INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Vowel Onset Point Detection using Sonority Information

Bidisha Sharma and S. R. Mahadeva Prasanna

Dept. of Electronics and Electrical Engineering, Indian Institute of Technology Guwahati, Guwahati-781039,

{s.bidisha, prasanna}@iitg.ernet.in

Abstract

Vowel onset point (VOP) refers to the starting event of a vowel, that may be reflected in different aspects of the speech signal. The major issue in VOP detection using existing methods is the confusion among the vowels and other categories of sounds preceding them. This work explores the usefulness of sonority information to reduce this confusion and improve VOP detection. Vowels are the most sonorant sounds followed by semivowels, nasals, voiced fricatives, voiced stops. The sonority feature is derived from the vocal-tract system, excitation source and suprasegmental aspects. As this feature has the capability to discriminate among different sonorant sound units, it reduces the confusion among onset of vowels with that of other sonorant sounds. This results in improved detection and resolution of VOP detection for continuous speech. The performance of proposed sonority information based VOP detection is found to be 92.4%, compared to 85.2% by the existing method. Also the resolution of localizing VOP within 10 ms is significantly enhanced and a performance of 73.0% is achieved as opposed to 60.2% by the existing method.

Index Terms: Vowel onset point, Sonority, Vocal-tract system, Excitation source, Suprasegmental.

1. Introduction

Vowels are produced by less constricted vocal-tract configuration with very low pressure drop across the constriction. This results in excitation source in the vicinity of glottis with negligible noise [1,2]. Based on the vocal-tract shape and point of constriction, vowels can be divided into low-vowels, mid-vowels and high-vowels with varying acoustic properties [3]. Due to distinct features compared to other sound units (consonants), vowels are manifested as high energy carrier and highly informative in extraction of various attributes required in different applications, such as speech recognition [4, 5], speaker recognition [6], language identification, expressive speech analysis, detection of endpoint of an utterance [7, 8] and syllable rate detection. Therefore, detection of VOP from the speech signal is regarded as an important task in the area of speech processing.

There are several approaches in the literature towards this direction [9–14]. In [9] vowel strength derived from peaks and valleys of amplitude spectrum is used for detection of VOPs. The method explained in [10] uses a product function obtained from the appropriate wavelet and scaling coefficients of input speech signal, that gives higher values in case of vowels. Authors of [11, 12] employed statistical approaches like autoassociated neural network and hierarchical neural network models to detect VOPs. [14] uses spectral energy associated only with glottal closure region for VOP detection that is useful for low bit rate coded speech. In [13], smoothed Hilbert envelope (HE) of Linear prediction (LP) residual (SHE), energy of spectral peaks and modulation spectrum energy (MSE) are used in combination and significant improvement is achieved. Here majority of errors in VOP detection are reported, when vowels are preceded by semivowels, dipthongs, nasals. Due to these errors, in [6] the onsets of semivowels and dipthongs are also included in VOP and termed as vowel like region onset point. Nevertheless, the confusion among vowels and liquid, glide, nasals is not well studied. The classes of sound units like low-vowels, mid-vowels, high-vowels, glides, liquids, nasals share some common attributes in terms of production behavior. These sounds are produced with less vocal-tract constriction (VTC) and more energy associated with glottal vibration. This makes regular structured, high energy and periodic regions in the speech signal. These classes of sound units are together referred as sonorant sounds. Due to the common characteristics, there may be many confusions of the vowels with the other sonorant classes. The sonority associated with a sound unit depends on extent of VTC and energy associated with glottal vibration [15]. It makes vowels as the most sonorant sound. The degree of sonority decreases in the order of glides, liquids and nasals [16]. As the existing features for VOP detection do not have the capability to discriminate between different sonorants, in this work an effort is made to detect VOPs using sonority information. The sonority feature proposed in [17] is based on the study of difference in production behavior of different sonorant sounds and its reflection on vocal-tract spectrum, excitation source, suprasegmental behavior of speech signal. Therefore, it has the ability to bring out difference among sonorants. It can be hypothesized that, adoption of the sonority feature in VOP detection may reduce the confusion among onset of vowels and that of other sonorants.

The rest of the paper is organized as follows: in Section 2, the extraction of the sonority feature is explained. Section 3 describes use of sonority feature in VOP detection. In Section 4, experimental evaluation of proposed features is shown. Conclusion and discussion are reported in Section 5.

2. Sonority Feature Extraction

Despite of some common properties of sonorant sounds with comparatively open vocal-tract configuration, there is variation among the sonorants due to deviation in length, volume and position of constriction. These deviations are reflected in the properties of excitation source along with vocal-tract spectrum (VTS) of the produced speech signal [3]. Based on this fact, a 7-dimensional sonority feature is proposed in [17], which has efficacy to discriminate sounds with varying degree of sonority. In the sonority feature, first 5-dimensions represent VTS characteristics, 6^{th}-dimension represents strength of excitation (SoE) information and 7^{th}-dimension is for representing suprasegmental behavior. These three aspects characterize formant prominence, SoE and periodicity respectively. The feature provides higher values for more sonorant sounds and

http://dx.doi.org/10.21437/Interspeech.2017-790

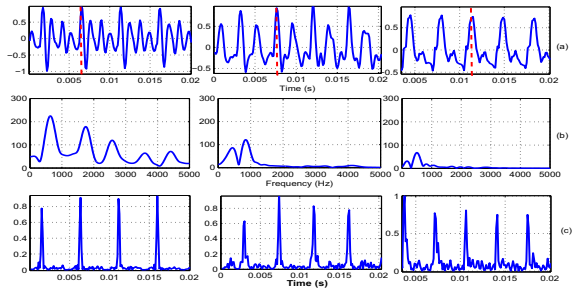

Figure 1: *(a) 20 ms speech segments; (b) HNGD spectra for the speech samples around epoch location in (a); (c) HE of LP residual of speech signal corresponding to vowel, semivowel and nasal respectively from left to right.*

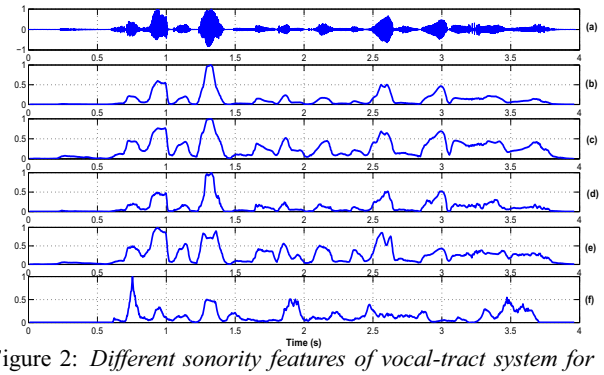

Figure 2: *Different sonority features of vocal-tract system for the utterance "she had your dark suit in greasy wash water all year" taken from TIMIT database. (a) Speech signal, (b) formant peak values, (c) formant peak deviation, (d) amplitude of spectral valleys, (e) slope associated with formant peaks, (f) bandwidth associated with formant peaks*

lower values for sounds with less spectral and source prominence [17]. Due to the coupling between vocal-tract and excitation source during speech production, it is better to consider the VTS only at glottal closed phase. HE of differenced numerator of group delay spectrum (HNGD) is found to have potential in deriving VTS for a very short segment of speech signal around glottal closure instants (GCIs) as reported in [18]. Given the speech signal, GCIs are extracted from zero frequency filtered (ZFF) signal as mentioned in [19] and HNGD spectrum is derived for 5 ms window around each GCI location, that mostly represents glottal closed phase. The HNGD spectra for vowel, semivowel and nasal are shown in Figure 1(b) for the corresponding epoch locations shown in Figure 1(a). A five-dimensional feature representing different attributes of VTS is derived from the HNGD spectrum. The 7-dimensional sonority feature is explained briefly below:

(1) The mean of first three spectral peak values increases with the increase in sonority associated with a sound unit which is represented as f_1, where $f_1 = \frac{1}{3}\sum_{i=1}^{3} P_i$; P_1, P_2, P_3 are the first three spectral peak values.

(2) The mean of relative deviation between amplitudes of first three spectral peaks, $f_2 = \frac{1}{2}\sum_{i=1}^{2} D_i$ where D_1 and D_2 are differences in amplitude of first-second and second-third spectral peaks respectively.

(3) The mean value of amplitude of formant valleys preceding to each spectral peaks is calculated as $f_3 = \frac{1}{3}\sum_{i=1}^{3} Q_i$, where Q_1, Q_2, Q_3 are amplitudes of first three spectral valleys.

(4) In order to detect spectral prominence, mean slope associated with first three spectral peaks is measured as $f_4 = \frac{1}{3}\sum_{i=1}^{3} SP_i$, where

$$SP_1 = \frac{P_1 - Q_1}{F_1 - V_1}; SP_2 = \frac{P_2 - Q_2}{F_2 - V_2}; SP_3 = \frac{P_3 - Q_3}{F_3 - V_3}$$

here, F_1, F_2, F_3 are first three formant frequencies and V_1, V_2, V_3 are location of spectral valleys (in Hz) preceding to first three formant peaks.

(5) In contrast with above features, formant bandwidth increases with the decrease in vocal-tract openness, as it is directly proportional to the loss associated with vocal-tract. The mean of 3 dB bandwidths of first three spectral peaks (B_1, B_2, B_3) is calculated as $f_5 = \frac{1}{3}\sum_{i=1}^{3} B_i$. The contours of all the five vocal-tract features (f_1, f_2, f_3, f_4, f_5) are depicted in Figure 2(b),(c),(d),(e),(f), respectively for the utterance shown in Figure 2(a). It can be observed that features f_1, f_2, f_3, f_4 are directly proportional to the sonority associated with a sound unit, whereas f_5 is inversely proportional to the same. The first four dimensions are normalized and summed up; the resultant feature is added with normalized inverse of f_5 to derive the combined vocal-tract evidence representing sonority which is shown in Figure 3(b) with solid line.

(6) HE of LP residual shows discriminative characteristics between vowels, semivowels and nasals as shown in Figure 1(c), corresponding to the speech segments in 1(a). The feature of excitation source for sonority is defined as $f_6 = \frac{P}{\mu}$, where P is the value of central peak of HE of LP residual at the GCI location and μ is the mean of sample values from 2 ms to 3 ms duration in the 3 ms segment of HE of LP residual segment (1.5 ms left and 1.5 ms right of each GCI). This can be referred as *peak to side-lobe ratio* around the GCIs that can represent SoE. The normalized SoE derived from HE of LP residual is illustrated in Figure 3(c) with solid line.

(7) Sonorant sounds maintain regular structure of speech signal over longer interval. With the increase in sonority level, the correlation among successive pitch periods of speech signal increases. This behavior can be observed for more than a single speech segment. If there are M number of GCIs in the given speech signal, $x_1, x_2, ...x_{M-1}$ are the segments corresponding to $M-1$ number of cycles starting from one GCI to the next. The similarity over K number of cycles (pitch periods) is measured as follows:

$$f_7(i) = \frac{1}{K}\sum_{j=i+1}^{i+K} \frac{< x_i, x_j >}{\sum^{N_i} x_i^2 \sum^{N_j} x_j^2}; \quad i = 1, 2, ...M-1-K \tag{1}$$

where $f_7(i)$ is the correlation coefficient representing suprasegmental evidence of sonority, $< x_i, x_j >$ represents inner product between samples corresponding to x_i and x_j, which are i^{th} and j^{th} pitch cycles in the speech segment. N_i and N_j are the number of samples present in i^{th} and j^{th} cycles. M is the total number of GCIs in the given speech segment and K is the number of cycles over which the similarity measure is calculated. Here, $K = 10$ is considered. The derived suprasegmental feature seems to have more temporal variation

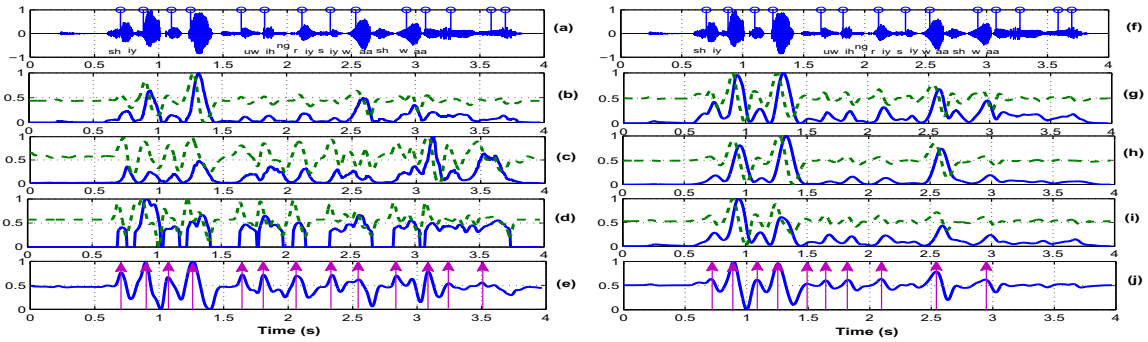

Figure 3: *Steps involved in VOP detection using sonority evidence for the utterance "she had your dark suit in greasy wash water all year" taken from TIMIT database, using sonority feature and existing feature in [13]. (a) Speech signal with reference VOPs, VOP evidence from (b) combined feature of vocal-tract system (c) feature of excitation source (d) suprasegmental feature (the dotted contour in (a), (b), (c) are corresponding FOGD convolved features) (e) combination of FOGD convolved signals in (a),(b),(c). (f)Speech signal with reference VOPs, VOP evidence from (g) energy of spectral peaks (h) modulation spectrum energy (i) smoothed Hilbert envelope (the dotted contour in (g),(h),(i) are corresponding FOGD convolved features) (j) VOP evidence from combination of FOGD convolved signals in (g),(h),(i).*

which effects in VOP detection. Hence in the normalized suprasegmental feature, the values less than 0.2 are made zero and resultant signal is smoothed over an window of 50 ms which is more than one pitch period. The contour of normalized post-processed suprasegmental feature for an utterance from TIMIT database is shown in Fig. 3(d), which carries significantly different information compared to features of vocal-tract system and excitation source.

The distributions corresponding to the normalized VTS system, excitation source and suprasegmental features for all the segments of vowels, semivowels and nasals for entire TIMIT test database are shown in Figure 4. It elucidates the ability of the sonority feature to differentiate between vowels, semivowels and nasals.

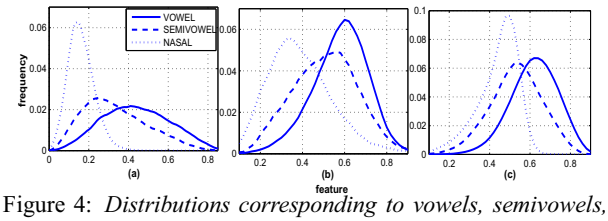

Figure 4: *Distributions corresponding to vowels, semivowels, nasals for (a) vocal-tract system feature, (b) excitation source feature and (c) Suprasegmental feature, for TIMIT test database*

3. VOP Detection Using Sonority Information

Detection of VOP is challenging in case of continuous speech and recent methods for detecting VOPs have high errors when the vowel is preceded by other sonorant sounds [13], as they are not capable to discriminate well between them. In all the three features demonstrated in Fig. 3(b), (c), (d), some changes in feature value can be observed with change of sound unit along the temporal axis. To track these changes each feature is convolved with a first order Gaussian differentiator (FOGD) of length 100 ms (800 samples for 8 kHz) and standard deviation as one sixth of the window length (134 for 8 kHz) [13]. The convolved normalized signals corresponding to each feature are shown with dotted line in Fig. 3(b), (c), (d), which shows peaks at the VOPs that are marked by solid lines in Fig. 3(a) over the speech signal.

The convolved signals corresponding to each of VTS, excitation source and suprasegmental features are added to derive combined evidence of sonority feature as shown in Fig. 3(e). The peaks in the combined evidence in Fig. 3(e) represent the VOPs, which are detected by finding the maximum value between two successive positive to negative zero crossings with some threshold to eliminate the spurious peaks. It can be observed from Fig. 3(a) and (e) that, the sonority feature has potentiality to correctly characterize VOP in continuous speech. After 3 s in Fig. 3(a) although the energy of the speech signal seems to be very low at vowel regions, the combined sonority evidence is able detect the VOPs. Moreover, it is significant to observe from Fig. 3(b), (c), (d) that each evidence carries discriminative information along the utterance irrespective of the wideband energy associated with a particular instant of the speech signal. For example around the instant 3.5 s in the speech signal shown in Fig. 3(a), the feature of VTS exhibits low values with minute variation. On the other hand, both excitation source and suprasegmental features show higher values and prominent variation around 3.5 s and thereby due to the combined effect, the VOP at around 3.5 s is correctly detected.

The VOP detection evidence used in [13] are SHE, MSE and spectral peaks energy which are shown in Fig. 3(g), (h), (i), respectively for the utterance in Fig. 3(f), which infers that all three features show less variation with the change in speech sound in the continuous utterance. Furthermore, for low energy regions the variation in feature value is less. The corresponding FOGD convolved signal is depicted in dotted line over each feature which shows peaks with less strength compared to that of sonority features in Fig. 3(b), (c), (d). Around 0.5 s in the speech signal, [sh] is followed by a high-vowel [iy] as shown in Fig. 3(a). At transition point from [sh] to [iy], the sonority features seem to have a sharp transition compared to features shown in Fig. 3(g), (h), (i). Similar observation can be made for the speech segments of around 2 s, where [ih], [ng], [r], [iy] sounds are continuously uttered which are sonorants. The variation of each sonority evidence can be distinctly observed compared to that of existing evidences. Although the VOPs seem to be detected in this case using existing features in Fig. 3(j), the detected VOP locations are apart from actual VOPs and there is a chance of missing the VOPs. Due to the very low variation

in the existing features, authors of [13] have enhanced peaks in the features and then convolved with the FOGD. Although this post processing adds to detection of some more VOPs, it may introduce many spurious VOPs. Due to these facts, use of combined sonority feature which has potential to discriminate between different sonorant sounds, found to give better performance in case of VOP detection from continuous speech.

4. Experimental Evaluation

For comparison of the sonority feature in VOP detection task with the existing methods, a set of 593 sentences from TIMIT test database (358 male voiced and 235 female voiced) having 6818 number of VOPs is used. These VOPs are manually adjusted by observing spectrograms and waveforms. The features used in [13] for VOP detection are distinctive from the sonority evidence. Detection rate and spurious rate for each of proposed and baseline methods (within braces) are demonstrated in Table 1 for tolerance of $\pm 10\ ms$, $\pm 20\ ms$, $\pm 30\ ms$ and $\pm 40\ ms$ around the true VOPs, which shows significant improvement while using sonority feature. It can be observed that, improvement is more significant in $\pm 30\ ms$ tolerance with 13.6% increase in detection rate. Moreover, when tolerance level changes from $\pm 30\ ms$ to $\pm 40\ ms$, there is highest change in detection rate compared to other tolerance levels.

Table 1: *Performance of sonority evidence in VOP detection for* 593 *sentences comprising of* 6818 *VOPs. Baseline result is shown within braces for method used in [13].*

Evaluation Parameter	**Proposed Method (Baseline Method) (VOPs within ms)**			
	±10	±20	±30	±40
Detection Rate (%)	73.0 (60.2)	77.5 (62.82)	82.4 (68.8)	92.4 (85.7)
Spurious Rate (%)	32.8 (40.3)	25.7 (32.9)	23.6 (28.5)	13.8 (21.1)

Table 2: *Performance of sonority evidence in VOP detection for different CV units with different tolerance levels. Baseline result is shown within braces for method used in [13]. Among* 6818 *VOPs,* 1916 *semi-vowels,* 1475 *fricatives,* 803 *nasals,* 80 *affricates and* 2544 *stops are present.*

Type of CV unit	**Detection Rate (%) of proposed Method (Baseline Method) (VOPs within ms)**			
	±10	±20	±30	±40
semi-vowel	**35.77 (22.55)**	**42.27 (25.96)**	**46.99 (28.86)**	**82.58 (66.84)**
Fricative	84.84 (71.97)	85.67 (74.79)	91.40 (78.38)	96.95 (92.01)
Nasal	73.31 (70.54)	82.83 (73.04)	86.44 (78.21)	89.17 (84.73)
Affricative	83.20 (62.72)	87.59 (64.59)	94.50 (79.67)	98.30 (92.59)
Stop	88.08 (73.02)	90.48 (75.73)	92.50 (79.29)	95.22 (92.53)

Different types of consonants are present in the CV units of 6818 VOPs which can be classified into five categories according to the consonant type. Among 6818 VOPs, 1916 semivowels, 1475 fricatives, 803 nasals, 80 affricates and 2544 stops are present. The detection rate for all the five different categories for both baseline and proposed methods are reported in Table 2. It can be inferred that most of the improvement using proposed method is obtained in terms of vowels which have preceding semi-vowels, although for this category detection rate is lowest. The average improvement in case of semivowels over all tolerance levels is 15.9%. Among all the CV units, most of the VOPs preceded by affricates and fricatives are correctly detected. The least detection accuracy is still in VOPs preceded by semivowels and nasals although the use of sonority evidence has increased the same to some extent. It is important to notice that, along with the increase in detection rate for VOPs preceding semivowels and nasals, it also increases in case of fricatives, affricates and stops while using sonority feature. In [13], energies of first 10 largest peaks of VTS are considered as a feature, but for some high energy fricatives, high amplitude peaks may be present at higher frequency region in the VTS which will yield high values of the feature in case of fricatives also, leading to miss detection of the following VOP in that CV unit. Whereas, in the sonority feature the statistics of first three formant peaks are considered which gives low values in obstruents. Moreover, in fricatives irregular sequence of peaks in HE of LP residual may be present, which will be manifested as higher values in the feature SHE used in [13]. In excitation source information of sonority feature, relation among peak of HE of LP residual at GCI and nearby peaks is considered which will give low values at regions with irregular peaks. These facts may lead to improved detection rate in case of vowels preceded by fricatives, affricates and stops. Along with the VOP detection accuracy, another necessary requirement is robustness of the detection method in noisy scenario. To demonstrate noise robustness of sonority feature in VOP detection, speech signal corresponding to same 593 sentences are added with factory noise, babble noise and white noise, each at different levels of signal to noise ratio (SNR) (0dB, 5dB, 10dB, 15dB) and same VOP detection algorithm is applied. Performance of VOP detection by sonority feature in degraded condition is shown in Fig. 5 with some acceptable level of reduction in performance. Further analysis shows that the noise robustness in the sonority feature is due to the fact that, both VTS and excitation source features are extracted from glottal closed phase, which is less affected by noise. Moreover, the HNGD spectrum is demonstrated to have noise robustness as in [18]. As in suprasegmental feature, the correlation is computed over samples of speech signal, it may have some effect of noise.

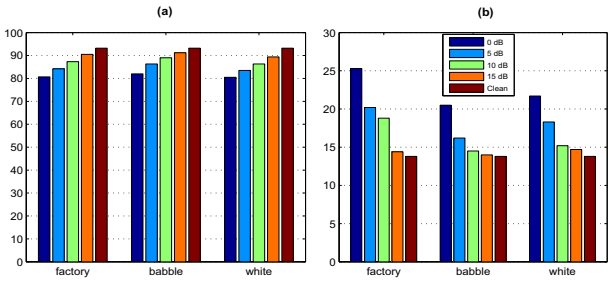

Figure 5: *Bar plot representing (a) detection rate (%), (b) spurious rate (%) of VOP detection within* $\pm 40\ ms$ *tolerance for different types and levels of noise.*

5. Conclusion

This work explores effectiveness of sonority feature in VOP detection. The basic confusion in VOP detection is in nasals and semivowels in CV units which are two classes of sonorant sounds that have less sonority compared to vowels. The sonority feature has ability to discriminate between different sonorants. The fact behind miss detection of VOPs using features in [13] are explained in detail and effective use of different dimensions of sonority feature in those cases is discussed. Apart from sonorants the sonority feature seems to improve performance of VOP detection in case of fricatives, affricates and stops. Moreover, the usefulness sonority feature derived in [17] is explored in case of continuous utterance.The significance of this improvement may be tested in different tasks like consonant-vowel recognition, end point detection and so on.

6. References

[1] D. O'shaughnessy, *Speech communication: human and machine*. Reading, MA: Addison-Wesley,, 1987.

[2] J. R. Deller Jr, J. G. Proakis, and J. H. Hansen, *Discrete time processing of speech signals*. Prentice Hall PTR,, 1993.

[3] K. N. Stevens, *Acoustic phonetics*. MIT press, 2000, vol. 30.

[4] S. Furui, "On the role of spectral transition for speech perception," *The Journal of the Acoustical Society of America*, vol. 80, no. 4, pp. 1016–1025, 1986.

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

[6] S. R. M. Prasanna and G. Pradhan, "Significance of vowel-like regions for speaker verification under degraded conditions," *IEEE Transactions on Audio, Speech, and Language Processing,*, vol. 19, no. 8, pp. 2552–2565, 2011.

[7] S. R. M. Prasanna, J. M. Zachariah, and B. Yegnanarayana, "Begin-end detection using vowel onset points," in *Workshop on Spoken Language Processing*, 2003.

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

[9] D. J. Hermes, "Vowel-onset detection," *The Journal of the Acoustical Society of America*, vol. 87, no. 2, pp. 866–873, 1990.

[10] J.-F. Wang and S.-H. Chen, "A C/V segmentation algorithm for mandarin speech signal based on wavelet transforms," in *ICASSP*, vol. 1. IEEE, 1999, pp. 417–420.

[11] S. V. Gangashetty, C. C. Sekhar, and B. Yegnanarayana, "Detection of vowel onset points in continuous speech using autoassociative neural network models." in *INTERSPEECH*, 2004.

[12] J.-F. Wang, C.-H. Wu, S.-H. Chang, and J.-Y. Lee, "A hierarchical neural network model based on a C/V segmentation algorithm for isolated Mandarin speech recognition," *Signal Processing, IEEE Transactions on*, vol. 39, no. 9, pp. 2141–2146, 1991.

[13] S. R. M. Prasanna, B. S. Reddy, and P. Krishnamoorthy, "Vowel onset point detection using source, spectral peaks, and modulation spectrum energies," *IEEE Transactions on Audio, Speech, and Language Processing,*, vol. 17, no. 4, pp. 556–565, 2009.

[14] A. K. Vuppala, J. Yadav, S. Chakrabarti, and K. S. Rao, "Vowel onset point detection for low bit rate coded speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 6, pp. 1894–1903, 2012.

[15] S. Parker, "Sound level protrusions as physical correlates of sonority," *Journal of phonetics*, vol. 36, no. 1, pp. 55–90, 2008.

[16] S. G. Parker, "Quantifying the sonority hierarchy," Ph.D. dissertation, University of Massachusetts Amherst [Published by the GLSA.], 2002.

[17] B. Sharma and S. R. M. Prasanna, "Sonority measurement using system, source, and suprasegmental information," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 3, pp. 505–518, March 2017.

[18] B. Yegnanarayana and D. N. Gowda, "Spectro-temporal analysis of speech signals using zero-time windowing and group delay function," *Speech Communication*, vol. 55, no. 6, pp. 782–795, 2013.

[19] K. Murty and B. Yegnanarayana, "Epoch extraction from speech signals," *IEEE Transactions on Audio, Speech, and Language Processing,*, vol. 16, no. 8, pp. 1602–1613, 2008.