SIGNIFICANCE OF VOWEL ONSET POINT FOR SPEECH ANALYSIS

S. R. Mahadeva Prasanna, Suryakanth V. Gangashetty, and B. Yegnanarayana

Speech and Vision Laboratory
Department of Computer Science and Engineering
Indian Institute of Technology Madras
Chennai-600036, India
email: {prasanna, svg, yegna}@speech.iitm.ernet.in

ABSTRACT

In this paper we discuss the significance of Vowel Onset Point (VOP) for speech recognition in particular, and for accurate speech analysis in general. Vowel onset point is the instant at which the vowel part starts in a Consonant-Vowel (CV) utterance. The syllable-like CV utterances are basic units for both speech production and speech recognition. Recognition performance of isolated utterances of CV units improves significantly, if the parameters/features are extracted using VOPs as anchor points. Even manual identification of VOP requires acoustic cues specific to each subgroup of CV units. We present some acoustic cues which help in manual identification of VOPs in CV utterances. A database of CV units with VOPs marked, together with the knowledge of acoustic cues, will be useful to develop an algorithm for automatic detection of VOPs.

1. INTRODUCTION

Speech analysis is usually performed using short (10-30 ms) segments or frames of speech. The analysis frame is positioned arbitrarily with respect to the speech signal. Consequently the effects of truncation due to windowing produce variations in the extracted parameters/features in successive frames, in addition to the natural variations due to the dynamic nature of speech production. Most production and perception units of speech sounds are dynamic in nature, and therefore syllable-like units are appropriate to describe the production and recognition aspects of speech [1]. Isolating syllable-like units in continuous speech is difficult with the conventional fixed-frame analysis. On the other hand, if certain events of significance for each unit can be identified and detected, the analysis for parameter/feature extraction can be anchored around such events. Such an event-based analysis is likely to produce consistent representation of speech information for applications like speech recognition.

For syllable-like units such as CV type, the onset of vowel is an important event which marks the transition from the consonantal part to the vowel part [2]. The significance of VOP for speech analysis may be summarized as follows. VOPs can be used as anchor points for: (1) Extraction of fixed length patterns so that the neural networks can be used for CV classification, (2) Spotting CV segments in continuous speech, (3) Detecting other important acoustic events like burst, aspiration, formant transition, which also plays important role in speech recognition, (4) Automatic transcription of pitch movements, automatic prosodic analysis [2], (5) Segmenting speech into vowel- and nonvowel-like intervals [3], (6) Finding durations of the vowels, which are important for recognition in Indian languages.

Our preliminary studies on recognition of isolated utterances of CV units show that the recognition accuracy is significantly better with VOP-based analysis of speech, compared to the usual fixed-frame nonevent-based analysis. Of course the recognition performance depends critically on the accuracy of VOP detection in the first place. Careful observation of the characteristics of CV units shows that there are subsets of CV units which are difficult even for manual marking of the VOPs. The features presently being used for manual marking are mostly the amplitude, onset of voicing, and pitch periodicity of the vowel. For difficult cases of CV units such as semivowels and aspirated sounds, it is necessary to use other suitable acoustic cues for manual marking.

The objective of this paper is to discuss the significance of VOP for accurate analysis of speech for various applications, including speech recognition, and to suggest a few acoustic cues to identify the VOPs for difficult cases of CV units. In Section 2 results of recognition experiments are given to demonstrate the importance of VOPs for recognition of CV units. These studies suggest that, even for manual marking of VOPs, there is a need for exploring suitable acoustic cues. The acoustic-phonetic description for the VOPs in the individual subgroups of the CV units is given in Section 3. This description helps in the identification of suitable acoustic cues for the detection of VOPs. The acoustic cues based on the instants of significant excitation, that aid in manual marking of VOPs in CV units are described in Section 4. Section 5 gives a summary of this paper.

2. RECOGNITION OF ISOLATED UTTERANCES OF CV UNITS

Speech recognition experiments were conducted for units of Hindi. The different categories of the 145 CV units considered are shown in Table 1. Note that the CV units are shown only for the vowel ending /a/. Similar CV units exist for the vowel endings /i/, /u/, /e/, and /o/. Twelve utterances of each of these CV units from each of three speakers are collected as speech data for this study. Recognition performance for a subset of 80 CV units was obtained under two different conditions: (a) Parameter extraction for every 10 ms frame and (b) Parameter extraction for 60 ms before and 140 ms after the VOP, where the VOP for each CV utterance is manually marked. The manual marking is based on amplitude, strength of voicing, and pitch periodicity features of the vowel.

Table 1. The table lists different categories of CV units of Hindi language. Out of 145 classes, the ones ending with vowel /a/ only are shown. The stop-consonant-vowels (SCV) can be further grouped based on place of articulation (POA) as velar, alveolar, dental, and bilabial sound units. Depending on the manner of articulation (MOA), the SCVs can be grouped as unvoiced-unaspirated (UVUA), unvoiced-aspirated (UVA), voiced-unaspirated (VUA), and voiced-aspirated (VA).

Category	Sub-Category	Sound units
Stop-consonant-vowels	UVUA	/ka/ (क), /ṭa/ (ट), /ta/ (त), /pa/ (प)
	UVA	/k^ha/ (ख), /ṭ$^h a$/ (ठ), /$t^h a$/ (थ), /$p^h a$/ (फ)
	VUA	/ga/ (ग), /ḍa/ (ड), /da/ (द), /ba/ (ब)
	VA	/$g^h a$/ (घ), /ḍ$^h a$/ (ढ), /$d^h a$/ (ध), /$b^h a$/ (भ)
Nasals		/na/ (न), /ma/ (म)
Semivowles		/ya/ (य), /va/ (व)
Trill		/ra/ (र)
Lateral		/la/ (ल)
Fricatives		/sa/ (स), /ṣa/ (श), /ha/ (ह)
Affricates		/ca/ (च), /$c^h a$/ (छ), /ja/ (ज), /$j^h a$/ (झ)

The main limitation of the Neural Network (NN) models is their inability to model temporal sequences. Hence, for the case (a), Hidden Markov Model (HMM) models are used for recognition [4]. For the case (b), Feed Forward Neural Network (FFNN) models are used [4], [5]. In all the experiments, one-third of the data was used for training, and two-third data for testing the systems. The average recognition performance of the subnets for the subgroups of the 80 SCV classes in these cases is shown in Table 2. It is obvious from the table that VOP information helps in producing better recognition performance.

Table 2. Average recognition performance of subnets for subgroups of the 80 SCV classes using HMMs for case (a) and FFNNs for case (b).

Grouping Criterion	HMM	FFNN manual VOP
MOA	46.9	83.9
POA	57.4	79.3
Vowel	42.7	72.3

Careful examination of the manually marked VOP data shows that there are errors even in this marking, especially in nasals, semivowels, and aspirated sounds. To find the effect of error in the marking of VOP on the performance, recognition experiments were conducted for the patterns extracted with the anchor points at different deviations (in ms) from the manually marked VOPs during testing. The results of the recognition experiment are given in Table 3. As shown in the table, significant degradation in the performance of the recognition system is observed with the increase in the deviation of the anchor point with reference to manually marked VOPs. This is due to the loss of crucial information for recognition. It is conjectured that accurate VOP marking of data would produce significantly better recognition performance. For this purpose we need to determine suitable acoustic cues which is discussed in the following sections.

Table 3. Average recognition performance of subnets for subgroups of the 80 SCV classes using FFNNs for patterns extracted with the Anchor Points (AP) at different deviations (in ms) from the manually marked VOPs.

Grouping Criterion	AP = Manual VOP	AP = Manual VOP $\pm$10 ms	AP = Manual VOP $\pm$20 ms	AP = Manual VOP $\pm$30 ms
MOA	83.0	75.6	61.3	49.6
POA	79.3	69.5	48.0	36.2
Vowel	72.3	59.3	37.2	25.2

3. ACOUSTIC-PHONETIC DESCRIPTION OF VOWEL ONSET POINTS

In this section we give the acoustic-phonetic description for the VOPs in the individual subgroups of the CV units. This description helps in the identification of suitable acoustic cues for the detection of VOPs. Speech signal is the result of exciting a time-varying vocal tract system with time-varying excitation. The changes in the source and system characteristics around the VOP play a significant role in detection of VOPs. Therefore, it is necessary to characterize the changes for each subgroups of CV units. First we describe the VOPs in each subgroup in terms of changes occurring in the source and system characteristics. After this, the acoustic-phonetic description is given for these changes.

In the case of UVUA CV units, the VOPs are characterized by the changes in the source as well as the system characteristics. The change in the source is from the burst release to the quasi-periodic laryngeal source. The change in the system characteristics is from total closure to wide opening. The consonant region is characterized mainly by a burst region, and the vowel region is characterized by pitch periodicity and formant structure.

The VOPs in the case of UVA CV units are characterized by the change in the source characteristics only. This is because the change in the system characteristics occurs at the beginning of the aspiration. The change in the source is from the noise due to turbulent airflow at glottis to the quasi-periodic laryngeal source. The waveform may have high energy in the aspiration, but without pitch periodicity. Formant transition starts in the aspiration region and the formant structure will be approximately same as that of the following vowel, but with some irregularity due to noise source.

For the VUA CV units, change in both the source as well as system characteristics occurs at the VOPs. The change in the system characteristics is from total closure to wide opening. Eventhough the quasi-periodic laryngeal source remains same during the production of consonant as well as vowel, its characteristics changes due to the loading of the system on the source. The consonant region is characterized by a low pitch frequency and low strength of voicing. The vowel region is characterized by the formant structure and high strength of voicing.

The VOPs in the VA CV units are characterized by change in the source characteristics only. The change in the system characteristics occurs at the beginning of the aspiration region. In the initial part of the consonant, we have only quasi-periodic laryngeal source. This will be superimposed by additional noise source at the glottis in the aspiration region. This is followed by change only to the quasi-periodic laryngeal source at the VOP. The strength of voicing may be ambiguous because of the high strength in the aspiration. The onset of formant transition takes place in the aspiration region, and the formant structure will be approximately same as the following vowel but with some irregularity due to the noise source.

In the case of nasal CVs, the VOPs are characterized by the changes in the source as well as system characteristics. To produce nasal consonants, oral cavity is the side chamber and nasal-pharyngeal tube is the main cavity. In the case of nasal vowels, nasal pharyngeal tube is the side chamber and the oral cavity is the main cavity. The change in the source characteristics is due to the loading effect of system on the quasi-periodic laryngeal source. Thus the consonant region is characterized by the low strength of voicing and low formant structure. The vowel region is characterized by regular formant structure and high strength of voicing.

The VOPs in the fricative and affricate CVs may be characterized by the change in the source as well as system characteristics. The change in the source is from noise due to turbulent airflow at a narrow constriction to quasi-periodic laryngeal source. The change in the system characteristics is from narrow constriction to wide opening. The consonant region is characterized by a high frequency formant structure due to frication followed by a regular formant structure in the vowel region.

The VOPs in the semivowels are also characterized by the change in the source as well as system characteristics. The change in the system characteristics is from partial closure to wide opening. The change in source characteristics is due to the loading effect of the system on the quasi-periodic laryngeal source during the production of the consonants. The semivowels bear the closest similarity to vowels and diphthongs, and they are characterized by the formant structure both in the consonant as well as vowel region.

In the case of Trill, VOPs are characterized by the change in source as well as system characteristics. The change in the system characteristics is from constriction at the hard palate to wide opening of the vocal tract system at the VOP. The trill /r/ is characterized by small pulses of vowel-like structure followed by regular vowel structure.

The VOPs in the Lateral are characterized by the change in source as well as system characteristics. The change in the system characteristics is from the constriction at the back of the teeth (alveolar ridge) to wide opening at the VOP. The change in source characteristics is from low strengths of excitation of vocal folds due to loading of system during the production of consonant to high strengths of excitation during the production of the vowel.

4. ACOUSTIC CUES FOR VOP DETECTION

In the previous section we discussed about the acoustic-phonetic description of different subgroups of CV units.

With these descriptions we can manually mark the VOPs using simple acoustic cues available in the waveform or spectrogram of the sound units. For example in case of UVUA CV /ka/, we can mark the VOP using the amplitude and pitch periodicity cues as shown in Fig. 1. But there are difficult cases like nasals, semivowels, and aspirated sounds for example, /mi/, /yo/, and /ghu/ shown in Fig. 2, where it is difficult to manually mark the VOPs. This is because the cues amplitude, pitch periodicity, and strength of voicing shows a similar trend both in consonant regions as well as vowel regions. Hence there is a need for exploring new acoustic cues.

We can obtain acoustic cues based on the instants of significant excitation [6], [7] for marking the VOPs. The advantages of these acoustic cues are: better temporal resolution and robust. The instant of significant excitation primarily refers to the instant of glottal closure within a pitch period in the voiced speech, and to some random instant in the unvoiced speech. The different acoustic cues based on the instants of significant excitation which we are considering for the present study are, formant transition ($F_{tr}(t)$), formant structure ($F_s(t)$), epoch intervals ($E_i(t)$), strength of instants ($S_i(t)$), symmetric Itakura distance ($I_d(t)$), and ratio of the signal energy to residual energy ($S_r(t)$). From the acoustic cues point of view, we characterize VOP as the instant at which there is a significant change in one or more of the above mentioned cues. These acoustic cues are selected after a study of their behavior in several cases of isolated CV utterances. In this section we describe methods for extracting these acoustic cues for VOP marking, and how they can be used for manually marking the VOPs.

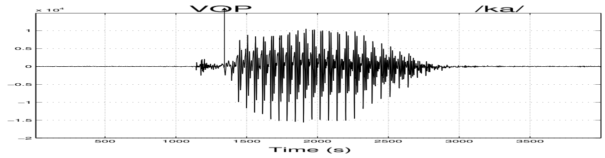

Fig. 1. UVUA CV /ka/ with VOP marked using amplitude and pitch periodicity acoustic cues.

4.1. Formant Transition ($F_{tr}(t)$)

There are, for each consonant, characteristic frequency positions, or loci at which the formant transitions begin, or to which they may be assumed to point [8]. On this basis, the transitions may be regarded simply as movements of the formants from their respective loci to the frequency levels appropriate for the next sound units, wherever these levels might be [9]. For a given sound unit, the formant transition can be seen in its formant contours plot, which is obtained as described in [10]. $F_{tr}(t)$ can be identified as the instant at which uniform formant contours begins. Formant transition

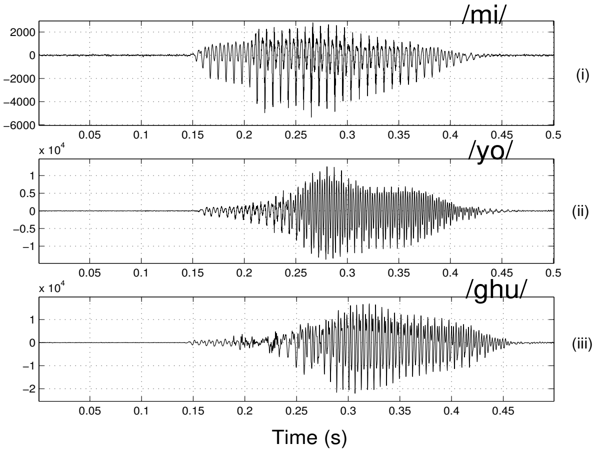

Fig. 2. (i)Nasal /mi/, (ii) Semivowel /yo/, and (iii) Voiced Aspirated CV /ghu/. In these sound units, the acoustic cues amplitude, pitch periodicity, and strength of voicing shows a similar trend both in consonant as well as vowel regions. Hence it is difficult to manually mark the VOPs accurately.

is significant during transition from consonant to vowel, that is, at the VOP. This is due to change in the system characteristics occurring at this point. Hence this can be used as an acoustic cue for VOP marking. Beginning of the $F_{tr}(t)$ (Hz) can be seen in the formant contours plot which is shown in Fig. 3 for the UVUA CV /ka/. Fig. 3 also shows the VOP marked using $F_{tr}(t)$ as a cue. The VOP is identified as the instant at which transition to the following vowel begins.

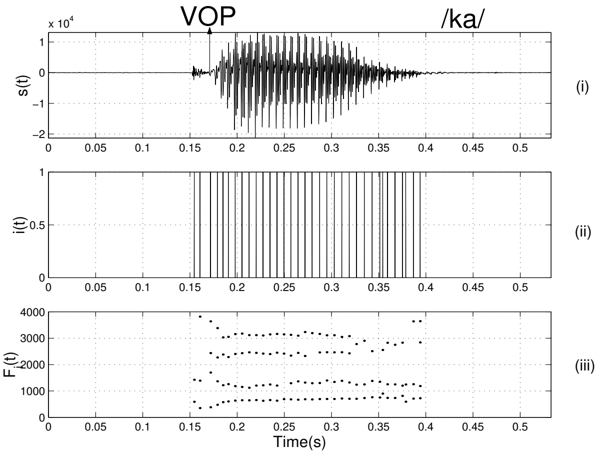

Fig. 3. UVUA CV /ka/. (i) Waveform ($s(t)$), (ii) Instants of significant excitation ($i(t)$), and (iii) Formant contours ($F_i(t)$).

4.2. Formant Structure ($F_s(t)$)

In CV units like nasals, semivowels, and aspirated sounds, trace of formant contours will begin ahead of VOP. In such cases formant transition at VOP may be ambiguous. Hence, we may use change in formant structure ($F_s(t)$) as acoustic cue for manually marking the VOPs. For illustration, let us consider the nasal /ma/ as shown in Fig. 4. A low formant frequency exists in the nasal consonant /m/ region. There is a change in the formants from the low frequency to high frequency at the instant of opening of the oral cavity. This change in $F_s(t)$ (Hz) can be used as an acoustic cue for VOP marking.

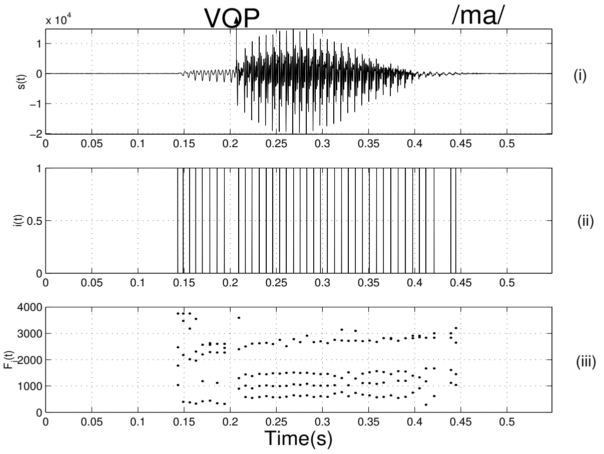

Fig. 4. Nasal CV /ma/. (i) Waveform ($s(t)$), (ii) Instants of significant excitation ($i(t)$), and (iii) Formant contours ($F_i(t)$).

4.3. Epoch Intervals ($E_i(t)$)

The distance between two successive instants of significant excitation is the epoch interval. After extracting the instants, the $E_i(t)$ plot is obtained by plotting at each epoch, the distance of the next epoch as amplitude. The $E_i(t)$ value varies randomly in the case of unvoiced sounds, but remains almost uniform in the case of voiced sounds. Hence this plot will have uniform structure for the voiced sounds. The beginning of such a structure may indicate the VOP, and hence it can be used as an acoustic cue for marking the VOP. Fig. 5 shows the VOP marking for the fricative CV /ṣa/ using the $E_i(t)$ plot.

4.4. Strength of Instants ($S_i(t)$)

The strengths of instants that is, the strengths of excitation mainly depends on the amount of loading of the vocal tract system on the source. The strengths of the instants for voiced sounds is higher compared to the strengths of the

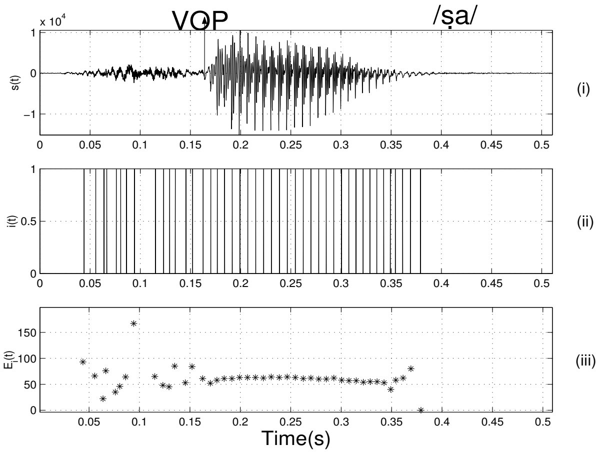

Fig. 5. Fricative CV /ṣa /. (i) Waveform ($s(t)$), (ii) Instants of significant excitation ($i(t)$), and (iii) Epoch intervals ($E_i(t)$).

random instants present in the unvoiced sound. Also the $S_i(t)$ shows a significant change at the transition from unvoiced to voiced sounds. There are CV utterances like UVA CV in which the VOP coincides with the change in $S_i(t)$. The short-term energy computed for a frame of 3 ms speech samples around each instant is considered as the strength of the instant ($S_i(t)$) at that instant. Fig. 6 shows the manually marked VOP for the UVA CV /$k^h a$/ using $S_i(t)$ as the acoustic cue.

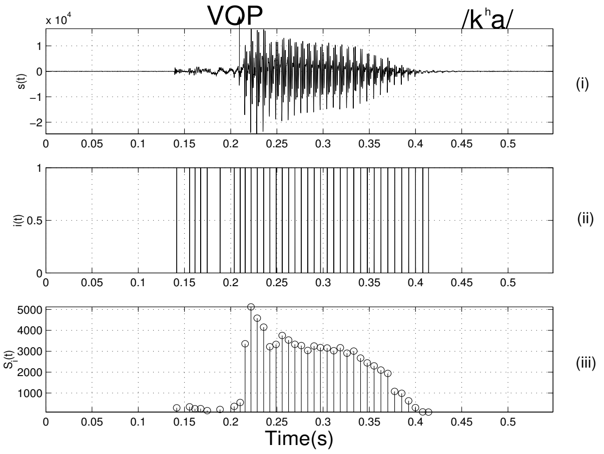

Fig. 6. UVA CV /$k^h a$/. (i) Waveform ($s(t)$), (ii) Instants of significant excitation ($i(t)$), and (iii) Strength of instants ($S_i(t)$).

4.5. Symmetric Itakura Distance ($I_d(t)$)

The change in the system characteristics during the production of the sound units will be manifested as spectral change. The amount of spectral change or distortion can be measured using the Itakura distance [11]. The spectral change is significant at the VOP and hence $I_d(t)$ can be used as an acoustic cue for VOP marking. Fig. 7 shows manually marked VOP for the UVUA CV /ka/ using $I_d(t)$ as the acoustic cue.

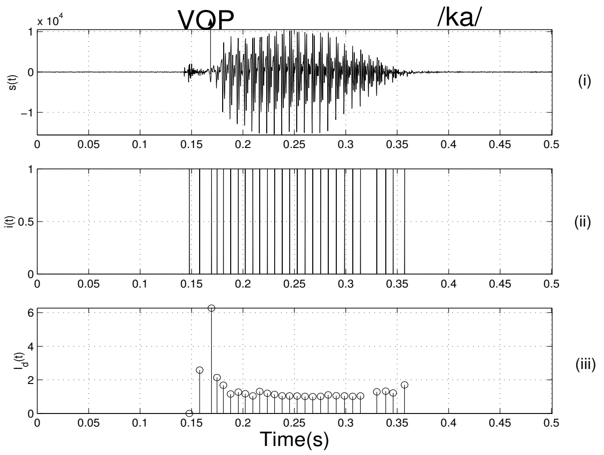

Fig. 7. UVUA CV /ka/. (i) Waveform ($s(t)$), (ii) Instants of significant excitation ($i(t)$), and (iii) Symmetric Itakura distances ($I_d(t)$).

4.6. Ratio of the Signal Energy to Residual Energy ($S_r(t)$)

Energy of the signal is generally higher in the voiced region as compared to the energy in the unvoiced region. In the case of residual signal, the energy level may be higher even in the unvoiced region, comparable to that of the energy in the voiced case. Therefore ratio of the signal energy to the residual energy is low in the consonant region and high in the vowel region, thus indicating a significant change at the VOP. For extracting this acoustic cue short-term energies for 3 ms of the speech signal and residual signal around each instant are computed. The ratio of these two short-term energies is the required acoustic cue. Fig. 8 shows the manually marked VOP for fricative CV /ha/, using this cue.

The summary of descriptions of the acoustic cues is given in Table 4. From the above discussion we can conclude the following: The change occurring in an acoustic cue at the VOP depends on the type of CV utterance. It is obvious that any one acoustic cue is not sufficient for locating the VOP for all classes. We have to use combination of these cues even for manual marking of the VOP. The acoustic cues

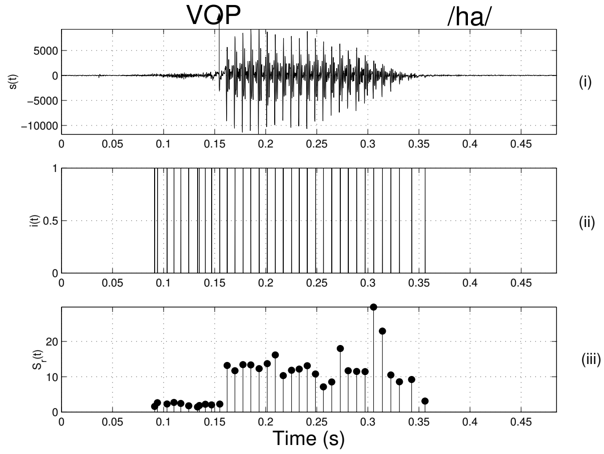

Fig. 8. Fricative CV /ha/. (i) Waveform ($s(t)$), (ii) Instants of significant excitation ($i(t)$), and (iii) Ratio of the signal energy to residual energy ($S_r(t)$).

which are suitable for each category of sound units are given in Table 5. Fig. 9-11 illustrates the proposed acoustic cues for the case of nasal /mi/, semivowel /yo/, and voiced aspirated CV /ghu/. As shown in the figures, we can mark the VOPs by using the acoustic cues as given in Table 5.

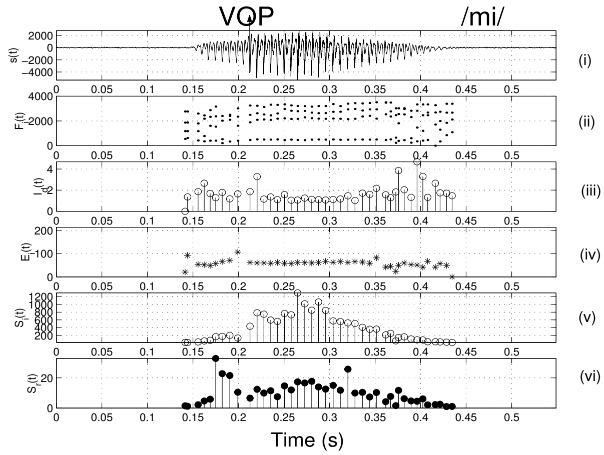

Fig. 9. Nasal /mi/. (i) Waveform (s(t)) with VOP marked, (ii) Formant contours ($F_i(t)$), (iii) Symmetric Itakura distance ($I_d(t)$) (iv) Epoch intervals ($E_i(t)$), (v) Strength of instants ($S_i(t)$), and (vi) Ratio of the signal energy to residual energy ($S_r(t)$).

5. SUMMARY AND CONCLUSIONS

In this study our objective is to discuss the significance of of VOP for accurate speech analysis and to suggest a few a-

Table 4. Summary of the acoustic cues for manual marking of VOP. For each acoustic cue, the entries in the third column, contain the sound units for which the change in the specified cue occurs at the VOP.

Sl. No.	Acoustic Cue	Category of CV units for which this cue is suitable
1	Formant Transition	UVUA CVs, Semivowels, Trill, Laterals
2	Formant Structure	UVA CVs, VA CVs Affricates, Fricatives, Nasals
3	Epoch Intervals	UVUA CVs, UVA CVs, VA CVs, Affricates, Fricatives
4	Strength of Instants	UVUA CVs, VUA CVs, Nasals,Trill
5	Symmetric Itakura Distance	UVUA CVs, VUA CVs, Fricatives, Affricates Nasals
6	Ratio of the Signal Energy to Residual Energy	UVUA CVs, UVA CVs, VUA CVs, VA CVs, Affricates, Fricatives

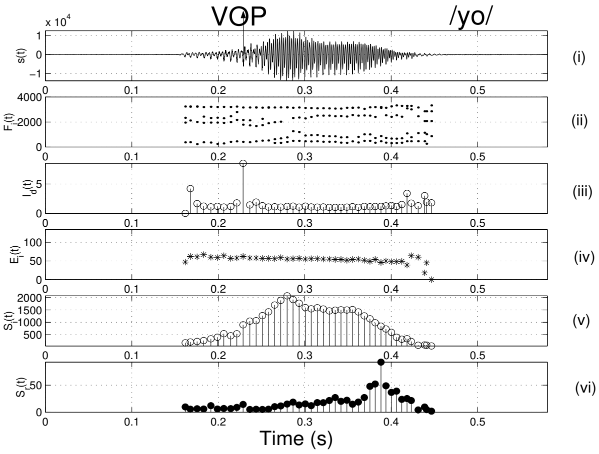

Fig. 10. Semivowel /yo/. (i) Waveform (s(t)) with VOP marked, (ii) Formant contours ($F_i(t)$), (iii) Symmetric Itakura distance ($I_d(t)$) (iv) Epoch intervals ($E_i(t)$), (v) Strength of instants ($S_i(t)$), and (vi) Ratio of the signal energy to residual energy ($S_r(t)$).

coustic cues to identify the VOPs for difficult cases of CV units. From the results of recognition experiments conducted, we have shown that the recognition performance is significantly better with VOP-based analysis of speech compared to the usual fixed-frame nonevent-based analysis. We have also shown the importance of accurate marking of VOPs. Significant degradation in the performance of the recognition system is observed with the increase in the deviation of the anchor point with reference to manually marked VOPs. It is conjectured that accurate marking of VOP would produce significantly better recognition performance. For this purpose we have proposed acoustic cues based on the instants of significant excitation. The use of proposed acoustic cues in manual marking of the VOPs is illustrated.

We are currently preparing VOP marked database of CV units using the proposed acoustic cues. We are also exploring the methods to use these cues for automatic detection of VOP. A VOP marked database of CV units, together with the knowledge of the acoustic cues, will be useful to develop an algorithm for automatic detection of VOPs.

6. REFERENCES

[1] Steven Greenberg, "Speaking in shorthand – A syllable-centric perspective for understanding pronunciation variation," *Speech Communication*, vol. 29, pp. 159–176, 1999.

[2] Dik J. Hermes, "Vowel-Onset Detection," *J. Acoust. Soc. of Amer*, vol. 87(2), pp. 866–873, 1990.

[3] Hideki Kasuya and Hisashi Wakita, "An approach to segmenting speech into vowel- and nonvowel-like intervals," *IEEE Trans. Acoust., Speech, And Signal Processing*, vol. ASSP-27, pp. 319–327, 1979.

[4] C. Chandra Sekhar, *Neural network models for recognition of stop-consonant-vowel (SCV) segments in continuous speech*, Ph.D. Thesis, Dept. of Computer Science and Engineering, IIT Madras, Chennai, India, 1996.

[5] J. Y. Siva Rama Krishna Rao, *Recognition of consonant-vowel (CV) utterances using modular neural network models*, M.S. Thesis, Dept. of Computer Science and Engineering, IIT Madras, Chennai, India, 2000.

[6] B. Yegnanarayana and R. L. H. M. Smits, "A robust method for determining instants of major excitations in voiced speech," *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, vol. 1, pp. 776–779, 1995.

[7] Roel Smits and B. Yegnanarayana, "Determination of instants of significant excitation in speech using group delay functions," *IEEE Transactions on Speech and Audio Processing*, vol. 3, pp. 325–333, 1995.

[8] P. C. Delattre, A. M. Liberman, and F. S. Cooper, "Acoustic loci and transitional cues for consonants," *J. Acoust. Soc. of Amer*, vol. 27, pp. 769–773, 1955.

[9] S. E. G. Ohman, "Coarticulation in VCV utterances: Spectrographic measurements," *J. Acoust. Soc. of Amer*, vol. 23, pp. 151–165, 1965.

[10] B. Yegnanarayana and Raymond N. J. Veldhuis, "Extraction of vocal-tract system characteristics from speech signals," *IEEE Transactions on Speech and Audio Processing*, vol. 6, pp. 313–327, 1998.

[11] P. E. Papamichalis, *Practical approaches to speech coding*, Englewood Cliffs, NewJersy: Prentice Hall, 1987.

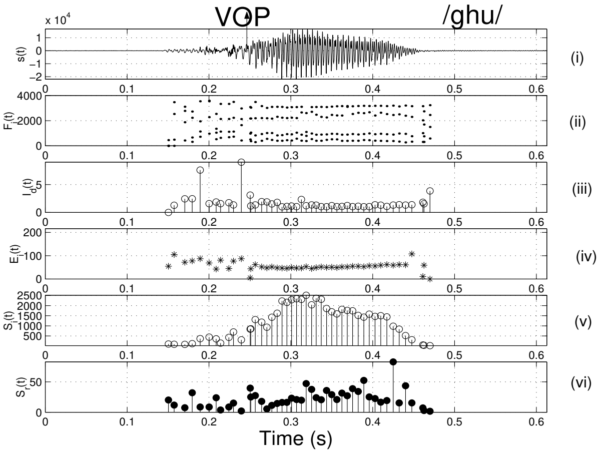

Fig. 11. Voiced Aspirated /mi/. (i) Waveform (s(t)) with VOP marked, (ii) Formant contours ($F_i(t)$), (iii) Symmetric Itakura distance ($I_d(t)$) (iv) Epoch intervals ($E_i(t)$), (v) Strength of instants ($S_i(t)$), and (vi) Ratio of the signal energy to residual energy ($S_r(t)$).

Table 5. Summary of different acoustic cues for manual marking of VOP for each category of sound units.

Sl. No.	Category of sound units	Unambiguous acoustic cues
1	Unvoiced-unaspirated CVs	formant transition, epoch intervals, symmetric Itakura distance, and ratio of the signal energy to residual energy.
2	Unvoiced-aspirated CVs	formant structure, epoch intervals, and ratio of the signal energy to residual energy.
3	Voiced-unaspirated CVs	formant structure, strength of instants, symmetric Itakura distance, and ratio of the signal energy to residual energy.
4	Voiced-aspirated CVs	formant structure and epoch intervals.
5	Nasal CVs	formant structure, epoch intervals, strength of instants, and symmetric Itakura distance
6	Fricative and Affricate CVs	formant transition, epoch intervals, strength of instants, symmetric Itakura distance, and ratio of the signal energy to residual energy.
7	Semivowels	formant transition, strength of instants, and ratio of the signal energy to residual energy.
8	Trill	formant transition, symmetric Itakura distance, and ratio of signal energy to residual energy
9	Lateral	formant structure, strength of instants, symmetric Itakura distance, and ratio of signal energy to residual energy