Detection of Vowel Onset Points in Voiced Aspirated Sounds of Indian Languages

Biswajit Dev Sarma and S. R. M. Prasanna

Dept. of Electronics and Electrical Engineering
Indian Institute of Technology Guwahati, Guwahati-781039, India
{s.biswajit, prasanna}@iitg.ernet.in

Abstract

Vowel onset point (VOP) is defined as the instant at which onset of vowel takes place. Accurate detection of VOP is useful in many applications like syllable unit recognition, end-point detection, speaker verification etc. Manually and automatically locating VOPs accurately in case of voiced aspirated (VA) sounds is found to be difficult and ambiguous. This is due to the complex nature of the speech signal waveform around the VOP. This work addresses this issue and a manual marking approach using electroglottograph (EGG) signal is described which accurately marks the VOPs without any ambiguity. The knowledge derived from this manual analysis is transformed into an automatic method for the detection of VOPs in VA sounds. An automatic method is proposed using both source and vocal tract information. VOP detection accuracy of the proposed method is found to be significantly higher than some of the state of the art techniques.

Index Terms: VOP, EGG, VA SCV, ZFFS

10.21437/Interspeech.2014-339

1. Introduction

Vowel onset point (VOP) is defined as the instant at which onset of vowel takes place [1]. The stop sounds form a major subset of sounds present in most of the Indian languages [2]. There are several attempts earlier to automatically detect the VOPs present in the stop consonant vowel (SCV) units [2] [3] [4] [5] [6]. Detected VOPs are used in different applications like CV unit recognition ([5], [7]), speaker verification ([4], [8]) etc. Major degradation in both manual and automatic VOP detection performance is reported for the case of voiced aspirated (VA) SCV units [6]. This is due to the complexity involved in the excitation component for the production of VA SCV units. Since it is voiced, there is glottal vibration and also aspiration is present at the glottis, as it is aspirated. The following vowel unit is having glottal vibration. Thus, there are three distinct types of excitation of VA SCV units, damped glottal vibration during the closure bar, aspiration overriding glottal vibration in the voice onset time (VOT) region and only glottal vibration in the following vowel region. VOT is the timing interval from the onset of burst till the onset of vowel in a given a SCV unit [9]. Due to high energy of speech signal in the VOT region, there is always ambiguity in manually marking and also automatically detecting VOPs of VA SCV units. VOP is an instant property and hence it should be possible to locate beginning of first glottal cycle associated with vowel region and mark it as VOP. However, this is seldom possible in the speech signal waveform.

The question behind this work is, can we locate the VOPs for VA SCV units, first manually and then automatically? This requires analysis of the excitation component of VA SCV units. The separation of excitation source component always has some degree of uncertainty due to the errors or approximations that may be present in the signal processing procedure employed for source-system decomposition. For instance, in case of linear prediction (LP) analysis, the LP residual obtained using appropriate LP order is treated as the best approximation to the excitation component [10]. However, since it is an error signal, this also includes the error that may create ambiguity in locating or detecting VOP. A velar VA stop unit $[g^h]$ with vowel $[a]$ sampled at 16 kHz and its LP residual using 20^{th} order are shown in Fig. 1(a) and 1(b), respectively. For comparison, the corresponding voiced unaspirated (VUA) counterpart $[g]$ with vowel $[a]$ are also shown in Fig. 1(d) and 1(e), respectively. The marking of VOP in case of VA is ambiguous compared to its VUA counterpart. This begs the need for a method for simultaneous recording of the excitation component directly during the production along with speech signal. Electroglottograph (EGG) device provides one such non-invasive approach for recording information related to the excitation component [11]. Even though EGG does not have one to one correspondence with glottal volume velocity or glottal signal [12], the information provided by it is sufficient to make the current study.

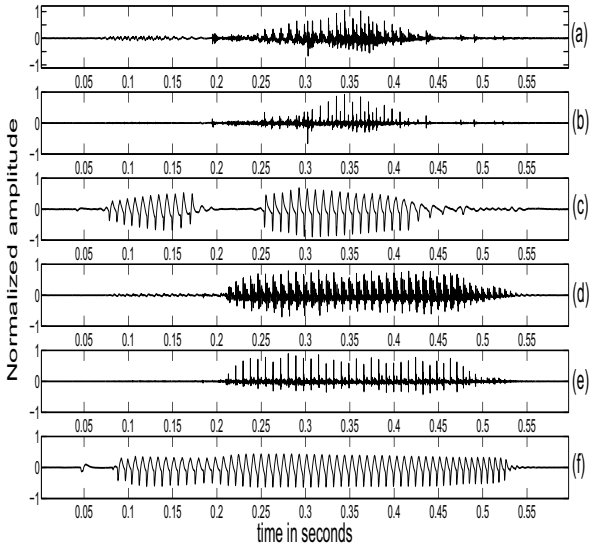

Figure 1: *(a) Speech signal of VA SCV unit $[g^h a]$, its (b) LP residual and (c) EGG signal. (d) Speech signal of VUA SCV unit $[ga]$, its (e) LP residual and (f) EGG signal.*

Fig. 1(c) shows EGG signal for VA SCV unit $[g^h a]$ and

Fig. 1(f) shows the EGG signal for VUA SCV unit $[ga]$. It is interesting to observe the distinction present in the EGG signal among the three regions of VA SCV unit and it is less obvious in the case of VUA SCV unit. This may be explained as follows: EGG records the glottal activity directly as impedance measurement [11]. During the VOT region of VA case, the aspiration component is dominant over the glottal vibration and hence it dictates the impedance measurement. Alternatively, since there is no aspiration, the glottal vibration alone decides the impedance measurement in VUA case. Thus from the EGG signal it is easy to locate the VOP in case of VA sounds compared to VUA case! Hence the use of EGG signal in understanding and marking the VOPs in case of VA SCV units. Zero Frequency Filtered signal (ZFFS) is the output of a zero hertz resonator followed by a trend removal operation (with trend removal window length equal to average pitch period) [13]. ZFFS contains source information and in time domain it looks like a sinusoidal signal which follows the pattern of EGG signal. Thus ZFFS can be used as an evidence to automatically detect VOPs. For accurate VOP detection ZFFS alone may not be sufficient and hence system information from short term energy is used. A combination of these two informations is explored and used as a evidence for VOP detection in case of VA sounds. The signal processing method used for detection of an event from the evidence plays an important role in accurate detection of the event. A signal processing method for accurate detection of VOP is described in this work and its potential is measured using the EGG signal itself as the evidence. Then the method is applied to the evidence extracted from the speech signal and evaluated.

The rest of the work is organized as follows: Section 2 describes the database collected for the study and some initial analysis. The process of manual detection of VOPs by the subjects and the evaluation of manual detection process are described in Section 3. Based on these studies, an automatic method for the detection of VOPs in case of VA units is proposed in Section 4 and its performance is evaluated using the manually marked database. The summary, conclusions and future scope of present work are mentioned in Section 5.

2. Database of VOPs in VA units

The first job was to collect a database of SCV units from different Indian languages. The VA SCV units, in both isolated utterance and embedding in continuous speech are collected. Continuous utterances are formed in Hindi Language. The VA SCV units from different number of speakers of six different Indian languages totaling to 21 speakers was collected. All of them were invited to the recording studio and explained about the procedure for recording the data. The written document containing information about the SCV units to be recorded are given to the subjects and asked them to practice. The VA SCV units are given in Table 1. The subjects were assisted to connect the EGG electrodes to the proper position around the larynx. The head mounted microphone was placed and adjusted in front of the mouth to receive maximum energy. Both the speech data and EGG are simultaneously recorded, sampled and stored in a computer at a sampling frequency of 16 kHz.

Table 1: *VA SCVs in Indian Languages*

	Labial	Dental	Alveolar	Velar
VA	b^ha	d^ha	ɖ^ha	g^ha

Voiced aspirated fricative $[z^ha]$ also has similar characteristics and problems regarding detection of VOP. So this sound is also used for our study along with the four VA SCV unit shown in Table 1. Each of the subjects have to give ten examples for each of the 5 units. From 21 speakers, a total of 1050 examples were collected. All these examples are collected when they are producing the VA units in isolated utterance fashion. The same VA units are also embedded in continuous speech sentences and 600 VA units are excised from the continuous speech for comparison with isolated case. Each of the examples have the speech signal and corresponding EGG signal.

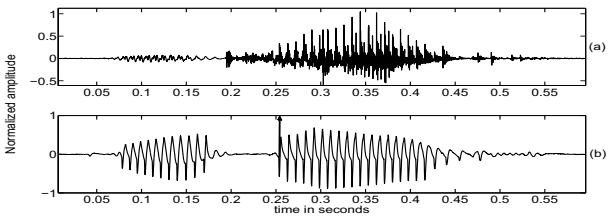

Figure 2: *(a) Speech signal for $[g^ha]$ and (b) its EGG signal. Arrow mark shows the manual VOP marking*

3. Manual Marking of VOPs in VA Units

Five subjects are involved in the process of manual marking of VOPs. The EGG and corresponding speech signal are loaded in two separate audacity panels and the subjects are explained about the characteristics of glottal signal in the closure, aspiration and vowel regions. After this the procedure for marking the VOPs in case of VA units is explained. The procedure suggested for manual marking was the following: (i) load the EGG and speech signal for a given VA unit into audacity waveform panels, (ii) zoom out to display only portion of closure, complete aspiration and a portion of vowel region till the first glottal cycle of the vowel region is clearly visible, (iii) mark the instant of beginning of first glottal cycle of the vowel as the VOP as illustrated in Fig. 2 for the SCV unit $[g^ha]$. In case of few VA units, due to week glottal vibration, the EGG signal may have few cycles with very low amplitude in the aspiration region. In such cases, marking should be done at the transition from low to high amplitude. All isolated and continuous VA units are equally divided among the subjects for marking ensuring that 100 units are common across all the five subjects. These 100 units are used for evaluating the agreement among all the subjects. Ideally it is expected that all the subjects mark the instant of beginning of the first glottal cycle of vowel region. However, due to human error, the marked instant may deviate from the actual instant. The agreement among the subjects are evaluated by calculating standard deviation. Standard deviation is found to be 4.2 ms indicating less ambiguity present in the EGG signal for marking the VOP. Thus EGG signal can be used for manual marking of VOPs in case of VA units.

4. Automatic Detection of VOPs in VA Units

To use the knowledge of VOPs in speech processing tasks like speech recognition and speaker recognition, a method for automatic detection of VOPs is needed. There are several methods in the literature [6] [2] [3] [4]. Careful analysis of the results indicate that one of the factors contributing to poor performance is the VA category. Hence new methods are required to take care of this category alone.

4.1. Basis for Automatic Detection

VOP is defined as a temporal event occurring at a resolution of about 100 ms, where event is an instant property [6]. In the

given EGG or speech signal, there are events occurring at lower resolution (as low as 10 ms) like glottal closure and glottal opening. These events need to be smoothed out and only changes occurring at about 100 ms resolution to be preserved. A differential Gaussian window of 100 ms duration will do this job as illustrated earlier [4]. Since VOP is an event and needs to be hypothesized with better temporal resolution, a small standard deviation is preferred. A gaussian window function is defined as,

$$g(n) = \frac{1}{2\pi\sigma} exp^{\frac{1}{2}(\frac{n-\mu}{\sigma})^2}. \quad (1)$$

Its first order difference is given by,

$$g_1(n) = g(n) - g(n-1). \quad (2)$$

Its second order difference is given by,

$$g_2(n) = g_1(n) - g_1(n-1). \quad (3)$$

The first order difference results in negative peak and its negative positive peak, at VOP [4]. Alternatively, the second order difference results in negative going zero crossing and its negative, a positive going zero crossing at VOP. This work, as a choice, propose to use positive going zero crossing as representative of hypothesized VOP. The signal evidence derived from the speech signal is convolved with the negative of second order difference of Gaussian window and positive zero crossings are hypothesized as possible candidates for VOPs.

Small amplitude variation of signal evidence may lead to some spurious zero crossings. Therefore these small variations needs to be smoothed before convolving with the Gaussian differentiator. One easy way of smoothing is by doing moving average filtering. But some evidences are already averaged over a single or multiple frames. In that case, further averaging will reduce the accuracy of VOP detection. Alternatively, the evidence is passed through a nonlinear operation where high variations at the VOP is further enhanced and small variations are reduced and smoothed. As a result spurious positive zero crossing in the output signal is reduced. The nonlinear operation is given by [6]:

$$E_n = \frac{1}{1 + e^{-(E-\theta)/\tau}} \quad (4)$$

where,E and E_n are the evidences before and after performing the nonlinear operation, θ (=0.2) and τ(=0.04) are the slope parameters. Fig. 3 shows the speech signal in (a) and its energy contour (as an evidence for VOP) in (b) along with the signal after convolving the evidence with the second order Gaussian differentiator (SOGD) in (c). There are many spurious zero crossings can be seen here. (d) shows the evidence after passing through the nonlinear operation and (e) shows the corresponding convolution output. After passing through the nonlinear operation most of the spurious zero crossings are eliminated. Remaining spurious zero crossings are eliminated by computing the slopes at each of the zero crossings and choosing the ones that are above certain threshold. The spurious ones can also be eliminated by increasing the window size by keeping the standard deviation fixed at 10 ms (say, 150, 200, 250 and 300 ms). Since the evidence at VOP is robust, if it is the candidate for VOP, then zero crossing will be present in the same vicinity for all the cases and may shift locations randomly in case of spurious ones.

4.2. Potential of proposed method: Testing using EGG

Since manual marking of VOPs in case of VA units is done using EGG signal in the previous section, the evidence in EGG at

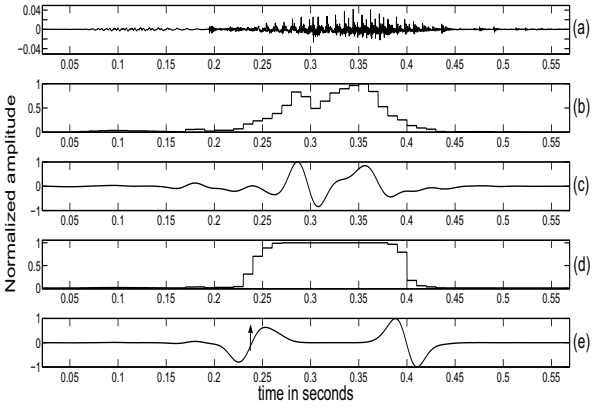

Figure 3: *(a) Speech signal for $[g^h a]$, (b) its short term energy contour, (c) convolved output of energy contour with the SOGD, (d) evidence after passing through nonlinear operation and (e) convolved output of nonlinearly mapped evidence with the SOGD. Arrow mark shows the detected VOP at the positive zero crossing.*

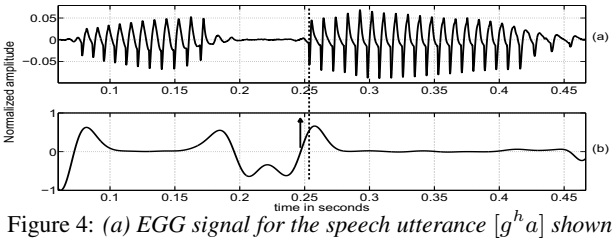

Figure 4: *(a) EGG signal for the speech utterance $[g^h a]$ shown in 3(a) and (b) Final VOP evidence obtained from the EGG signal after applying the signal processing method on it. Arrow mark shows the detected VOP at the positive zero crossing. The dotted line shows the manual VOP marking*

VOP is unambiguous. Hence it may be better to check the basis of proposed method described above using the EGG signal. Given a signal evidence containing VOP evidence, how accurately will the above mentioned procedure will detect VOPs. The method while using in EGG may detect spurious VOPs (due to the signal in the voice bar region), but the aim is to see the accuracy of the automatic method in detecting the true VOPs. The EGG signal of an isolated VA unit is passed through the nonlinear operation and convolved with the negative of SOGD which is shown in Fig. 4(b). We can observe a positive zero crossing at the VOP as indicated by the manually marked VOP shown in Fig. 4(a).

Table 2: *VOP detection accuracy of the method using EGG signal as the evidence.*

Deviations	10 ms	20ms	30 ms	40 ms
	91.45	96.65	98.82	99.53

Table 2 shows accuracy of detection in detecting the true VOPs. This results show that, if the evidence is strong and unambiguous, then we can achieve high performance using above suggested method.

4.3. Automatic Detection of VOPs using Source and System Information

The zero frequency filtering (ZFF) is proposed earlier for the detection of glottal closure instants (GCIs) from speech [13]. The

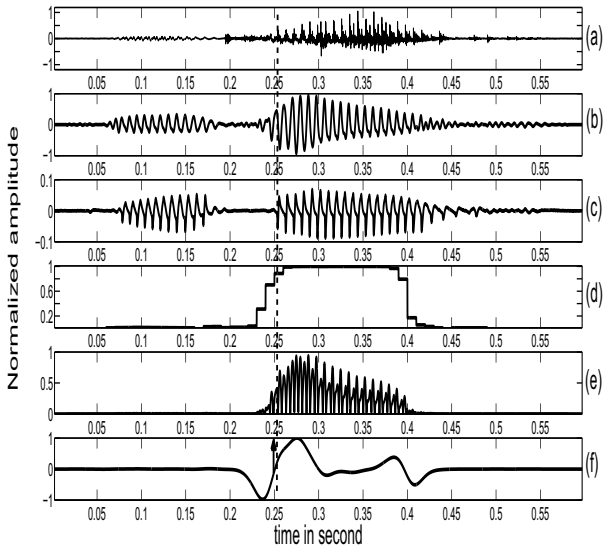

Figure 5: *(a) Speech signal for utterance 'gha'. (b) ZFFS of the speech signal in a). (c) EGG signal of the speech signal in a). (d) Short term energy after passing through a non linear operation. (e) ZFFS signal after multiplying with the signal in d). (f) Convolution output of the evidence in e) with SOGD. Arrow mark shows the detected VOP at the positive zero crossing and dotted line shows the manual marking*

speech signal is passed through the fourth order zero frequency resonator (ZFR) and the output of ZFR is trend removed using blocks of size equal to average pitch period resulting in ZFFS. The ZFFS only contains excitation source information, especially the glottal signal information. The speech signal corresponding to the EGG signal shown in Fig. 4(a) is plotted in Fig. 5(a) and its EGG in Fig. 5(c). The ZFFS of the speech signal is shown in Fig. 5(b). The ZFFS follows the nature of EGG signal. However, the nearly periodic nature begins slightly ahead of VOP. This is because, the trend removal with window size of average pitch period length emphasis changes at pitch period level. Since, the glottal vibration is weakly present in aspiration region it gets emphasis after ZFF. Therefore there is a need for exploring the vocal tract information. The short term energy of speech signal contains system information and normally energy increases at the VOP from low to high. To emphasize

Table 3: *Performance of VOP detection using source, spectral peaks and modulation spectrum (SSM) based method, excitation source (ES) based method and the proposed method*

Mode	Method	DR	SR
Isolated VA unit	SSM	95.66	6.58
	ES	96.13	4.72
	Proposed	97.25	4.37
VA unit in Continuous speech	SSM	91.75	9.93
	ES	93.92	8.73
	Proposed	94.68	8.08

the evidence it is passed through a nonlinear operation (given in Eqn. 4). The output of the nonlinear operation (in Fig. 5(d)) is multiplied with ZFFS and the output (in Fig. 5(e)) is convolved with SOGD which is given in Fig. 5(f). The hypothesis of VOP is slightly ahead of its actual occurrence. To remove spurious zero crossings, 1) the window size of SOGD is varied and deviation of zero crossing positions are calculated from the original

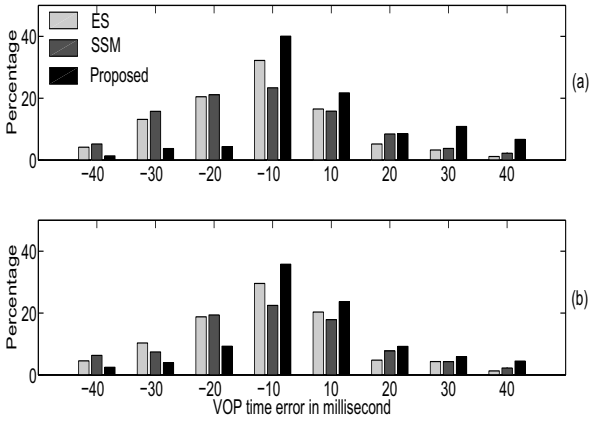

Figure 6: *VOP detection accuracy in terms of percentage of VOP detected within 10, 20 30 and 40 ms time errors. VOP detection accuracy for a) isolated VA unit, b) VA unit excised from continuous speech, using excitation source (ES) based method, source, spectral peaks and modulation spectrum (SSM) based method and the proposed method*

position. If average deviation is more than 10 ms it is hypothesized as spurious zero crossing. 2) slope at the zero crossing is computed and a threshold is set (0.2 to 0.3). If slope is less than the threshold it is a spurious zero crossing. Performance is evaluated for both isolated utterances of VA unit and the same in continuous speech. The performance of VOP detection for different factors is given in Table 3. The performance of VOP detection is measured using the following parameters:

- Detection rate (DR): Percentage of VOPs that are detected within 40 ms of manual markings;
- Spurious rate (SR): Percentage of VOPs that are detected beyond 40 ms of manual markings;
- Detection Accuracy: Percentage of VOPs that are detected within 10 ms, 10 to 20 ms, 20 to 30 ms and 30 to 40 ms. This is shown by plotting histograms.

Performance is compared with existing excitation source (ES) [4], and source, spectral peaks and modulation spectrum (SSM) based VOP detection techniques [3]. Detection rate and spurious rate are comparable and detection accuracy is much better for proposed method (shown by plotting histogram in figure 6). Percentage of VOP detected within 10 ms is increased by around 10%.

5. Summary and Conclusion

This paper describes a procedure for manual marking of VOPs in VA units in an unambiguous fashion using EGG signal. A method for automatic detection of VOPs in VA units is demonstrated. The basis of automatic detection method is verified using the EGG signal. Source information from ZFF signal and system information from short term energy signal are used as the evidences for VOPs in the automatic detection procedure. Result is compared with existing techniques and a significant improvement is achieved in terms of accuracy of detection. The proposed method can be used for the analysis of acoustic features around the detected VOPs.

6. Acknowledgement

This work is part of the ongoing project titled *Development of Prosodically guided phonetic Engine for Assamese language* funded by the TDIL, DeitY, MC&IT, GoI

7. References

[1] D. Herms, "Vowel onset detection," *J. Acoust.Soc. Amer*, vol. 87, pp. 886–873, 1990.

[2] C. Sekhar, "Neural network models for recognition of stop consonant-vowel (scv) segments in continuous speech," Ph.D. dissertation, Department of Computer Science and Engineering, IIT Madras, 1996.

[3] S. R. M. Prasanna, B. V. 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, May 2009.

[4] G. Pradhan and S. R. M. Prasanna, "Speaker verification by vowel and nonvowel like segmentation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 4, pp. 854–867, April 2013.

[5] 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, August 2012.

[6] S. R. M. Prasanna, "Event based analysis of speech," Ph.D. dissertation, Department of Computer Science and Engineering, IIT Madras, 2004.

[7] A. K. Vuppala, K. S. Rao., and S. Chakrabarti, "Improved consonant-vowel recognition for low bit-rate coded speech," *Int. J. Adapt. Control Signal Process*, vol. 26, pp. 333–349, 2012.

[8] 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, Nov 2011.

[9] P. Bhaskararao, "Salient phonetic features of indian languages in speech technology," *Sadhana*, vol. 36, pp. 587–599, 2011.

[10] T. Ananthapadmanabha and B. Yegnanarayana, "Epoch extraction from linear prediction residual," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing, USA*, April, 1978.

[11] A. S. Krishnamurthy and D. G. Childers, "Two-channel speech analysis," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 34, pp. 730–743, Aug. 1986.

[12] D. G. Childers, D. M. Hooks, G. P. Moore, L. Eskenazi, and A. L. Lalwani, "Electroglottography and vocal fold physiology," *J. Speech. Hear. Res.*, vol. 33, pp. 245–254, Jun. 1990.

[13] K. S. R. Murthy and B. Yegnanarayana, "Epoch extraction from speech signals," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 8, pp. 1602–1613, November 2008.