et al. 2011, 2012a, 2012b). There exists several methods for the detection of VOPs. The method presented in Hermes (1990) detects VOPs by identifying the events 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 in the amplitude spectrum, and the energy of a dip associated with the peak. This method requires unvoiced and voiced classification of the speech signal. In Wang and Chen (1999) a product function generated from the appropriate wavelet and scaling coefficients of input speech signal is used to determine the VOPs. The values of product function for vowel segments are much larger than consonant segments. The methods presented in Gangashetty et al. (2004a, 2004b), Wang et al. (1991) 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 the excitation source signal is presented in Prasanna and Yegnanarayana (2005). In Prasanna et al. (2009), a method has been proposed by combining the evidence from excitation source, spectral peaks energy, and modulation spectrum for the robust detection of VOP. Each of these evidence carries complementary information with respect to VOP. In this paper, this VOP detection method is termed as COMB-ESM. The performance of the other existing methods are inferior compared to the COMB-ESM method. Hence, proposed method is compared with COMB-ESM method.

In real-time environment noise is one of the major degradation. In this work, we propose a method for robust de-

A.K. Vuppala (✉)
LTRC, International Institute of Information Technology-Hyderabad, Hyderabad, India
e-mail: anil.vuppala@gmail.com

K.S. Rao
School of Information Technology,
Indian Institute of Technology, Kharagpur, India
e-mail: ksrao@iitkgp.ac.in

tection of the VOPs under noise. Proposed method uses the spectral energy at formant frequencies of the speech segments present in glottal closure region for the detection of VOPs. In general, voiced regions contains most of spectral energy. Within voiced region, in each pitch cycle, speech energy is dominant in glottal closure phase compared to glottal open phase. This is due to instant of significant excitation at the instant of glottal closure. Within glottal closure region, most of the energy is concentrated at the formant frequencies. Therefore, the proposed method exploits the spectral energy at the formant frequencies of the speech signal present in glottal closure region, for the detection of VOPs.

In this work, epochs or instants of significant excitation correspond to the instants of glottal closure in the case of voiced speech, and some random excitations, like the onset of burst, in the case of nonvoiced speech. In this study, glottal closure regions are determined by using epochs, which are extracted using zero frequency filter method (Murty and Yegnanarayana 2008). Formants in the glottal closure region are extracted using group delay function (Joseph et al. 2006). Performance of the VOP detection methods is analyzed by using TIMIT database for white and vehicle noise at different signal-to-noise ratios (SNRs). The rest of the paper is organized as follows. Baseline methods for VOP detection, epoch extraction and formant extraction are briefly explained in Sect. 2. Section 3 discusses the proposed VOP detection method. Performance evaluation of the proposed method is presented in Sect. 4. Summary of the present work is discussed in Sect. 5.

2 Baseline methods for vowel onset point detection, epoch extraction, and formant extraction

2.1 Vowel onset points detection using COMB-ESM method

In this study, performance of the proposed VOP detection method is compared with COMB-ESM method (Prasanna et al. 2009). COMB-ESM method combines the multiple evidences from excitation source, spectral peaks, and modulation spectrum (Prasanna et al. 2009). The basic steps in the COMB-ESM method are as follows:

- Derive the VOP evidence from excitation source, spectral peaks, and modulation spectrum. Here the evidence from excitation source information is obtained from the Hilbert envelope of the linear prediction residual signal. The evidence from the spectral peaks is obtained by summing the ten largest spectral peaks of each speech frame. The evidence due to modulation spectrum is derived by passing the speech signal through a set of critical band pass filters, and summing the components corresponding to 4–16 Hz.
- The above evidences are further enhanced by computing their slope using first order difference.
- These enhanced evidences are convolved with the first order Gaussian difference operator for deriving the final VOP evidences.
- Individual VOP evidences derived from excitation source, spectral peaks and modulation spectrum are added sample by sample to get the combined VOP evidence plot.
- The positive peaks in the combined VOP evidence signal are hypothesized as the locations of VOPs.

Figure 1 shows the detection of VOP using individual and combined methods. Figure 1(a) shows the speech signal with manually marked VOPs for an utterance "*Don't ask me to carry an*". VOP evidences correspond to excitation source, spectral peaks and modulation spectrum are shown in Figs. 1(b), 1(c) and 1(d), respectively. Figure 1(e) shows the VOP evidence by combining the evidence shown in Figs. 1(b), 1(c) and 1(d). The peaks in the combined VOP evidence plot (Fig. 1(e)) are marked as the VOPs obtained from the combined method. VOP detection performance of combined method for clean speech is about 96 % within 40 ms deviation, and only 45 % within 10 ms deviation (Prasanna et al. 2009).

2.2 Epoch extraction using ZFF method (Murty and Yegnanarayana 2008)

Among the existing epoch extraction methods, zero frequency filter (ZFF) method is known for determining the epoch locations accurately (Murty and Yegnanarayana 2008). It is also interesting to note that the ZFF method is robust against degradations such as white noise, babble, high-frequency channel, and vehicle noise (Murty and Yegnanarayana 2008). ZFF method exploits the discontinuities due to impulse excitation reflected across all the frequencies including the zero frequency. The ZFF method consists of following sequence of steps:

(1) Difference the input speech signal to remove any time-varying low frequency bias in the signal

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

(2) Compute the output of cascade of two ideal digital resonators at 0 Hz i.e.,

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

where $a_1 = +4$, $a_2 = -6$, $a_3 = +4$, $a_4 = -1$.

(3) Remove the trend i.e.,

$$\hat{y}(n) = y(n) - \bar{y}(n) \tag{3}$$

where

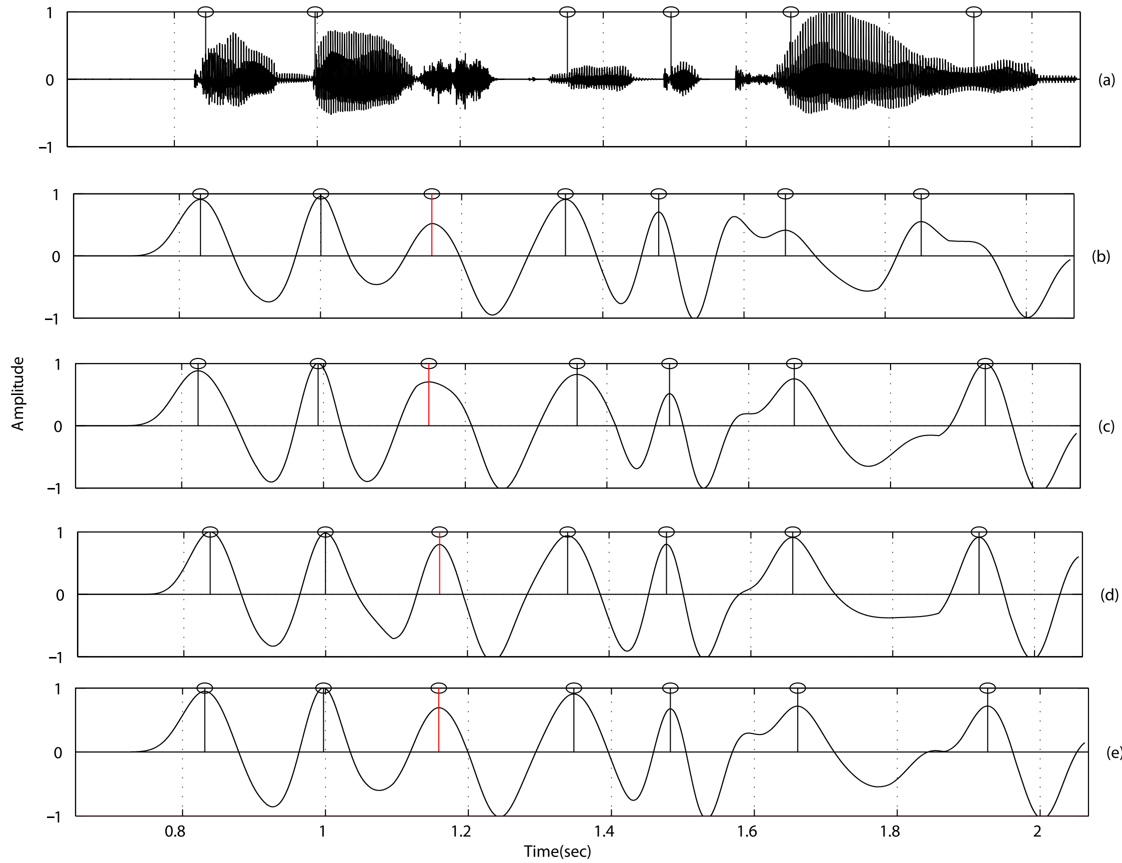

Fig. 1 VOP detection using combination of all three evidence for utterance "*Don't ask me to carry an*". (**a**) Speech signal. VOP evidence plot for (**b**) excitation source. (**c**) Spectral peaks. (**d**) Modulation spectrum. (**e**) Combined [1(b) + 1(c) + 1(d)]

$$\bar{y}(n) = \frac{1}{2N+1} \sum_{n=-N}^{N} y(n) \tag{4}$$

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

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

Epoch extraction using ZFF for the segment of voiced speech is shown in Fig. 2. Figure 2(a) shows the segment of voiced speech, its ZFF signal and the derived epoch locations are shown in Figs. 2(b) and 2(c), respectively.

2.3 Formant extraction using group delay function

Formant extraction from short segments of speech signal using group delay functions presented in Joseph et al. (2006) is used in this work. Short segmental analysis based on conventional spectral methods suffer from the problem of poor resolution in the frequency domain. Hence, high resolution property of group delay can be used for extracting formant frequencies from short segments of speech (Joseph et al. 2006). Group delay ($\tau_g(\omega)$) is defined as

$$\tau_g(\omega) = -\frac{d\phi(\omega)}{d\omega} \tag{5}$$

where $\phi(\omega)$ is phase function and ω is frequency variable. τ_g can be computed directly from signal $x(n)$ as

$$\tau_g(\omega) = \frac{X_i(\omega)X_r'(\omega) + X_r(\omega)X_i'(\omega)}{X_r(\omega)^2 + X_i(\omega)^2} \tag{6}$$

where $X_i(\omega)$ and $X_r(\omega)$ are the imaginary and real parts of Fourier transform of $x(n)$, and $X_i'(\omega)$ and $X_r'(\omega)$ are their derivatives.

$$\tau_g(\omega) \propto \left|X(\omega)\right|^2 \tag{7}$$

It is known that group delay function $\tau_g(\omega)$ of signal around resonant frequency is proportional to square of the magnitude of Fourier transform $|X(\omega)|^2$. Issues associated with the calculation of group delay function are primarily due to zeros present in the denominator term of (6). The denominator term corresponds to the magnitude spectrum of the signal, which is typically large around the formant locations, and hence, it decreases the value of numerator

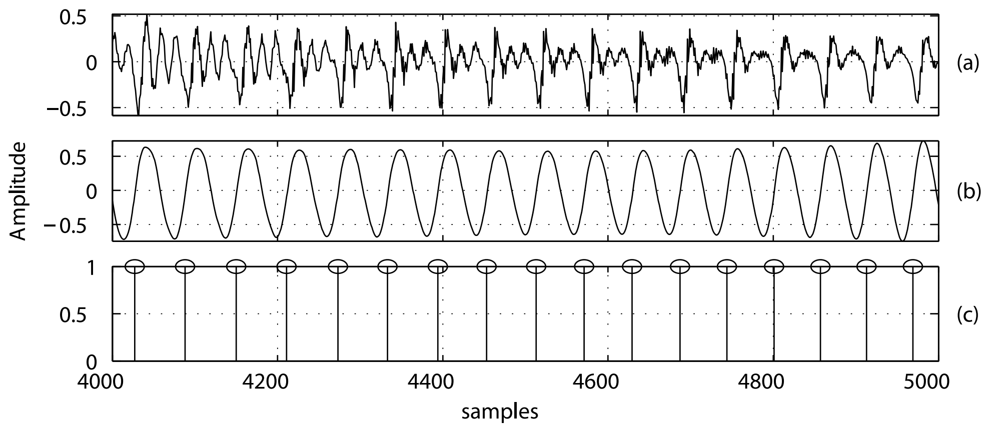

Fig. 2 Epoch extraction using zero frequency filter method. (**a**) Segment of voiced speech signal, (**b**) local mean subtracted ZFF signal and (**c**) epoch locations from ZFF signal

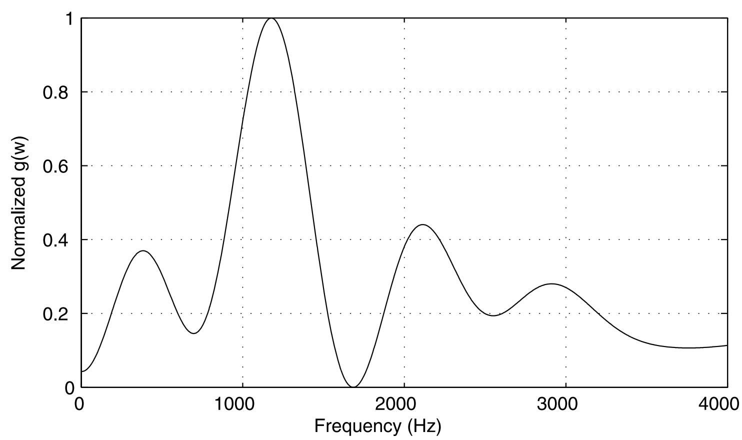

Fig. 3 Numerator $g(\omega)$ of group delay function computed for voiced speech segment of 3 ms duration

around the formant locations. Therefore, numerator ($g(\omega)$) of group delay function is considered for the calculation of formants from short speech segments. Numerator of group delay function $g(\omega)$ is

$$g(\omega) = X_i(\omega)X_r'(\omega) + X_r(\omega)X_i'(\omega) \tag{8}$$

$$g(\omega) \propto \left|X(\omega)\right|^4 \tag{9}$$

At resonant frequencies $g(\omega)$ is proportional to $|X(\omega)|^4$, so $g(\omega)$ gives sharper peaks at resonances than $\tau_g(\omega)$. Numerator $g(\omega)$ of group delay function computed for a voiced speech segment of 3 ms duration is shown in Fig. 3. The peaks in the $g(\omega)$ signal corresponds to the formant locations.

If we synchronize the analysis windows with the epochs, the variation in the configuration of the vocal tract can be captured through the variation in the formant frequencies from one pitch cycle to another. Formant extraction using group delay based method is carried out with the following sequence of steps (Joseph et al. 2006):

- Consider a speech segment present in the glottal closure phase.
- Filter the segment of the speech signal using a half Hanning window of length less than pitch period.
- Compute the $g(\omega)$ function.
- Pick the largest N number of peaks in the computed $g(\omega)$ function.
- Repeat above steps at all glottal closure instants.

3 Proposed VOP detection method

The proposed method exploits the spectral energy at formant frequencies present in the glottal closure region. The reasons for choosing the speech segments at the glottal closure region for deriving the spectral energy are (i) speech during glottal closure phase has high signal to noise ratio, hence high spectral energy, (ii) vocal tract resonances during glottal closure phase are more accurate. There will be a significant change in spectral energy at the formant frequencies from consonant segments to vowel segments. The sequence of steps in the proposed method are:

(1) Determine the epoch locations (glottal closure instants) by using ZFF method.
(2) Compute formants using group delay based method (Joseph et al. 2006) for the speech samples present in 30 % of glottal cycle starting from the GCI. The reason for choosing 30 % of glottal cycle is to ensure that the

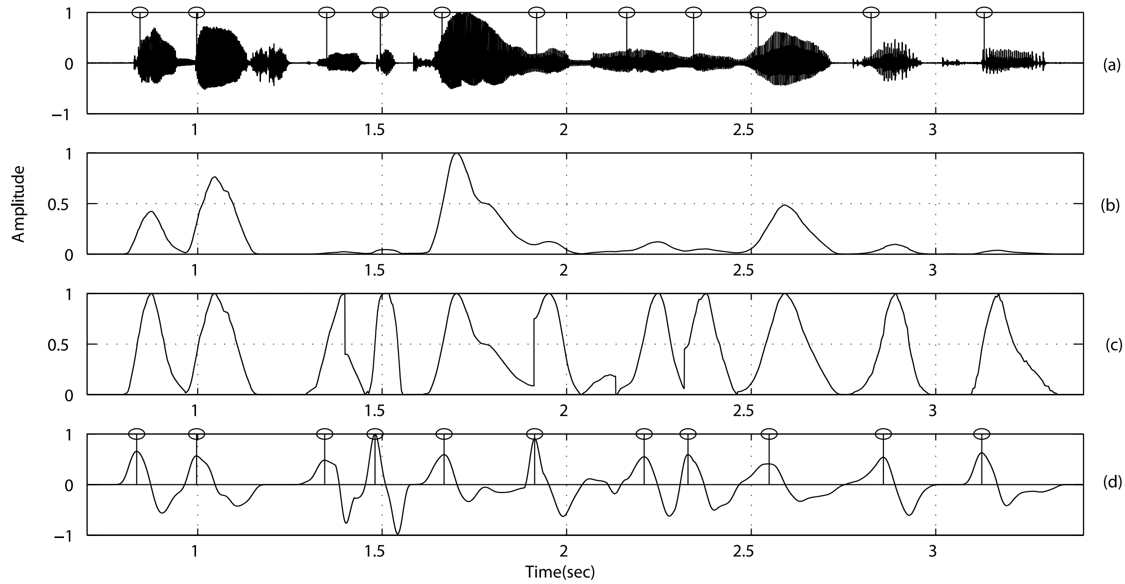

Fig. 4 VOP detection using the proposed method for a speech utterance "*Don't ask me to carry an oily rag like that*". (**a**) Speech signal, (**b**) mean smoothed plot of spectral energy at formant frequencies around each epoch, (**c**) enhanced spectral energy signal, (**d**) proposed VOP evidence plot

chosen speech segment should generate during glottal closure phase.

(3) Determine the sum of spectral energies present at first 3 formant frequencies.

(4) Spectral energy computed in step 3 is plotted as a function of time. Fluctuations in the spectral energy contour are smoothed by using mean smoothing with 50 ms window.

(5) Change at the VOP present in the smoothed spectral energy is enhanced by computing its slope using a first-order difference (FOD). Enhancement of VOP evidence using FOD is described in Prasanna et al. (2009).

(6) The significant changes in the spectral characteristics present in the enhanced version of the smoothed spectral energy are detected by convolving with first order Gaussian difference operator of length 100 ms.

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

The output of each of the steps in the proposed VOP detection method are shown in Fig. 4. Figure 4(a) shows the speech signal "*Don't ask me to carry an oily rag like that*" with manually marked VOPs. Smoothed plot of spectral energy at formant frequencies around each epoch location is shown in Fig. 4(b) (step 4). Figure 4(c) shows the enhanced plot of Fig. 4(b) (step 5). VOP evidence plot obtained from the proposed method is shown in Fig. 4(d) (step 6). We can observe that manual marked VOPs in Fig. 4(a) and detected VOPs marked in Fig. 4(d) are close to each other.

Robustness of the proposed VOP detection method compared to COMB-ESM method is illustrated in Fig. 5 by using white noise added (SNR of 10 dB) speech utterance "*Don't ask me to carry an oily rag like that*". Figure 5(a) shows the speech signal with manually marked VOPs. VOP evidence plots for the speech signal shown in Fig. 5(a) by using the COMB-ESM and proposed methods are shown in Figs. 5(b) and 5(c) respectively. From the Figs. 5(b) and 5(c), we can observe that 4 spurious VOPs are detected in case of COMB-ESM VOP evidence plot, and only 1 spurious VOP is detected in case of proposed VOP evidence plot.

4 Results and discussion

Effectiveness of the proposed VOP detection method is analyzed for two different noise types by using TIMIT database. TIMIT database contains speech files with manual marked phoneme boundaries, and these phoneme boundaries are used for marking the reference VOPs. Detected VOPs by the automatic methods are compared with reference ones to find the deviations, spurious and missed VOPs. Noises considered in this study are white and vehicle noises from Noisex-92 database at SNR levels of 0, 5, 10, 20 dBs. The VOP detection method presented in this work uses speech specific information such as epochs, so present work is suitable only for background noises other than speech-specific noises like babble noise. Hence, babble noise is not considered in this work for analysis.

Performance of the proposed VOP detection method for noisy speech is compared with COMB-ESM method. Table 1 shows the performance of VOP detection methods on TIMIT database under noise. Columns 1 and 2 indicate the type of noise and SNR levels, respectively. Column-3 in Ta-

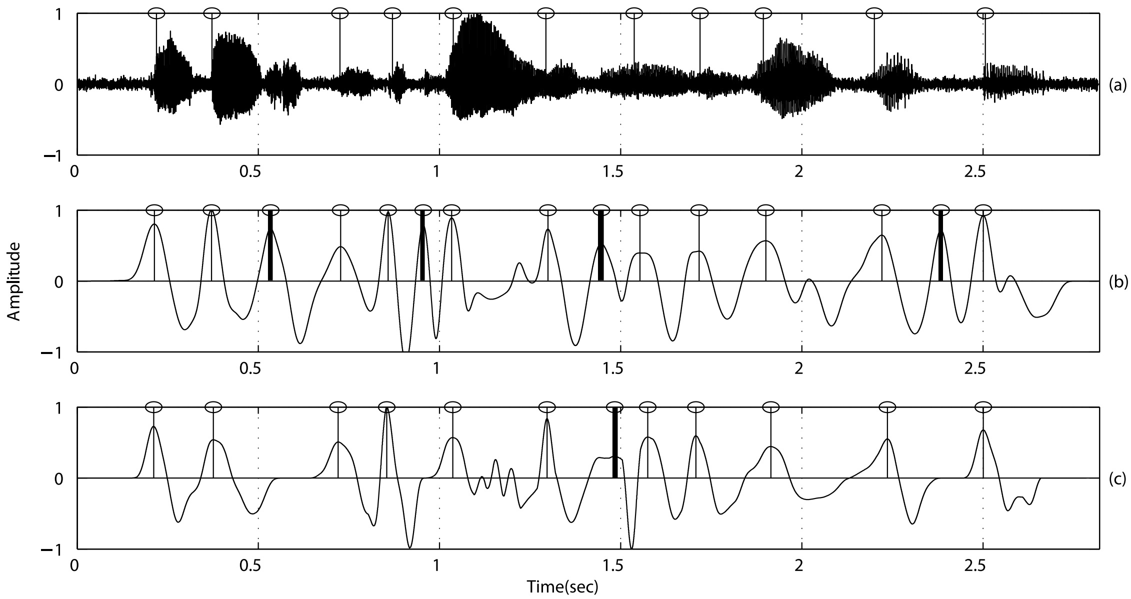

Fig. 5 VOP detection using COMB-ESM and proposed VOP detection methods for the white noise added (SNR of 10 dB) speech utterance "*Don't ask me to carry an oily rag like that*". (**a**) Speech signal with manually marked VOPs, (**b**) VOP evidence plot using the COMB-ESM method, and (**c**) VOP evidence plot using the proposed method

Table 1 Performance of the VOP detection on TIMIT DATABASE using proposed and COMB-ESM VOP detection methods in presence of noise

Noise	SNR	VOP detection method	VOPs detected within ms (%)				AVG dev. (ms)	MISS VOPs (%)	SPU VOPs (%)
			10	20	30	40			
Clean		COMB-ESM	51	59	70	95	16	5	3
		Proposed	65	83	91	95	12	5	2
White	0 dB	COMB-ESM	39	49	62	85	24	15	39
		Proposed	58	71	84	90	19	10	7
	5 dB	COMB-ESM	42	52	65	88	22	12	32
		Proposed	60	72	85	94	14	6	4
	10 dB	COMB-ESM	45	58	68	92	17	8	30
		Proposed	62	77	85	94	11	6	3
	20 dB	COMB-ESM	48	59	71	94	16	6	18
		Proposed	64	77	86	96	10	4	2
Vehicle	0 dB	COMB-ESM	43	54	64	89	21	11	32
		Proposed	59	74	85	92	15	8	6
	5 dB	COMB-ESM	46	57	68	93	19	7	28
		Proposed	62	75	86	94	15	6	4
	10 dB	COMB-ESM	50	59	70	94	17	6	25
		Proposed	64	75	86	95	11	5	4
	20 dB	COMB-ESM	52	60	71	96	16	4	18
		Proposed	65	81	89	96	11	4	2

ble 1 indicates the VOP detection methods considered in this study.

Columns 4 to 7 in Table 1 indicate the percentage of VOPs detected in 10, 20, 30 and 40 ms deviation. Column 8 indicates the average deviation with respect to the manual marked VOPs. Columns 9 and 10 in Table 1 show the percentage of miss and spurious VOPs, respectively. From the results, it is observed that VOP detection performance is severely effected due to noise in-terms of spurious VOPs and average deviation (see Table 1). Performance of the proposed VOP detection method is superior compared to COMB-ESM method under both clean and noisy conditions. Average deviation in VOP detection using the proposed method is reduced around 4 to 8 ms compared to COMB-ESM method. In case of COMB-ESM method, number of spurious detections are very high due to noise at low SNR values (see Table 1), and spurious VOPs are reduced significantly by using the proposed method. This is because of exploiting the high SNR characteristics present at the formant frequencies in the glottal closure phase.

5 Summary and conclusions

In this paper, we proposed a method for detecting the VOPs under noise using spectral energies at formant frequencies of the speech segments present in glottal closure region. Spectral energy at formant frequencies in glottal closure region is high and robust. 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. Performance of the proposed VOP detection method is compared with existing VOP detection method which combines the evidence from excitation source, spectral peaks and modulation spectrum. White and vehicle noises at different SNR values are used to study the performance of VOP detection in presence of background noise. From the conducted studies, it is observed that performance of the VOP detection in presence of noise is severely affected due to spurious VOPs at low SNR values. Performance of the proposed method was observed to be superior compared to existing method, and spurious VOPs were also reduced significantly.

References

Gangashetty, S. V., Sekhar, C. C., & Yegnanarayana, B. (2004a). Detection of vowel onset points in continuous speech using autoassociative neural network models. In *Proc. int. conf. spoken language processing* (pp. 401–410).

Gangashetty, S. V., Sekhar, C. C., & Yegnanarayana, B. (2004b). Extraction of fixed dimension patterns from varying duration segments of consonant-vowel utterances. In *Proc. of IEEE ICISIP* (pp. 159–164).

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

Joseph, M. A., Guruprasad, S., & Yegnanarayana, B. (2006). Extracting formants from short segments of speech using group delay functions. In *Proc. of interspeech* (pp. 1009–1012).

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

Prasanna, S. R. M., & Yegnanarayana, B. (2005). Detection of vowel onset point events using excitation source information. In *Proc. of interspeech* (pp. 1133–1136).

Prasanna, S. R. M., Gangashetty, S. V., & Yegnanarayana, B. (2001). Significance of vowel onset point for speech analysis. In *Proc. of int. conf. signal processing and communications* (pp. 81–88).

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

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

Vuppala, A. K., Rao, K. S., Chakrabarti, S., Krishnamoorthy, P., & Prasanna, S. R. M. (2011). Recognition of consonant-vowel (cv) units under background noise using combined temporal and spectral preprocessing. *International Journal of Speech Technology*, *14*(1).

Vuppala, A. K., Rao, K. S., & Chakrabarti, S. (2012a). Improved consonant–vowel recognition for low bit-rate coded speech. *Wiley International Journal of Adaptive Control and Signal Processing*, *26*(4), 333–349.

Vuppala, A. K., Rao, K. S., & Chakrabarti, S. (2012b). Spotting and recognition of consonant-vowel units from continuous speech using accurate vowel onset points. *Circuits, Systems, and Signal Processing*, *31*(4), 1459–1474.

Wang, J.-H., & Chen, S.-H. (1999). A c/v segmentation algorithm for mandarin speech using wavelet transforms. In *Proc. IEEE int. conf. acoust., speech, signal processing* (pp. 1261–1264).

Wang, J.-F., Wu, C. H., Chang, S. H., & Lee, J. Y. (1991). A hierarchical neural network based C/V segmentation algorithm for mandarin speech recognition. *IEEE Transactions on Signal Processing*, *39*(9), 2141–2146.