Interspeech 2018
2-6 September 2018, Hyderabad

Identification and classification of fricatives in speech using zero time windowing method

RaviShankar Prasad and B. Yegnanarayana

Speech Processing Laboratory,
International Institute of Information Technology, Hyderabad, India
ravishankar.prasad@research.iiit.ac.in, yegna@iiit.ac.in

Abstract

Fricatives are produced by creating a turbulence in the air–flow by passing it through a stricture in the vocal tract cavity. Fricatives are characterized by their noise–like behavior, which makes it difficult to analyze. Difference in the place of articulation leads to different classes of fricatives. Identification of fricative segment boundaries in speech helps in improving the performance of several applications. The present study attempts towards the identification and classification of fricative segments in continuous speech, based on the statistical behavior of instantaneous spectral characteristics. The proposed method uses parameters such as the dominant resonance frequencies, the center of gravity along with the statistical moments of the spectrum obtained using the zero time windowing (ZTW) method. The ZTW spectra exhibits a high temporal resolution and therefore gives accurate segment boundaries in speech. The proposed algorithm is tested on the TIMIT dataset for English language. A high identification rate of 97.5% is achieved for segment boundaries of the sibilant fricative class. Voiced nonsibilants show a lower identification rate than their voiceless counterparts due to their vowel–like spectral characteristics. A high classification rate of 93.2% is achieved between sibilants and nonsibilants.

Index Terms: fricatives, sibilant, nonsibilant, zero time windowing, dominant resonances, skewness, kurtosis.

1. Introduction

Fricatives comprise the largest set of consonants in English language, given the different manners of articulation. Fricative sounds in speech are produced by forcing the air through a construction in the oral cavity. A turbulence is generated in the air–flow at this constriction, which serves as the source for fricative sounds. The cavity following the constriction or the place of articulation dictates the system response. Different classes of fricatives are dependent of the pair of articulators forming the constriction. The English language has fricatives produced with constrictions mainly at four places, namely, labiodental (voiced /v/, voiceless /f/), linguadental (voiced /dh/, voiceless /th/), alveolar (voiced /z/, voiceless /s/) and palatal (voiced /zh/, voiceless /sh/) [1–3]. The present study spans around these classes of fricatives. The glottal fricative sound (voiceless /h/) exhibits spectral characteristics closer to the abutting vowels and therefore is not included in the present study. The alveolar and palatal fricative classes are called sibilants, and the other two classes are called nonsibilants. Sibilants are identified with relatively high intensity and a defined spectral structure. The distinction in the acoustic properties of these two classes arise due to the difference in cavity structures. Fricatives are also distinguished as voiced or voiceless depending on the presence of phonation (vocal fold vibration) during their production.

Attempts have been made to derive acoustic characteristics of fricatives for the identification of segment boundaries, and classification based on the place of articulation [3–9]. Fricative identification in speech is an important module for several speech applications, such as speech recognition, voice activity detection, speaker identification, audio search etc. Spectral properties of fricatives are dictated by the dimensions of the cavity beyond the place of articulation. A longer the cavity length leads to a well–defined spectrum. It is therefore that alveolars and palato–alveolars have distinct spectral shape, whereas dentals and labiodentals exhibit a relatively flat spectrum. The peak frequency location for palato–alveolar /sh/ and /zh/ occur mostly in mid–frequency band (around 2.5–4.5 kHz). Alveolars /s/ and /z/ are produced with a shorter anterior cavity, and therefore exhibit spectral prominence in the 4.5–6 kHz range. Labiodentals /f, v/ and linguadentals /th, dh/ are characterized by a relatively flat spectrum with no particular range of spectral prominence [4].

Studies have been attempted to delineate the acoustic correlates of fricatives in continuous speech and isolated utterances in English and other languages [3, 4, 10–12]. Parameters such as the durations of frication noise, spectral moments, spectral slope, amplitude energy, and zero crossing rate, have been used to classify fricatives based on their place of articulation [13–15]. Energy in different spectral bands, formant transition behavior around consonant–vowel (CV) boundaries, and other parameters have also been popular cues for the task. Features derived from methods based on filter–bank analysis, critical–band filtering, short–term adaptation, forward masking and synchrony detection have also been proposed [3, 5–7]. Another study uses the zero–crossing rate of the speech to detect unvoiced fricatives in continuous speech [17]. The linear prediction (LP) spectral peak locations have also been used as parameters to classify fricatives [18] along with the features based on presence of voicing, intensity of frication, and spectral shape. Generic automatic speech recognition (ASR) modules are also used to determine the onset and offset of fricatives in continuous speech [19, 20]. A detailed study of the acoustic characteristics for fricatives in English language has been presented in [4, 11]. A study based on the first four statistical moments of the fricative spectra employed to classify different fricative classes yielded a result of 80% [16]. The noise like structure in fricative regions makes it difficult to characterize the production behavior. The present study is motivated towards studying the production cavity behavior for different fricatives for the task of identification and classification.

The paper introduces spectral and statistical parameters derived from a high resolution spectra, obtained from the zero time windowing (ZTW) analysis of speech. The ZTW method estimates the spectral characteristics at each sample, using a small duration window. The spectral characteristics of the windowed

10.21437/Interspeech.2018-1958

segment is derived using the numerator of group delay function. The high temporal resolution helps in resolving accurate boundaries for fricative segments, while the spectral moments help to highlight the distinction in spectral structure. The paper is organized as follows: Section 2 discusses the zero time windowing (ZTW) method. Section 3 discusses the identification of fricative segments in speech. Section 4 presents the classification for fricative classes, and discusses the results. Section 5 gives a summary of the paper.

2. ZTW method and dominant resonance based representation the system response

Zero time windowing (ZTW) method [23] uses a heavily decaying window, given by

$$w_1[n] = \begin{cases} 0, & n = 0, \\ 1/(4sin^2(\pi n/2N)), & n = 1, 2, \ldots, N-1, \end{cases} \tag{1}$$

where N is the length of the window (in samples) corresponding to a duration of l ms. The net windowed segment is given as $x[n] = s[n]w[n]$, where $s[n]$ is the speech signal. The overall window $w[n] = w_1^2[n]w_2[n]$ has a component $w_2[n]$, which helps to reduce the window truncation effect and is given by,

$$w_2[n] = 4cos^2(\pi n/2N); \quad n = 0, 1, \ldots, N-1. \tag{2}$$

The application of $w_1[n]$ is equivalent to integration in the spectral domain. The spectral features are obtained by differentiating the numerator of the group delay (NGD) function, which is given as,

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

where $X(\omega)$=$X_R(\omega)$+$jX_I(\omega)$ is the discrete–time Fourier transform (DTFT) of $x[n]$, and $X'(\omega)$=$X'_R(\omega)$+$jX'_I(\omega)$ is the DTFT of $nx[n]$. Spectral characteristics of $x[n]$ are represented using Hilbert envelope of the differenced NGD (HNGD) function, which has a good resolution around the formants [24, 27]. HNGD spectra is computed at each sample leading to a high temporal resolution.

3. Identification of fricative segment boundaries in speech

The high temporal resolution obtained with the ZTW method can help in accurate demarcation of acoustic segment boundaries in speech [26]. The dominant resonance frequency (DRF) value is derived as the location of the strongest resonance peak in the HNGD spectra. The DRF is concise and efficient representation to study the significant changes in the production cavity length during the production of speech [28]. This can be attributed to correspondence of DRFs with the length of the dominant cavity, if the entire length of the vocal tract is approximated using a single tube. The DRF contour is used to identify the fricative segment boundaries in continuous speech. The present study uses a window size of 10 ms to compute the HNGD spectrum of the analysis segment, and hence obtain the DRF contour.

Figure 1 shows the spectral information representation with short–time Fourier transform (STFT) method, and the DRF contour with ZTW method which highlight the acoustic segment boundaries for fricative segments. Figures 1(a1) and 1(a2) show spectrograms for utterance *'situation'*, obtained from female and male speaker in TIMIT, respectively. The spectrogram is computed using STFT method with a window duration of 10 ms, with a shift of 5 ms. The phonetic boundaries for fricatives are observed in the figure at segments with an increased high frequency energy concentration. However, the spectrogram representation based on STFT suffers from limited resolution, and hence it is difficult to demarcate the accurate phonetic boundaries. Figures 1(b1) and 1(b2) show the corresponding DRF contours obtained using the ZTW analysis with a window length l = 10 ms. The consistency of the DRF contour to transit to higher frequency range, and the contrast in their behavior for different fricative classes, can be noticed in both the figures. It can be observed that the vowel segments have DRFs in a relatively lower frequency range (upto 1.5 kHz). This can be attributed to the fact that voiced sounds are produced with relatively open vocal tract cavity which acts as the dominant resonance cavity, resulting in a lower value for DRF. Fricatives exhibit a relatively smaller cavity which results in DRFs in higher frequency range. A significant change in the length of the cavity for the production of fricatives is reflected in the abrupt transition at their onset and coda regions, which presents the ability of DRF contours to accurately identify phone boundaries.

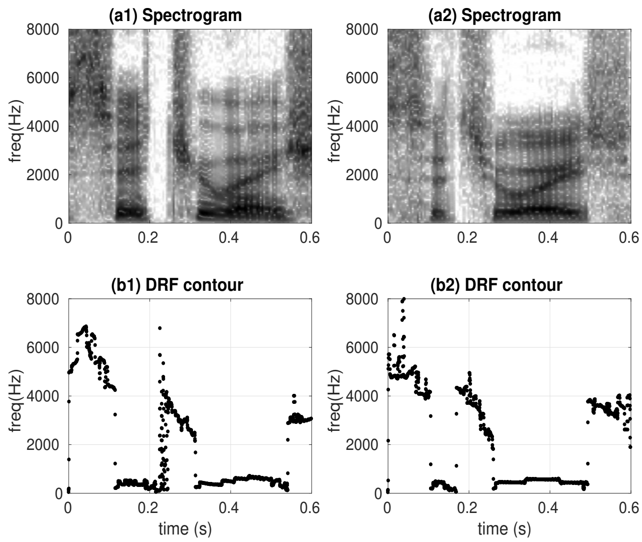

Figure 1: *(a1) and (a2) Spectrograms for utterance 'situation' by female and male speakers. (b1) and (b2) DRF contours obtained from the ZTW analysis.*

Based on the contrast in DRF contours for voiced sounds, fricative regions are identified as segments with DRFs in higher frequency range. The DRF contour is obtained using the ZTW analysis, using a window duration l = 10 ms. The segments with DRFs beyond the range of 1.5 kHz are demarcated as fricative segments. There is a likelihood of silence and stops to also be detected along with fricatives. A silence removal pre–processing based on the measure of single frequency filter spectrogram is implemented [30]. The resonance strength for the silence segments is generally very low compared to fricative segments, and therefore silence can also be removed by setting a threshold on the DRF strength values. Performance for frication region detection in speech segments is evaluated on a subset of the TIMIT database with 300 utterances, comprising of 68 speakers (33 male and 35 female)

Table 1: *Results obtained for fricative boundary (P_{phn}), and duration (P_{dur}), identification in continuous speech.*

Phones	# phones	P_{phn}	P_{dur} (%)
$/s/_{\text{sib}}$	246	97.5	90
$/sh/_{\text{sib}}$	93	100	94
$/z/_{\text{sib}}$	142	90.8	80
$/zh/_{\text{sib}}$	12	100	87
$/th/_{\text{non–sib}}$	33	30.3	77.5
$/f/_{\text{non–sib}}$	134	95	78.4
$/dh/_{\text{non–sib}}$	99	40	76.5
$/v/_{\text{non–sib}}$	68	12	68.4
Sib	493	97.5	89
Non–sib	179	45.3	75.2
Overall	334	72	82

[29]. Table 1 shows the results obtained using the proposed fricative boundary identification method based on DRF contour. The evaluation uses the manual boundary demarcations provided in the dataset, as a reference for performance assessment. The segments marked corresponding to the set of fricatives $\{/s/, /z/, /sh/, /zh/, /f/, /v/, /th/, /dh/\}$ are evaluated. The affricate segments are also detected using the given method due to a similarity in their production characteristics. But these segments have a relatively shorter duration and can be avoided using a duration based threshold. The table gives the results in terms of parameters P_{phn}, indicating the total number of corresponding phone onset and coda boundaries detected correctly (in %), and P_{dur} indicating the proportion of duration detected (in %). It is observed in the table that the segment boundary identification rate is higher for sibilants compared to nonsibilants. This is due to the absence of a proper cavity structure for nonsibilants, which lead to their flat noise–like behavior with a lower segment energy. The boundary identification for the sibilants are closed to 97%, with 89% of the duration of phones detected correctly. Nonsibilant segments with low SNR are eliminated as silence which leads to a lower rate of their identification. The nonsibilants /dh/ and /v/ are majorly voiced in nature and hence not detected as fricatives. The results show that the tracking of transition in DRFs from low to high frequency range is a good measure to identify fricative segments in speech.

4. Classification of fricatives based on spectral parameters

Difference in the cavity structure results in a contrast in the spectral structure for different fricative classes. Sibilants have a relatively well–defined cavity structure and therefore the corresponding DRFs appear within tight clusters in high frequency range. Nonsibilants lack a well–defined cavity beyond the place of excitation, and therefore the DRF locations in these segments are not bounded to a particular frequency range, but are spread across the spectral plane. This distinction in the behavior of DRFs can also be noted in Figs. 1(b1) and 1(b2) for different fricative classes. DRFs corresponding to the segment for phone $/s/$ (0–0.1 s region) are clustered beyond 4 kHz range, for both the speakers. Similarly, DRFs for the phone $/sh/$ ($\sim$ 0.5–0.6 s region) are clustered between 3–4 kHz range. DRFs are clustered in a relatively narrow frequency range for $/sh/$, compared to $/s/$ where the spread is larger. The presence of voicing does not essentially alter this behavior. However, voicing does introduce a prominent low frequency component which sometime appears dominant in the instantaneous spectra.

The present study uses DRFs along with other spectral parameters, to classify fricatives in broad classes based on their place of articulation. The following parameters are extracted from the HNGD spectra.

- Dominant resonance frequencies (DRFs): The first two dominant resonance frequencies (ρ_{D_1} and ρ_{D_2}) and their respective strengths (ρ_{S_1} and ρ_{S_2}) are given by,

$$\rho_{D_1} = \underset{\omega_i}{\operatorname{argmax}}(H(\omega_i)), \quad \rho_{S_1} = |H(\rho_{D_1})|, \tag{4}$$

  where ω_i is the frequency location of i^{th} peak in HGND spectrum $H(\omega)$. The peak locations are identified at zero–crossings of the differenced HNGD spectrum. The value of ρ_{D_2} and ρ_{S_2} are computed for the resonance peak with dominance behavior next to ρ_{D_1}.

- Spectral center of gravity (ρ_C): The center of gravity (COG) is computed to discriminate the concentration region of spectral energy. It is given by,

$$\rho_C = \frac{\sum_\omega \omega H(\omega)}{\sum_\omega H(\omega)}. \tag{5}$$

- Spectral skewness (ρ_K): Skewness is a measure of symmetry. It is the third central moment of the HNGD spectra and is given by,

$$\rho_K = \frac{(1/N)\sum_\omega (H(\omega) - \mu_\omega)^3}{\sigma_\omega^3}, \tag{6}$$

  where N is the total number of frequency bins, μ_ω is the first moment of the HGND spectrum, and σ_ω is the standard deviation.

- Spectral kurtosis (ρ_R): Kurtosis is a measure of the data being a heavy–tailed or light–tailed relative to a normal distribution. Data sets with high kurtosis exhibit a heavy tail, or outliers, and with a low kurtosis exhibit light tails. It is given by,

$$\rho_R = \frac{(1/N)\sum_\omega (H(\omega) - \mu_\omega)^4}{\sigma_\omega^4}. \tag{7}$$

- Ratio of spectral energies (ρ_E): The ratio of energies in frequency range [0 – 2 kHz] vs. [2 kHz – 4 kHz], [2 kHz – 4 kHz] vs. [4 – 6 kHz], and [4 kHz – 6 kHz] vs. [6 kHz – 8 kHz] is used to highlight the spectral density.

Every HNGD spectrum is understood as a probability distribution for a random variable, which helps to obtain the first four moments, namely the mean, variance, skewness, and kurtosis [4]. Skewness also refers to spectral tilt which dictates the slope of the spectrum. A spectrum with a higher energy density in the low frequency range exhibits a negative tilt in the spectrum. A spectrum with positive tilt and has more energy concentration towards the high frequency range. Positive kurtosis values indicate a spectrum with relatively high peakedness, while negative values indicate a relatively flatter distribution. Figures 2(a) and 2(b) show histograms for the parameter ρ_C obtained from segments of sibilants and nonsibilants, respectively, in TIMIT dataset. The nonsibilants exhibit a flatter spectrum and therefore the center of gravity tends to appear in the center of the spectral axis. The values of ρ_C appear centered

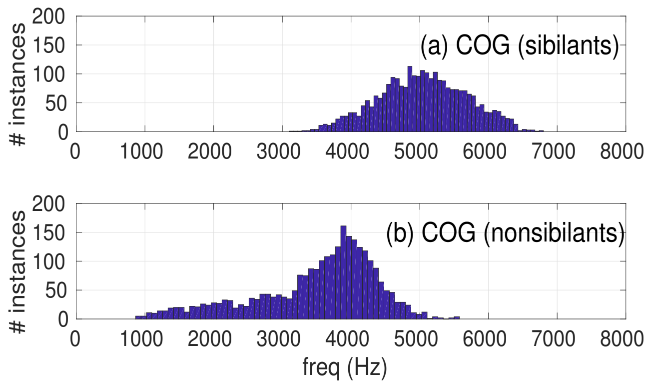

Figure 2: *Histograms for parameter ρ_C for (a) sibilants and (b) nonsibilants.*

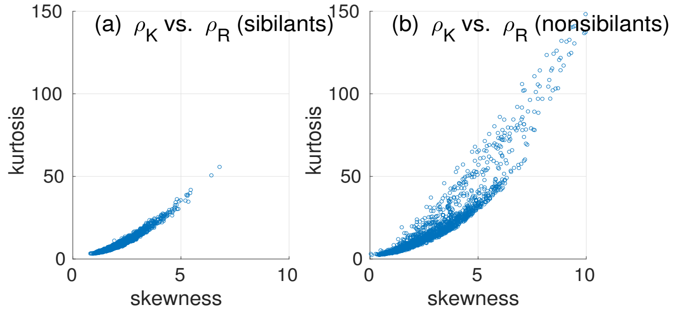

Figure 3: ρ_K *vs.* ρ_R *for (a) sibilants and (b) nonsibilants.*

around 5 kHz which suggests a shift in the spectral energy density to higher frequency range. Similar distinction in the behavior of higher spectral moments can be noticed from Figs. 3(a) and 3(b) which show the scatter plots of skewness and kurtosis, for sibilants and nonsibilants respectively. Both the classes exhibit similar skewness values bounded in a small positive range $[0-5]$. The values for kurtosis show relatively higher values for nonsibilants due to the higher number of spectral peaks.

The present study uses SVM classification method to discriminate between different fricative classes. The LibSVM package implementation is used [31]. The proposed set of parameters is obtained for 500 instances of each fricative class, uttered by different male and female speakers, in the TIMIT dataset. The speech segments corresponding to fricatives are obtained by the manual transcription provided in the dataset. A support vector machine (SVM) is trained using the parameter set. The classification rate (α), obtained on the test dataset of TIMIT using the proposed parameter set and MFCC is shown in Table 2.

Table 2: *Classification results between sibilant and nonsibilant fricatives.*

Methods	Proposed	MFCC
α (in %)	93.2	70

The results show that the task of fricative classification is performed with a good rate using the proposed parameter set. The classification between sibilants and nonsibilants is obtained at a high rate of 93%, compared to MFCCs, which result in a rate of 70%. The classification rate for different place of articulation for sibilants results in 73% accuracy as shown in Table 3. Further studies are planned to improve segment boundary identification and classification of nonsibilants.

Table 3: *Classification results between place of articulation (POA) for sibilant fricatives.*

Methods	palatal vs. alveolar	/s/ vs. /sh/	/z/ vs. /zh/
α (in %)	73	70	72

5. Summary and conclusions

A new method for the identification of fricative regions in continuous speech is proposed in this paper. This DRF based representation highlights the production characteristics with good resolution which gives accurate segment boundaries. Performance of the proposed features for the detection of fricatives is studied on the TIMIT database. A good performance is obtained for sibilant fricatives.

6. Acknowledgments

The first author would like to thank the Department of Electronics and Information Technology, Ministry of communication and IT, Govt. of India for granting PhD Fellowship under Visvesvaraya PhD Scheme.

7. References

[1] Kenneth N. Stevens(1999), “Acoustic Phonetics”,The MIT Press,Cambridge, Massachusetts, USA.

[2] Gunnar Fant(1960), “Acoustic Theory of Speech Production”,Mouton & Co. The Hague, Netherlands, pp. 1–328

[3] Ahmed. M. Abdelatty Ali and Jan Van der Spiegel,“Acoustic-phonetic features for the automatic classification of fricatives”, JASA, vol. 109, no. 5, pp. 2217–2235, May 2011.

[4] Allard Jongman and Ratree Wayland and Serena Wong(2000), “Acoustic characteristics of English fricatives”, JASA, vol. 108, no. 3, pp. 1252–1263.

[5] Stephanie Seneff(1985), “Pitch and spectral analysis of speech based on an auditory synchrony model”, Ph.D. Thesis, Dept. of Electrical Engineering and Computer Science, Massachusetts Institute of Technology.

[6] S. B. Davis and P. Mermelstein(1980), “Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences”, IEEE Trans. Acoustics Speech and Signal Processing,vol. 28,no. 4, pp. 357–366,

[7] M. J. Hunt and C. Lefebvre, “A comparison of several acoustic representations for speech recognition with degraded and undegraded speech”, ICASSP, Glasgow, Scotland, May 1989, pp. 262–265,

[8] Chorkin Chan and Ng, K. “Separation of fricatives from aspirated plosives by means of temporal spectral variation”, IEEE Trans. Acoustics Speech and Signal Processing,vol. 33,no. 5, pp. 1130–1137, Oct 1985, ISSN 0096–3518.

[9] Fujisaki, H. and Kunisaki, O., “Analysis, recognition, and perception of voiceless fricative consonants in Japanese”, IEEE Transactions on Acoustics, Speech and Signal Processing, vol. 26,no. 1, pp. 21–27, Feb 1978, ISSN 0096–3518.

[10] Alwan, A., Jiang, J., Chen, W. (2011). Perception of place of articulation for plosives and fricatives in noise Speech Communication, 53(2), 195–209. Elsevier B.V. doi:10.1016/j.specom.2010.09.001

[11] K. Maniwa, A. Jongman, and T. Wade, "Acoustic characteristics of clearly spoken English fricatives", Journal of Acoustic Society of America, vol. 125, no. 6, pp. 3962-3973.

[12] Ruinskiy, D; Dadush, N; Lavner, Y; , "Spectral and textural feature-based system for automatic detection of fricatives and affricates," Electrical and Electronics Engineers in Israel (IEEEI), 2010 IEEE 26th Convention of Israel, 17-20 Nov. 2010.

[13] You, H.Y., 1979.An Acoustical and Perceptual Study of English Fricatives. Master Thesis, University of Edmonton, Edmonton, Canada.

[14] Shadle, C.H., Mair, S.J., 1996. Quantifying spectral characteristics of fricatives. In: Proc. Internat. Conf. on Spoken Language Processing. Philadelphia, PA, pp. 1521-1524.

[15] Jongman, A., 1988. "Duration of frication noise required for identification of English fricatives." J. Acoust. Soc. Amer. vol. 85, no. 4, 17181725.

[16] Forrest, K., Weismer, G., Milenkovic, P. and Dougall, R.N., 1988. "Statistical analysis of wordinitial voiceless obstruents: preliminary data." The Journal of the Acoustical Society of America, vol. 84, no. 1, pp.115-123.

[17] D. A. MacKinnon and H. C. Lee,(April 1976). "Realtime recognition of unvoiced fricatives in continuous speech to aid the deaf", ICASSP, Philadelphia, USA, pp. 586–589.

[18] Lee Molho(April 1976), "Automatic acoustic-phonetic analysis of fricatives and plosives", ICASSP, Philadelphia, USA, pp. 182-185.

[19] Sharlene A. Liu, "Landmark detection for distinctive feature-based speech recognition", JASA, vol. 100, no. 5, Nov. 1996, pp. 3417–3430.

[20] Ariel Salomon and Carol Y. Espy-Wilson and Om Deshmukh, "Detection of speech landmarks: Use of temporal informoation", JASA, vol. 115, no. 3, March 2004, pp. 1296–1305.

[21] Alan V. Oppenheim and Ronald W. Schafer(1975), "Digital Signal Processing", Prentice Hall, Englewood Cliffs, New Jersey, USA, pp. 1-585,

[22] Sri Rama Murty K. and B. Yegnanarayana, "Epoch Extraction From Speech Signals", IEEE Trans. Audio, Speech and Language Processing, vol. 16, no. 8, pp. 1602–1613, Nov. 2008.

[23] B. Yegnanarayana, Dhananjaya N. Gowda, "Spectro–temporal analysis of speech signals using zero-time windowing and group delay function." Speech Communication, 55.6 (2013): 782–795.

[24] Anand Joseph M, Guruprasad S, and B. Yegnanarayana, "Extracting formants from short segments of speech using group delay functions", in Proc. Int. Conf. Spoken Language Processing (INTERSPEECH), 2006.

[25] A. Frid, "Acoustic-Phonetic Analysis of Fricatives for classification using SVM Based Algorithm", 2010 IEEE 26-th Convention of Electrical and Electronics Engineers in Israel, pp. 751–755, 2010.

[26] Prasad, RaviShankar, and B. Yegnanarayana. "Acoustic segmentation of speech using zero time liftering (ZTL)." In Proc. Int. Conf. Spoken Language Processing (INTERSPEECH), pp. 2292–2296. 2013.

[27] Sebastian, Jilt, Manoj Kumar, and Hema A. Murthy. "An analysis of the high resolution property of group delay function with applications to audio signal processing." Speech Communication 81 (2016): 42–53.

[28] R. S. Prasad and B. Yegnanarayana, "Determination of glottal open regions by exploiting changes in the vocal tract system characteristics, The Journal of the Acoustical Society of America, vol. 140, no. 1, pp. 666–677, 2016.

[29] John S. Garofolo, et al., (1993), "TIMIT Acoustic-Phonetic Continuous Speech Corpus", Linguistic Data Consortium, Philadelphia, USA,

[30] Aneeja, G., and B. Yegnanarayana. "Single frequency filtering approach for discriminating speech and nonspeech." IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP) 23.4 (2015): 705–717.

[31] Chang, Chih–Chung, and Chih–Jen Lin. "Training v–support vector regression: theory and algorithms." Neural computation 14.8 (2002): 1959-1977.