Speech Prosody 2016
31 May - 3 Jun 2106, Boston, USA

Syllable nucleus and boundary detection in noisy conditions

Sreedhar Patha, Yegnanarayana Bayya and Suryakanth V Gangashetty

Speech and vision lab, International Institute of Information Technology, Hyderabad, India
sreedhar.patha@research.iiit.ac.in, yegna@iiit.ac.in and svg@iiit.ac.in

Abstract

In this paper, utilizing the feature of the sonority, the problems of detecting the syllable nuclei and the syllable boundaries are explored. As there is minimal obstruction in the oral cavity in the production of sonorant sounds, sonority feature is used. A method is proposed for extracting the sonority profile from the strength of the dominant resonance frequency using the zero time windowing method. The strength of the dominant resonance frequencies in the specific band of 500 Hz to 1700 Hz is used. The proposed method is evaluated on SVL-DD-NEWS-corpora. The results are compared with the Fourier transform method and the energy-based method of extracting the sonority profile. The proposed method performed better compared to other methods, especially in the presence of noise such as white, pink and babble. The method is also evaluated on TIMIT test database and the performance is on par with the current methods in detecting the syllable nuclei.

Index Terms: syllable nuclei, syllable boundary, sonority, dominant resonance strength.

1. Introduction

Segmenting speech into syllables helps in many speech applications such as speech recognition, synthesis and spoken term detection [1]. Detecting the syllable nuclei and the syllable boundary has been a challenging task. Syllables are considered as basic elements for study of speech prosody [2]. Many studies have addressed the syllable nuclei detection separately [3][4], and only a few detected both the syllable nucleus and the syllable boundary by a single feature [2][5].

In Mermelstein's method, loudness measure obtained from short-time power spectrum is used to derive the sonority profile and convex-hull algorithm is applied recursively to get final segmentation results [6]. Most of the current syllable segmentation methods are derived from Mermelstein's framework and are evaluated on TIMIT database [7]. As TIMIT database does not have the syllable level transcription, they are obtained using phoneme level transcription with some available softwares like Tsylb [8]. The results were posted over these syllabified transcription. While many of the approaches considered intensity, a few of them incorporated voicing decision along with supervised techniques to detect the syllable nucleus and the syllable boundary. In [2], the syllable nucleus and the syllable boundary are detected, using the sonority profile generated by fusing intensity and voicing profiles from various frequency regions. In [5] the sonority profile was extracted from the temporal envelope represented by intensity in the band ranging from 300 Hz to 1000 Hz.

1.1. Syllable and Sonority

The best description of the syllable is from the speech production point of view. It is said that *"speaking is modified breathing"* [9]. This syllable producing movement of respiratory muscles has been called as chest / breath / syllable pulse. Syllable is also described as *"minimal pulse of initiatory activity bounded by a momentary retardation of initiator, either self imposed, or more usually, imposed by a constant type of articulatory stricture"* [10]. As it is difficult to define the syllable, in [11] Ladefoged tries to define the syllable in terms of the inherent sonority of each sound. He also suggests that *"the syllables must be marked not by the peaks in the sonority but by the peaks in the prominence"*, which means that, in a syllable there must be only one sonority peak, but if there is more than one peak, then the syllable nuclei has to be marked according to their level of the sonority. The general form of the syllable is given by $C^*\mathrm{V}C^*$, where C^* is a consonant or consonant cluster and V is a vowel. Every syllable must consist of a nucleus, and it optionally consists of onset and/or coda. The onset and coda are segments which occurs before and after the nucleus respectively. Sievers first attributed the sonority as sound fullness and explained it as relative loudness of speech [12]. Ladefoged further described the sonority as the loudness relative to other sounds having the same length, stress and pitch [11]. A sonorant sound tends to have low degree of acoustic loss leading to reduction in the formant bandwidth [13]. The hierarchy of sonorant sounds is given by :

low vowels $>$ mid vowels $>$ high vowels $>$ nasals $>$ fricatives

The behavior of sonority within the syllable is explained by the sonority sequencing principle, which states that sonority increases from onset to nucleus and decreases to coda [12]. This hill-shaped profile signifies the sonority, and can be used for detection of the syllable nucleus and the syllable boundary. Sonority is mostly linked to the intensity in a particular frequency band. In [5], 300 Hz to 1000 Hz is used as the sonority band for American English and in [14] 500 Hz to 1700 Hz is used for British English as the sonority band.

2. Speech database

The proposed method is evaluated on SVL-DD-NEWS-corpora [15]. This database consist of hand-labeled syllable level marking for three Indian languages, namely, Telugu, Tamil and Hindi. Each language has about five hours of news bulletins broadcasted by Doordarshan. The audio signal was digitized at 16 kHz sampling rate with 16 bits per sample. The digitized data was spliced into sentences, and further into phrases if the sentence exceeds 3 secs. ITRANS code was used for transcription [16]. As Indian languages have more syllable structure than other languages, SVL-DD-NEWS-corpora is a perfect set for evaluation. Moreover this database has manual syllable level markings which is more appropriate ground truth for evaluation than any tool generated ground truth. But for comparison with other methods in literature, the proposed method is also evaluated on TIMIT database.

 doi: 10.21437/SpeechProsody.2016-74

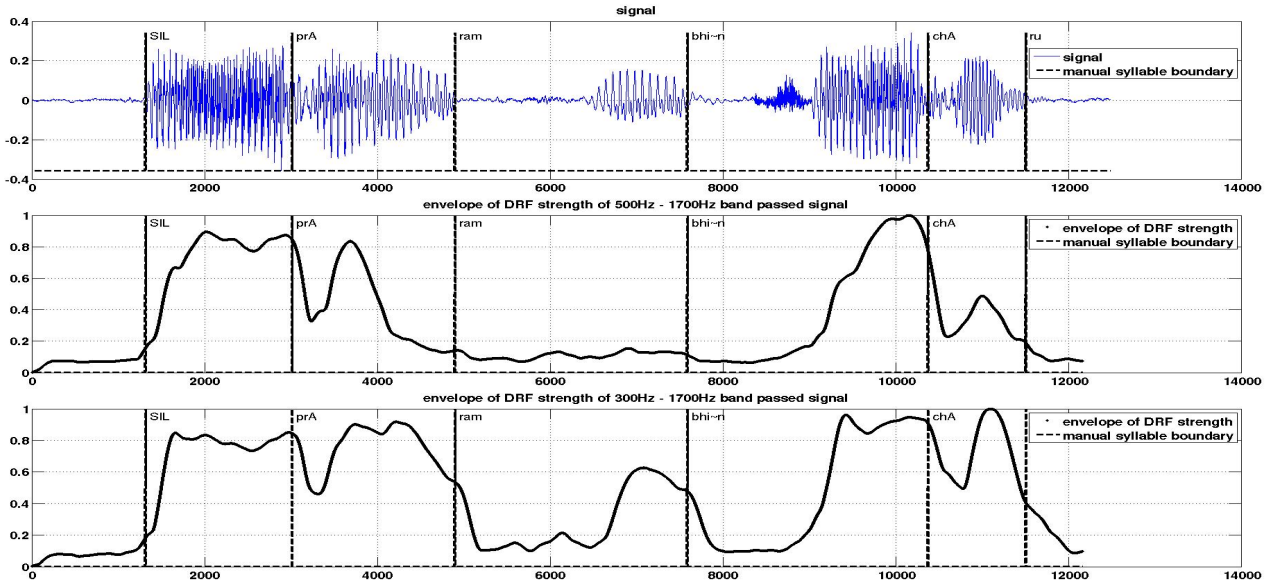

Figure 1: (a) Speech signal of Telugu utterance "prArambhi~nchAru", (b) The sonority contour in the 500 Hz - 1700 Hz frequency band, and (c) The sonority contour in the 300 Hz - 1000 Hz frequency band

Table 1: Syllable nucleus detection in the band of 500 Hz to 1700 Hz for SVL-DD-NEWS-corpora

Method	Insertion(%)	Deletion(%)	TER(%)
Proposed	19.71	07.09	26.80
FFT based	12.59	11.04	23.63
Energy-based	05.02	14.45	**19.47**

Table 2: Syllable boundary detection in the band of 500 Hz to 1700 Hz for SVL-DD-NEWS-corpora

Method	Insertion(%)	Deletion(%)	TER(%)
Proposed	22.29	14.66	36.95
FFT based	15.44	18.34	33.78
Energy-based	08.76	22.29	**31.05**

Table 3: Syllable nucleus detection in the band of 300 Hz to 1000 Hz for SVL-DD-NEWS-corpora

Method	Insertion(%)	Deletion(%)	TER(%)
Proposed	18.42	10.39	28.81
FFT based	09.59	18.77	28.36
Energy-based	03.96	23.14	**27.10**

Table 4: Syllable boundary detection in the band of 300 Hz to 1000 Hz for SVL-DD-NEWS-corpora

Method	Insertion(%)	Deletion(%)	TER(%)
Proposed	21.49	18.87	40.36
FFT based	12.89	26.81	39.70
Energy-based	07.83	31.56	**39.39**

3. Proposed method for the syllable nucleus and the syllable boundary detection

In the proposed method, the envelope of the amplitude of the dominant resonance frequency (DRF) obtained from ZTW method is used for the sonority profile.

In the ZTW method [15][17][18], a heavily decaying window is multiplied with the speech signal. The windowed signal is represented by,

$$x[n] = s[n] * h[n] \tag{1}$$

where $s[n]$ is the speech signal and $h[n]$ is the window function, given by,

$$h[n] = \begin{cases} 0, & n = 0 \\ \frac{1}{8sin^4(\pi n/N)}, & n = 1, 2...N-1 \end{cases} \tag{2}$$

The analysis window $h[n]$ is shifted for every sample. The spectral characteristics of the windowed signal $x[n]$ are obtained using the Hilbert envelope of the numerator group delay (HNGD) function. The numerator group delay is given by,

$$g(\omega) = X_I(\omega)X'_R(\omega) - X_R(\omega)X'_I(\omega) \tag{3}$$

where $X(\omega) = X_R(\omega) + jX_I(\omega)$ is the **DTFT** of $x[n]$ and $X'(\omega) = X'_R(\omega) + jX'_I(\omega)$ is the **DTFT** of $nx[n]$.

These spectral characteristics can be interpreted as instantaneous because of the heavily decaying window and single sample shift. The strongest resonance in the HNGD spectrum is called as the DRF. The DRF can be equivalently associated with dimensions of the prominent cavity in the vocal tract responsible for the production of the sound [17]. The amplitude of the DRF is called the dominant resonance strength (DRS).

3.1. Procedure

The speech signal is passed through a band pass filter with cut-off frequencies 500 Hz and 1700 Hz. The Dominant resonance

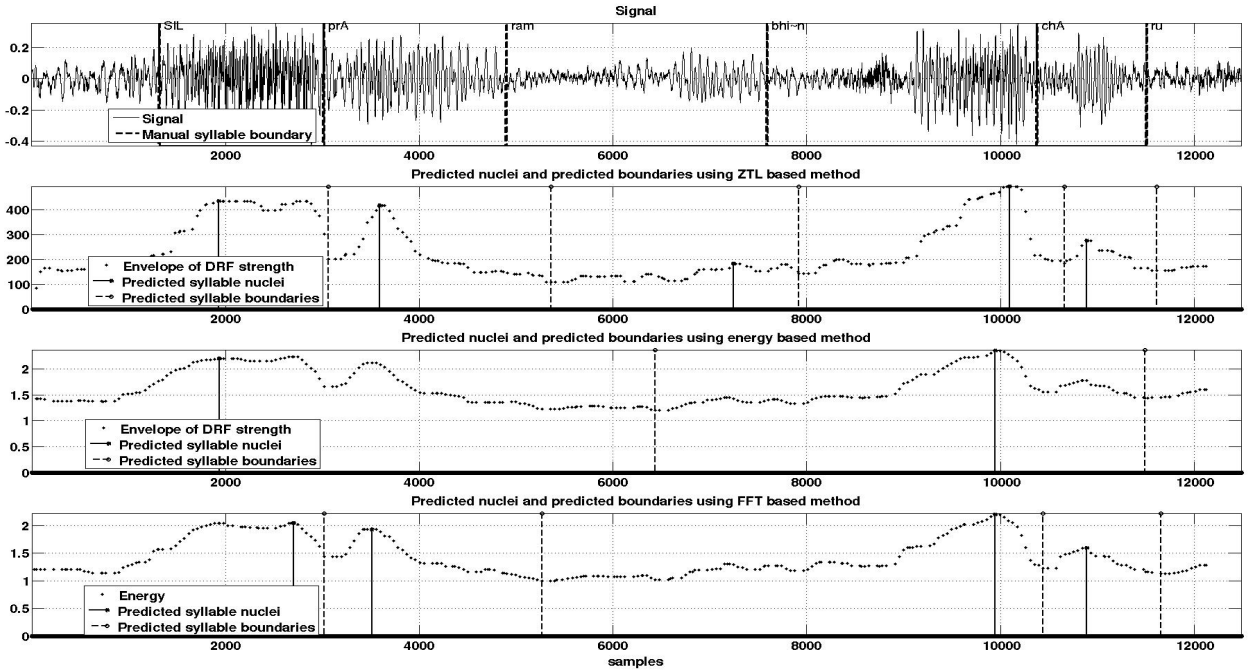

Figure 2: (a) Speech signal of Telugu utterance *"prArambhi~nchAru"* with 5 dB Babble noise. Syllable nuclei and their boundaries detection of corresponding utterance using (b) the proposed method, (c) the Fourier transform method and (d) the energy-based method.

Table 5: Syllable nucleus detection in the presence of noise for SVL-DD-NEWS-corpora

Noise	SNR (dB)	Method	Insertion (%)	Deletion (%)	TER (%)
White	5	Proposed	07.40	17.88	**25.28**
		FFT based	01.14	31.64	32.78
		Energy-based	00.12	51.90	52.02
	0	Proposed	04.40	29.63	**34.03**
		FFT based	00.46	47.54	48.00
		Energy-based	00.01	71.60	71.61
	-5	Proposed	02.04	46.66	**48.70**
		FFT based	00.08	67.37	67.45
		Energy-based	00.00	91.90	91.90
Pink	5	Proposed	07.46	18.61	**26.07**
		FFT based	01.18	34.09	35.27
		Energy-based	00.07	54.89	54.95
	0	Proposed	05.04	29.91	**34.95**
		FFT based	00.36	50.42	50.78
		Energy-based	00.00	75.13	75.13
	-5	Proposed	03.12	43.56	**46.68**
		FFT based	00.06	69.20	69.26
		Energy-based	00.00	93.78	93.78
Babble	5	Proposed	18.84	13.85	**32.69**
		FFT based	02.13	35.61	37.74
		Energy-based	00.19	52.11	52.30
	0	Proposed	15.52	20.75	**36.27**
		FFT based	01.28	49.54	50.82
		Energy-based	00.08	68.79	68.87
	-5	Proposed	12.89	28.67	**41.56**
		FFT based	00.78	62.84	63.62
		Energy-based	00.04	82.38	82.42

Table 6: Syllable boundary detection in the presence of noise for SVL-DD-NEWS-corpora

Noise	SNR (dB)	Method	Insertion (%)	Deletion (%)	TER (%)
White	5	Proposed	10.91	24.85	**35.76**
		FFT based	04.06	37.06	41.12
		Energy-based	01.38	55.66	57.04
	0	Proposed	07.34	35.48	**42.82**
		FFT based	02.02	51.58	53.60
		Energy-based	00.45	73.97	74.42
	-5	Proposed	03.79	50.64	**54.43**
		FFT based	00.71	69.99	70.70
		Energy-based	00.04	92.80	92.84
Pink	5	Proposed	10.99	25.47	**36.46**
		FFT based	03.63	39.33	42.99
		Energy-based	01.18	58.43	59.61
	0	Proposed	07.72	35.35	**43.07**
		FFT based	01.87	54.27	56.14
		Energy-based	00.39	77.31	77.70
	-5	Proposed	04.93	47.44	**52.37**
		FFT based	00.65	71.68	72.33
		Energy-based	00.02	94.49	94.51
Babble	5	Proposed	21.19	17.96	**39.15**
		FFT based	04.60	40.23	44.83
		Energy-based	01.63	55.85	57.48
	0	Proposed	17.48	23.92	**41.40**
		FFT based	02.65	52.72	55.37
		Energy-based	00.69	71.24	71.93
	-5	Proposed	14.39	30.84	**45.23**
		FFT based	01.37	64.72	66.09
		Energy-based	00.18	83.77	83.95

Table 7: Syllable nucleus detection for TIMIT Test database

Method	Insertion(%)	Deletion(%)	TER(%)
Proposed	04.74	17.07	21.81
FFT based	01.59	24.19	25.78
Energy-based	00.72	25.05	25.77
Syll-o-matic [2]	10.10	10.50	**20.60**

Table 8: Syllable boundary detection for TIMIT Test database

Method	Insertion(%)	Deletion(%)	TER(%)
Proposed	21.16	22.03	43.19
FFT based	15.17	26.09	41.26
Energy-based	11.53	29.17	40.70
Syll-o-matic [2]	11.20	12.80	**23.90**

strengths are extracted using the ZTW method with 20 ms window and one sample shift. Median filter with the size of the average pitch period obtained from Zero frequency filtering (ZFF) of the original speech signal [19][20] is applied on the DRS contour to get the sonority profile. The peaks in the sonority profile are considered as the syllable nuclei and the valleys as the syllable boundaries.

From Figure 1 [1], in the band 500 Hz to 1700 Hz, there are less number of peaks in the syllable compared to the other band. which signifies that this band is more sonorant in nature than the other band. Hence this band is used in the proposed method to highlight sonorant characteristics.

A peak can be a nucleus if it is voiced and follows the sonority sequencing principle. As the syllable nuclei are voiced, voicing decision obtained from the ZFF method [21] is applied over the sonority profile to ignore false peaks. According to sonority sequencing principle, the sonority strength has to be maximum at the nucleus and minimum towards the boundaries. Hence if 70% ($\approx$ RMS) drop of DRS not found on both the sides of a peak, then that peak is ignored. As there is only one nucleus in the syllable, so if the drop of 70% between two peaks is not found, then the peak with lesser DRS is ignored. The remaining peaks are considered as the syllable nuclei. The valley points in between the syllable nuclei are considered as the syllable boundaries.

In the case of the Fourier based extraction, instead of using the ZTW method to get the DRS, a 512 point discrete Fourier transform with Hamming window of size 20 ms and a sample shift is used, while in the case of energy-based extraction, energy is computed for a window of size 20 ms with a sample shift. The remaining procedure for detecting the syllable nuclei and its boundary is same.

4. Results

The proposed method is compared with the Fourier based and the energy-based sonority extraction methods.

The detected syllable nucleus is said to be correct if one nucleus is detected in the syllable. If more than one is detected it is classified as an insertion error and if no nucleus is detected, then it is classified as a deletion error. For the syllable boundary, the same strategy is followed, except that a margin of 50 ms on either side of the ground truth is also considered. For an easy comparision with previous literature, this evaluation metric is chosen.

Figure 2 shows the syllable nuclei and the syllable boundaries detected using different methods. It is observed that the proposed method is useful in detecting the syllable nuclei and boundaries even in the presence of noise. The heavily decaying window used in the ZTW method gives more dynamic range compared to that of the Hamming window used in the Fourier analysis. In the energy-based extraction, there is an averaging phenomenon, which occurs over the size of the window and hence the dynamic range decreases.

From Tables 1 and 2, it is evident that the insertion rate for the proposed method is high compared to the other methods. Reducing the insertions by utilizing some other features like the DRF has to be exploited. The band 500 Hz to 1700 Hz appeared to be more useful for characterizing the sonority compared to 300 Hz to 1000 Hz band as seen from Tables 1, 2, 3 and 4. Three kinds of noises, namely, white, pink, and babble are added at three different SNR levels of 5 dB, 0 dB and -5 dB from NOISEX database [22]. The proposed method performed better than the other methods even in the presence of noise as seen in Tables 5 and 6. It is also observed that as SNR level decreases, the increase in the total error rate (TER) of the proposed method is lesser than the other methods. For the purpose of comparison with [2], the proposed method is also evaluated on TIMIT test database. From Tables 7 and 8 the TER is almost equal for both the methods (i.e., proposed and method from [2]) for the syllable nuclei detection. But the TER for detecting the syllable boundary is high for this method, because the ground truth is from Tsylb software, which is C^*V based segmentation, whereas this method is for C^*VC^* type of segmentation. Syllable nuclei are detected more accurately than the syllable boundaries because the valley in the sonority contour sometimes occurred after 50 ms margin. Some of the insertions and deletions occurred due to improper voicing decisions. The sonority level of some sounds was comparable to that of the nucleus, which led to some insertions. Some deletions occurred due to the consideration of 70% drop in the DRS.

5. Conclusions

Sonority profile extraction using the proposed method to detect syllable nucleus and its boundary was found to be better than other methods because of its performance even in noisy conditions. Its superior performance is attributed to the high resolution both in time and frequency domains, large dynamic ranges and heavily decaying window provided by the zero time windowing method. The performance on TIMIT database in detecting the syllable nuclei is on par with the latest in the literature, however the syllable boundary detection fails because this method is for C^*VC^* type of segmentation where as the ground truth from Tsylb is of C^*V type segmentation. The Sonority profile using the proposed method also maintained the sonority hierarchy, which may help in the classification of sounds.

6. Acknowledgements

Authors sincerely thank Dr. Nicolas Obin from IRCAM, Paris, France for providing the syllable markings for TIMIT test database which is Tsylb generated and manually corrected.

[1] The dynamic range of the sonority contour in the Figures 1 and 2 has been reduced by applying quadroot over the DRS for illustration purpose.

7. References

[1] T. Ohno and T. Akiba, "Incorporating syllable duration into line-detection-based spoken term detection," in *IEEE Spoken Language Technology Workshop (SLT), Miami, Florida, USA*, Dec 02-05, 2012, pp. 204–209.

[2] N. Obin, F. Lamare, and A. Roebel, "Syll-o-matic: An adaptive time-frequency representation for the automatic segmentation of speech into syllables," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Vancouver, BC, Canada, May 26-31*, 2013, pp. 6699–6703.

[3] Y. Zhang and J. Glass, "Speech rhythm guided syllable nuclei detection," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Taipei, Taiwan, April 19 - 24*, 2009, pp. 3797–3800.

[4] A. R. Arrabothu, N. Chennupati, and B. Yegnanarayana, "Syllable nuclei detection using perceptually significant features." in *INTERSPEECH 2013, Lyon, France, August 25-29*, 2013, pp. 963–967.

[5] R. Ng and K. Hirose, "Syllable: A self-contained unit to model pronunciation variation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Kyoto, Japan, March 25 - 30*, 2012, pp. 4457–4460.

[6] P. Mermelstein, "Automatic segmentation of speech into syllabic units," *The Journal of the Acoustical Society of America*, vol. 58, no. 4, pp. 880–883, 1975.

[7] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "TIMIT acoustic phonetic continuous speech corpus LDC93S1," 1993.

[8] W.Fisher, "Tsylb syllabification package," 1996. [Online]. Available: http://www.itl.nist.gov/iad/mig/tools/tsylb2-11tarZ.htm

[9] D. Abercrombie, *Elements of General Phonetics*. Edinburgh University Press, 1967.

[10] J. Catford, *A Practical Introduction to Phonetics*. Clarendon Press, 1988.

[11] P. Ladefoged and K. Johnson, *A Course in Phonetics*. Cengage Learning, 2014.

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

[13] G. Clements, "Does sonority have a phonetic basis? Comments on the chapter by Vaux." in *Contemporary Views on Architecture and Representations in Phonological Theory*, E. Raimy and C. Cairns, Eds. MIT Press, 2009, pp. 165–175.

[14] Y. Nakajima, K. Ueda, S. Fujimaru, and Y. Ohsaka, "Sonority in British English," *In Proceedings of Meetings on Acoustics*, vol. 19, no. 1, pp. 1–5, 2013.

[15] N. Dhananjaya, "Signal proceesing for excitation-based analysis of acoustic events in speech," Ph.D. dissertation, IIT Madras, Chennai, India, October 2011.

[16] A.Chopde, "ITRANS - Indian language transliteration package, version 5.2." 2000. [Online]. Available: http://www.aczoom.com/itrans/

[17] R. Prasad and B. Yegnanarayana, "Acoustic segmentation of speech using zero time liftering (ZTL)," in *INTERSPEECH 2013, Lyon, France, August 25-29*, 2013, pp. 2292–2296.

[18] Y. Bayya 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," *In IEEE Transactions on Audio, Speech, and Language Processing, Las Vegas, USA, March 30 - April 4*, vol. 16, no. 8, pp. 1602–1613, 2008.

[20] B. Yegnanarayana, S. Prasanna, and S. Guruprasad, "Study of robustness of zero frequency resonator method for extraction of fundamental frequency," in *In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Prague, Czech Republic, May 22- 27*, 2011, pp. 5392–5395.

[21] N. Dhananjaya and B. Yegnanarayana, "Voiced/nonvoiced detection based on robustness of voiced epochs," *IEEE Signal Processing Letters*, vol. 17, no. 3, pp. 273–276, 2010.

[22] A. Varga and H. J. Steeneken, "Assessment for automatic speech recognition: Ii. noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, vol. 12, no. 3, pp. 247 – 251, 1993.