Detection of Vowel Onset Point Events using Excitation Information

S.R. Mahadeva Prasanna

Dept. of Electronics and Communication Engg.,
IIT Guwahati, Guwahati-781 039, India
Email:prasanna@iitg.ernet.in

B. Yegnanarayana

Dept. of Computer Science and Engg.,
IIT Madras, Chennai-600 036, India
Email:yegna@cs.iitm.ernet.in

Abstract

This paper proposes a method for the detection of Vowel Onset Point (VOP) events in speech using excitation information. VOP event is defined as the instant at which the onset of vowel takes place. For syllable-like units such as Consonant Vowel (CV) type, VOP event is the instant at which the consonant ends and the vowel begins. The speech signal is processed by the Linear Prediction (LP) analysis to extract the LP residual. The LP residual mostly contains the excitation information. The Hilbert envelope of the LP residual is derived using the analytic signal concept. A method is developed for detecting the VOP events using the Hilbert envelope of the LP residual and a modulated Gaussian window function. The performance of the proposed method is evaluated using reference VOP markings. The performance of the proposed method is also compared with the existing methods based on the vocal tract system features. The comparison shows that the excitation source also contains significant information about the VOP events.

1. Introduction

Speech analysis is usually performed using short (10-30 ms) segments or frames of speech [1]. The analysis frame is positioned arbitrarily with respect to the speech signal. Consequently, the effects of truncation due to windowing may produce variations in the features extracted from successive frames, in addition to the natural variations due to the dynamic nature of speech production. If certain events of significance can be identified and detected, the analysis for feature extraction can be anchored around such events. Such an event-based analysis is likely to produce consistent representation of information for speech analysis [2]. This event-based analysis is motivated by the nature of human speech production. A sequence of changes takes place in the speech production system and are manifested as events in the speech signal. For instance, Vowel Onset Point (VOP) is an event at which their will be significant change in both the vocal tract and the excitation source. The first step in the event-based analysis is to develop a method for automatic detection of the proposed event. The goal of this work is to develop a method for detecting VOP events in speech.

Vowel Onset Point (VOP) event is defined as the instant at which the onset of vowel takes place [3, 4]. It is also termed as Consonant Vowel (CV) segmentation point [5]. Important and discriminatory information for the analysis of speech lies around the VOP event, and hence a reliable method for automatic detection of the VOP event is essential. Kewely-Port et al., [6] have observed that information about the place of articulation in the transition of a stop consonant to a Vowel (V) is confined to a speech segment as short as 20-40 ms. Tekieli and Cullinan [7] demonstrated that the first 10-30 ms of vowels of V and CV units contain important information for analysis and about 40-50 ms duration segment contain enough information to allow the vowel to be perceived properly. Furui [8] showed that a speech wave of approximately 10 ms duration that includes maximum spectral transition bears the most important information for consonant and syllable perception.

Several methods have been proposed in the literature for the detection of the VOP events [3–5, 9]. In all these methods, mainly the vocal tract system features are used in one form or the other to detect the vowel onset points. However, there is significant information in the excitation features, which can be exploited for the detection of vowel onset points. Also it is to be noted that the onset of vowel is an instant property. Most of the existing methods treat VOP as a region property, and hence results in poor resolution, as the VOP is detected by cues with frame level resolution. The aim of this study is to explore the usefulness of excitation features for the detection of VOP events. As the excitation features are complementary to system features, by combining the existing vocal tract features with the proposed excitation features, it may be possible to develop a reliable and robust method for the detection of VOP events.

This paper is organized as follows: In Section 2 extraction of the excitation information from the speech signal is explained. A method for the detection of VOP events using the excitation information is proposed in Section 3. The experimental results and comparison with the results of some of the existing methods is given Section 4. Section 5 gives summary of the various issues discussed in this work and the scope for future work.

2. Extraction of Excitation Information

Linear Prediction (LP) analysis is an approach where the dependencies among adjacent samples of the speech signal are estimated, and then removed from the speech signal to obtain a residual signal. The LP residual signal represents significant characteristics of the excitation source [10]. In the LP analysis each sample is predicted as a linear combination of the past p samples, where p is the order of prediction. The predicted sample $\hat{s}(n)$ at the n^{th} instant is given by

$$\hat{s}(n) = -\sum_{k=1}^{p} a_k s(n-k), \tag{1}$$

where $\{a_k\}$ are the Linear Prediction Coefficients (LPCs).

The error between the actual sample and its predicted value is given by

$$\begin{aligned} e(n) &= s(n) - \hat{s}(n) && (2) \\ &= s(n) + \sum_{k=1}^{p} a_k s(n-k) && (3) \end{aligned}$$

The LPCs are obtained as a result of minimizing the mean square error, which results in solving a set of normal equations given by

$$\sum_{k=1}^{p} a_k R(n-k) = -R(n), \qquad n = 1, ...p, \tag{4}$$

where $R(m) = \sum_n s(n)s(n-m)$ is the autocorrelation function.

The LPCs capture mostly information about the vocal tract system. The information about the excitation source can be obtained from the speech signal by passing the signal through the inverse filter, $A(z) = 1 + \sum_{k=1}^{p} a_k z^{-k}$. The resulting signal is termed as LP residual.

The analytic signal $x(n)$ (in discrete time case) corresponding to the real LP residual signal $e(n)$ is given by

$$x(n) = e(n) + je_h(n), \tag{5}$$

where $e_h(n)$ is the Hilbert transform of $e(n)$, and it is obtained as follows: Let $E(\omega) = E_r(\omega) + jE_i(\omega)$ be the Fourier transform of $e(n)$, where $E_r(\omega)$ and $E_i(\omega)$ are the real and imaginary parts of $E(\omega)$, respectively. The Fourier transform of the imaginary part of the analytic signal $x(n)$ is given by

$$E_h(\omega) = H(\omega)E(\omega), \tag{6}$$

where $H(\omega)$ is the Hilbert transformer, and is given by

$$\begin{aligned} H(\omega) &= -j, \qquad 0 \le \omega < \pi && (7)\\ &= j, \qquad -\pi \le \omega < 0. && (8) \end{aligned}$$

The Fourier inverse of $E_h(\omega)$ gives $e_h(n)$. Thus one can derive the analytic signal corresponding to a given real signal $e(n)$ using Eqn.(5).

An approximation to the Hilbert transformer in discrete frequency is implemented as follows: Let $E(k) = E_r(k) + jE_i(k)$, $0 \le k \le N-1$, be the N-point DFT of the real sequence $e(n)$. The discrete Hilbert transform of $E(k)$ is given by

$$\begin{aligned} E_h(k) &= -jE(k),\ k = 0, 1, \ldots, (N/2-1) && (9)\\ &= jE(k),\ k = N/2, (N/2+1), \ldots, (N-1) && (10) \end{aligned}$$

The N-point inverse DFT of $E_h(k)$ gives an approximation to $e_h(n)$.

The magnitude of the complex analytic signal in Eqn.(5) is called the envelope of the signal [19,20]. We call this magnitude function as the Hilbert envelope of the real signal $e(n)$. The Hilbert envelope $h(n)$ is given by

$$h(n) = \sqrt{e^2(n) + e_h^2(n)} \tag{11}$$

Figure 1 shows a segment of voiced speech, the corresponding LP residual, Hilbert transform of the LP residual and the Hilbert envelope of the LP residual. In this study the Hilbert envelope of the LP residual is used as the excitation information.

3. Detection of VOP Events using Excitation Information

The speech signal is preemphasized and low pass filtered to 2.5 kHz (5 kHz sampling frequency) to select only high SNR regions. The LP residual is computed for every frame of 20 ms with a shift of 10 ms using LP order as 8. The Hilbert envelope of the LP residual is computed using the analytic signal concept as described in the previous section. VOP event is associated with the instant at which their is significant change in the amplitude of the Hilbert envelope of the LP residual begins. For detecting such instants, a modulated Gaussian window function given by the relation $g(n) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{n^2}{2\sigma^2}} cos(\omega n)$ is used. In this study the parameter values are chosen as $\sigma = 100$, where σ is spatial spread of the Gaussian window and $\omega = 0.0114$, where ω is the angular frequency of the sinusoidal component, and window length $n = 800$. The parameters of the modulated Gaussian window function are chosen in such way that the negative part of the window is larger than the positive part, because the time course of suppression is longer than that of excitation for CV units [11]. The modulated Gaussian window function is shown in Figure 2 for the above parameters. The parameters of the modulated Gaussian window function are not very crucial, except that the general shape as shown in Figure 2 is to be maintained.

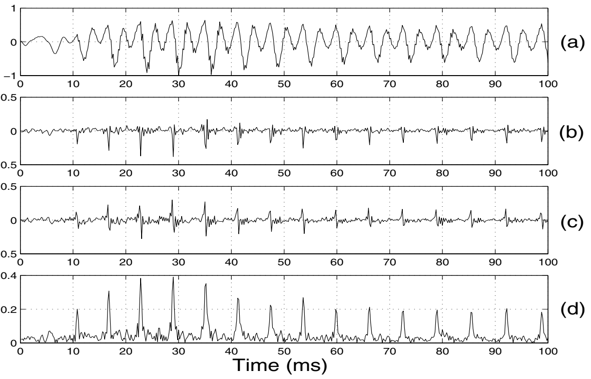

Figure 1: (a) Speech signal, (b) LP residual, (c) Hilbert transform of the LP residual, and (d) Hilbert envelope of the LP residual.

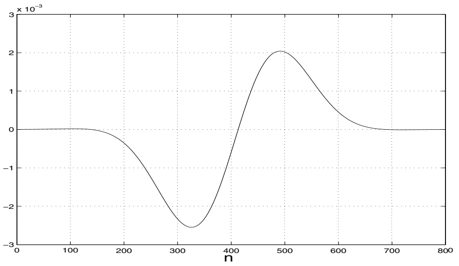

Figure 2: Modulated Gaussian window function for $\sigma = 100$, $\omega = 0.0114$ and $n = 800$.

The Hilbert envelope of the LP residual is convolved with the negative of the modulated Gaussian window function. The convolution output is termed as *VOP Evidence Plot* in this study. In the VOP evidence plot relative maxima occur at the instants where the amplitude in the Hilbert envelope of the LP residual starts rising sharply and this instant is detected as VOP event. The various steps involved in the proposed method for the detection of VOP event is illustrated for the CV unit /khi/ in Figure 3.

In case of isolated utterances of CV units, the maximum value in the VOP evidence plot is hypothesized as the VOP event. However, as there will be multiple VOP events in case of continuous speech, the heuristics for detecting the VOP events

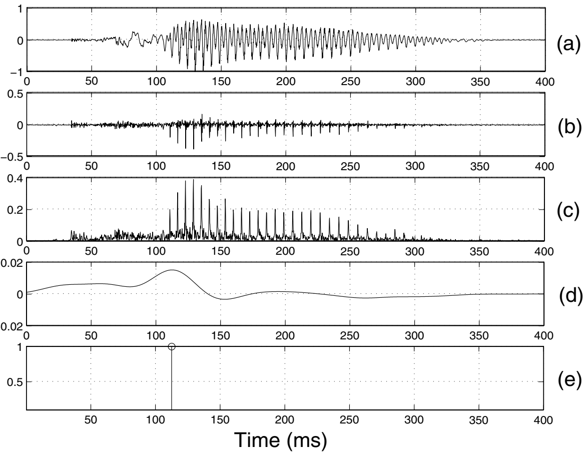

Figure 3: (a) Speech signal, (b) LP residual, (c) Hilbert envelope of LP residual, (d) VOP evidence plot, and (e) Hypothesized VOP.

is modified slightly. Thus for the case of continuous speech, in the VOP evidence plot the peaks are located using a peak picking algorithm. Spurious peaks are eliminated using the characteristics of the shape of the VOP evidence plot, namely, between two true VOPs, there exists a negative region of sufficient strength due to vowel region. For each peak, a check is made for the presence of such a negative region with respect to next peak to eliminate the spurious peaks. The remaining peaks are hypothesized as the VOP events.

The various steps involved in the proposed method for the detection of VOP events in continuous speech is illustrated for the Hindi sentence */antarashtriya bas seva pichale mahine shuru huyithi/*. In this sentence there are 16 VOP events, as marked in Figure 4 (a). The Hilbert envelope of the LP residual and the VOP evidence plots are shown in Figure 4 (b) & (c), respectively. The output of the peak picking algorithm is given in Figure 4 (d), and the hypothesized VOP events after eliminating the spurious ones are shown in Figure 4 (e). Comparing the manually marked VOP events and the hypothesized VOP events, it can be observed that the proposed method has hypothesized 14 VOP events correctly within a deviation of ± 20 ms, 2 VOP events are not detected and 1 is spurious.

4. Experimental Results and Discussions

To evaluate the performance of the proposed method, a reference database containing isolated CV units with manually marked VOP events is used [2]. The efficiency is found out by computing the differences between the hypothesized VOP events and the manually marked VOP events. The Performance of the proposed method for different deviations with respect to the manually marked VOP events is given in Table 1.

For comparison, algorithms based on vocal tract system features like energy derivative method and neural network method are briefly discussed here [4, 9]. In case of energy derivative method the hypothesis is that VOP event is the point at which there is a significant increase in energy in a CV utterance. This point may be detected by computing the derivative of the short-time energy of the speech signal and locating the point at which the positive derivative is maximum. The performance of the energy derivative method for the CV units from the reference database is shown in Table 1. The hypothesis for the neural network method is that the characteristics of the acoustic cues signal energy, LP residual energy and spectral flatness are significantly different in the regions immediately before and after the VOP event. A multilayer perceptron network is trained to detect the VOP event by using the trends in these parameters at the VOP event. The performance of the neural network method for the same reference database is given in Table 1.

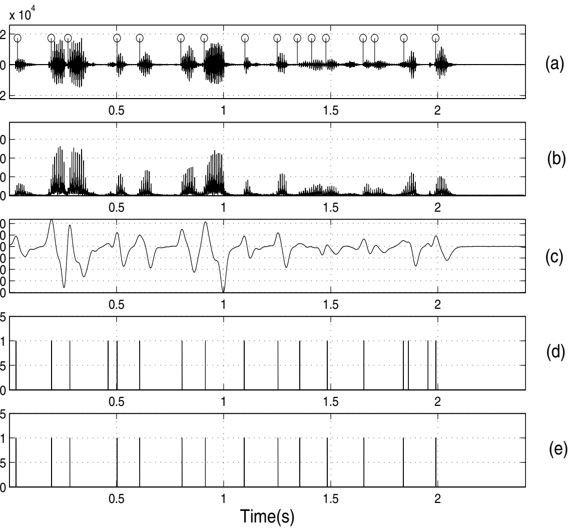

Figure 4: (a) Speech signal, (b) Hilbert envelope of LP residual, (c) VOP evidence plot, (d) Detected peaks as candidates for VOP events, and (e) Hypothesized VOP events.

The performance of the proposed method is better compared to the energy derivative and neural network methods [4, 9]. For a deviation of ± 30 ms, the proposed algorithm detects 88 % of the total 5220 VOP events correctly, whereas the energy derivative method detects only 75.0 % and neural network method detects 86.7 % of the VOP events. Also, the performance of the proposed method is significantly better for small deviation (± 10 ms), which indicates the robustness and high resolution property of the proposed method.

Table 1: The performance (%) of the proposed method using excitation information, Energy Derivative (ED) method and Neural Network (NN) method for VOP detection in isolated utterances of 145 CV classes for different deviations from the VOP. There are in total 5220 VOP events in the reference database.

Sl.No.	Deviation	ED	NN	Proposed
	(ms)	Method	Method	Method
1	± 10	51.5	58.5	61.5
2	± 20	65.4	78.7	81.8
3	± 30	75.0	86.7	88.0

To study the effectiveness of the proposed method in case of continuous speech, 25 sentences from five Hindi news bulletins from five different speakers (2 male and 3 female) are chosen. For each of these sentences, the VOP events are manually marked. There are 236 VOP events in the selected data. Out of the total 236 VOP events in the selected 25 sentences, 209 (88.5 %) are detected within a resolution of ± 20 ms, 22 (9.32 %) are missing and 29 (12.3 %) are wrongly hypothesized. The performance of the proposed method is significantly better for the continuous speech compared to the isolated utterances (88.5 % as opposed to 81.8 % at ± 20 ms deviation). This may be due to the poor articulation of aspirated and voiced CV units in continuous speech, which is common in case of continuous speech.

The energy derivative and neural network methods are mainly developed for the detection of VOP events in case of isolated utterances of CV units. Hence, to compare the performance of the proposed method with that of the method based on system features, the frame energies of the band-pass speech (500-2500 Hz) for blocks of 10 ms with every sample shift are computed. These energy values are used in place of Hilbert envelope of the LP residual. The band energy in the range 500-2500 Hz typically represents high SNR first two formants energy and hence it is assumed to represent the system features. The VOP evidence plot is obtained from the energies computed using the modulated Gaussian window function. The VOP events are hypothesized from the VOP evidence plot as explained earlier. This method is illustrated for the Hindi utterance */antarashtriya bas seva pichale mahine shuru huyithi/*. From Figure 4 and Figure 5, we can see that both the methods hypothesizes the VOP events approximately at the same places.

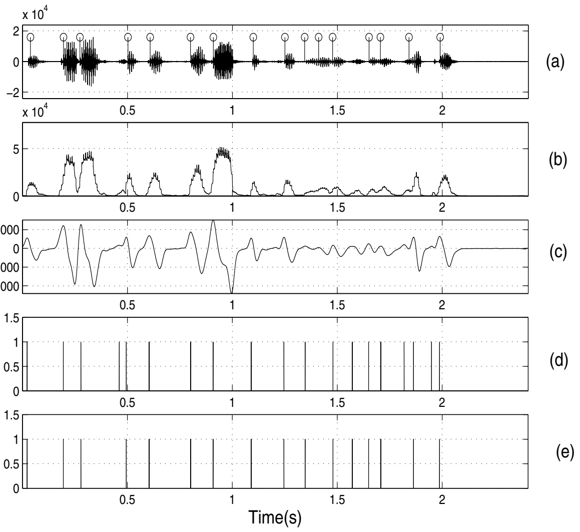

Figure 5: (a) Speech signal, (b)Energy of speech signal, (c) VOP evidence plot, (d) Detected peaks as candidates for VOP events, and (e) Hypothesized VOP events.

Table 2: Performance of proposed algorithm based on excitation features and the algorithm based on system features for VOP detection in continuous speech. In the table the abbreviation HYPO represents hypothesized.

Method	Total VOPs	HYPO VOPs	VOPs in ± 10 ms	VOPs in ± 20 ms	VOPs in ± 30 ms
Excitation feature	236	243	182 (77.1%)	209 (88.5%)	213 (90.2%)
System feature	236	237	186 (78.8%)	210 (89.0%)	211 (89.4 %)

5. Conclusions

In this study a method for the detection of VOP events using excitation information is proposed. In case of isolated utterances of CV units, it was found that 88% of the total 5220 VOP events are detected within a resolution of ± 30 ms. In case of continuous speech, for chosen 25 sentences having a total of 236 VOP events, 88.5% of the VOPs are correctly detected within a resolution of ±20 ms, 9.32% are missing and 12.30% are falsely hypothesized. The performance of the proposed method is comparable and even slightly better compared to the methods based on vocal tract system features.

In this study a method is proposed for the detection of VOP events in isolated and continuous speech using excitation information. Since source and system features are somewhat complementary in nature, it may be possible to combine the merits of both these features to obtain a robust and accurate method for detection of VOP events.

6. References

[1] J. R. Deller Jr., J. H. L. Hansen, and J. G. Proakis, *Discrete-Time Processing of speech Signals*. IEEE Press, 2000.

[2] S. R. M. Prasanna, *Event-based Analysis of Speech*. PhD thesis, Indian Institute of Technology Madras, Department of Computer Science and Engg., Chennai, India, 2004.

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

[4] C. C. Sekhar, *Neural network models for recognition of stop consonant-vowel (SCV) segments in continuous speech.* PhD thesis, Indian Institute of Technology Madras, Department of Computer Science and Engg., Chennai, India, 1996.

[5] J.-F. Wang and S.-H. Chen, "A C/V segmentation algorithm for Mandarin speech signal based on wavelet transforms," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 1, pp. 1261–1264, Sept. 1999.

[6] D. Kewley-Port and D. B. Pisoni, "Perception of static and dynamic acoustic cues to place of articulation in initial stop consonants," *J. Acoust. Soc. Amer.*, vol. 73(5), pp. 1779–1793, 1983.

[7] M. E. Teikeli and W. L. Cullinan, "The perception of temporally segmented vowels and consonant-vowel syllables," *J. Speech Hear. Res.*, vol. 22, pp. 103–121, 1979.

[8] S. Furui, "On the role of spectral transition for speech perception," *J. Acoust. Soc. Amer.*, vol. 80(4), pp. 1016–1025, 1986.

[9] J. Y. S. R. K. Rao, C. C. Sekhar, and B. Yegnanarayana, "Neural network based approach for detection of vowel onset points," in *Int. Conf. Advances in Pattern Recognition and Digital Techniques, (ISI Calcutta, India)*, pp. 316–320, 1999.

[10] J. Makhoul, "Linear prediction: A tutorial review," *Proc. IEEE*, vol. 63, pp. 561–580, Apr. 1975.

[11] R. L. Smith and J. J. Zwislocki, "Short-term adaptation and incremental responses of single auditory-nerve fibers," *Biol. Cybernetics*, vol. 17, pp. 169–182, 1975.