Interspeech 2018
2-6 September 2018, Hyderabad

Discriminating nasals and approximants in English language using zero time windowing

RaviShankar Prasad, Sudarsana Reddy Kadiri, Suryakanth V. Gangashetty and B. Yegnanarayana

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

Abstract

Nasals and approximants consonants are often confused with each other. Despite the distinction in the production mechanism, these two sound classes exhibit a similar low frequency behavior, and lack significant high frequency content. The present study uses a spectral representation obtained using the zero time windowing (ZTW) analysis of speech, for the task of distinction between these two. The instantaneous spectral representation has good resolution at resonances, which helps to highlight the difference in the acoustic vocal tract system response for these sounds. The ZTW spectra around the regions of glottal closure instants are averaged to derive parameters for their classification in continuous speech. A set of parameters based on the dominant resonances, center of gravity, band energy ratio, and cumulative spectral sum in low frequencies, is derived from the average spectrum. The paper proposes classification using a knowledge–based approach and training a support vector machine. These classifiers are tested on utterances from different English speakers in the TIMIT dataset. The proposed methods result in an average classification accuracy of 90% between the two classes in continuous speech.

Index Terms: Nasal consonants, approximant consonants, zero time windowing, dominant resonance frequency, numerator group delay, support vector machine

1. Introduction

Nasal (/m/ and /n/) and approximant (/l/, /r/, /j/ and /w/) consonants in English language belong to the family of sonorant consonants in phonetics. These consonants usually do not exhibit any frication noise. There have been attempts to study the acoustic characteristics of nasal segments and their identification in continuous speech. Approximants have not been studied much for their identification in continuous speech. In this paper, we derive the spectral features that help in distinguishing these two classes, from a production point of view.

Nasals are sonorant consonants, and are produced with a constriction in the oral cavity. Different nasal sounds are produced by closing the oral cavity at different articulatory positions. The point of constriction determines the corresponding spectral characteristics. The place of articulation for nasal consonants /m/ and /n/ in English language are bilabial and alveolar. Production of the nasal sounds also involves opening of the velopharyngeal port by lowering the velum. This results in a coupling of the nasal tract with the oral tract, resulting in a longer production cavity [1]. The coupled nasal tract introduces poles and zeros in the low frequency spectra, attributed to the presence of multiple sinus cavities and the closed oral tract [2, 3]. The temporal envelope of the nasal segments is generally characterized by a relatively lower energy content, and a smaller variance, in contrast to the adjacent vowel segment.

Identification of nasals in speech relies mostly on the behavior of the low frequency pole, zero, and the first formant (F1) [1]. Popular spectral cues to identify nasal segments in continuous speech are the presence of a low frequency pole (in 200–350 Hz) and a following zero (in 600–1100 Hz) in the spectrum, along with changes in the formant parameters such as location and bandwidth [4, 5, 6, 7, 8]. The location of the zero varies with the place of oral closure. A method based on the spectral parameters, such as, the energy in low frequency band (0–1 kHz), and the formant locations with corresponding bandwidths, resulted in a recognition of 80% for prevocalic and intervocalic cases and up to 60% for postvocalic cases [4]. Another method uses a set of parameters derived from different frequency bands in the spectrum, such as centroid in 0–500 Hz region, and average energy in 0–1, 1–2 and 2–5 kHz bands, to identify the onset of nasal segments [5]. These parameters resulted in a correct detection of 90% for nasal sounds present in different vocalic contexts. Another study uses locations of the peak amplitudes in multiple frequency bands (0–788 Hz, 788 Hz–2 kHz, 2–3 kHz, 3–4 kHz and 4–5 kHz), along with location of the lowest spectral peak, and average value of the difference among these parameters, to identify nasal segments in continuous speech [7]. The method resulted in a correct detection of 88% of nasal segments. Several other methods have focused on the study of nasalization in vowels [9, 10, 11, 12, 13]. Due to co–articulation, the nasal signature is mostly embedded as nasality in the adjacent sound. This phenomena leads to perception of nasal sounds in continuous speech.

Approximants are also sonorant consonants which are produced with a narrow constriction in the vocal tract. The rate of change of size of constriction is usually slower for these consonants leading to a slower formant transition from the adjacent vowels [14]. For example, the first and second formants for /w/ and /l/ merge together before/after transition from/to the vowel. Similarly /y/ exhibits a slower rise in amplitude for all the formants, whereas /r/ exhibits higher formants in low frequency range. The low frequency characteristics of approximants lie between those of vowel and nasal sounds, and hence the spectral changes during transition to these sounds is also subtle. There have not been many studies to characterize the acoustic properties of these sounds. A study using sonorant measures, consonant measures, and syllabic measures, is made to identify the acoustic behavior of different manners in approximants [14]. It uses energies in different frequency bands, and formant values, along with transitions within formants across the VC/CV conjunction regions.

There have been a few studies to distinguish nasal consonants among a set of sounds with similar acoustic behavior, such as semivowels, weak voiced fricatives, and voice bars [6, 8]. Parameters such as difference in the average energy between

10.21437/Interspeech.2018-1032

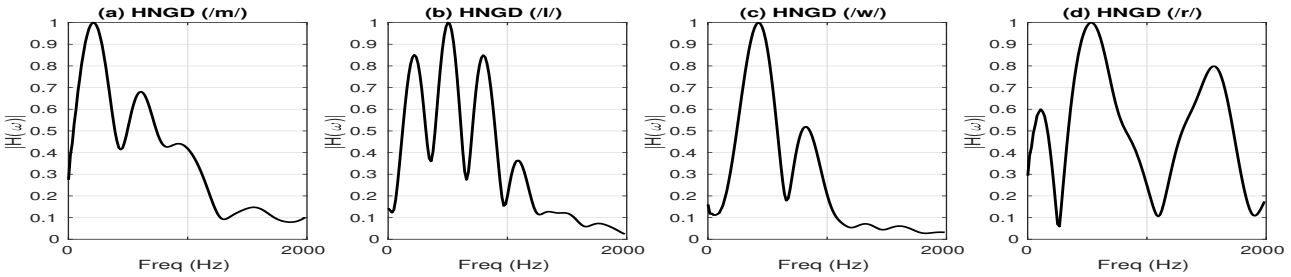

Figure 1: *Normalized HNGD spectrum (0–2 kHz) obtained with $l = 5$ ms for nasal (a) /m/, and approximants (b) /l/, (c) /w/ and (d) /r/.*

consonants and the adjacent vowels, average duration of resonance in 200–400 Hz, average strength of the resonance, and energy ratio between 0–350 Hz and 350–1000 Hz bands, are used to identify the nasal segments in speech resulting in a detection of 80%. Another study uses energy onset/offset measure, energy ratio based parameters, presence of a low spectrum peak, and variance of the amplitude envelope, to identify the nasal segments and distinguish them from approximant segments [8]. The spectral representation obtained using the numerator of group delay (NGD) has good resolution around the spectral peaks. The Hilbert envelope of NGD (HNGD) spectrum helps in resolving the low frequency nasal peaks, and therefore was used to identify the presence of nasal segments in voiced speech [19]. These studies have also highlighted the similarity between nasal and approximant segments, affecting the performance of nasal identification algorithms [8, 19]. The task of distinguishing between nasal and approximant consonants appears straight–forward owing to the presence of a spectral zero in the former case. However, the modeling and identification of a spectral zero is a difficult problem in signal processing.

The present study exploits differences in the production system response to discriminate nasal and approximant segments in continuous speech in English language. Acoustic characteristics of these sound classes are studied using segments of small duration around the high SNR regions in speech. A method is proposed to discriminate nasals and approximant segments in continuous speech. The proposed method is tested using utterances in the TIMIT dataset for different male and female speakers of English language [20]. The organization of this paper is as follows. Section 2 reviews the basic zero time windowing (ZTW) analysis of speech. Section 3 describes the parameters derived from the HNGD spectrum. Section 4 presents an algorithm to distinguish nasals and approximants, and discusses the results. Section 5 gives a summary of the paper.

2. ZTW based analysis of speech

The ZTW method uses a heavily decaying window for analysis [17]. The windowed segment is given as $x[n] = s[n]w[n]$, where $s[n]$ is the speech signal, and $w[n] = w_1^2[n]w_2[n]$ is the window function. The heavily decaying window function $w_1[n]$ is given by,

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

$w_2[n]$ is another window which helps to reduce the effect of truncation, and is given by,

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

N is length of the window (in samples) corresponding to a duration of l ms. Application of the window function $w_1[n]$ can be interpreted as an integration operation performed twice in the frequency domain [17]. The spectral characteristics of $x[n]$ are obtained by successive differentiation of the NGD function given by,

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

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 $nx[n]$. The Hilbert envelope of the differenced NGD (HNGD) function is computed to represent the spectral characteristics of $x[n]$. The HNGD exhibits good resolution of spectral peaks [17, 18]. The ZTW method uses a small window of duration less than a pitch period. The HNGD spectra can be interpreted as instantaneous spectral representation. The ability of the instantaneous spectra to discriminate between the acoustic vocal tract system response between glottal open and closed region was discussed in [21].

3. Analysis of nasals and approximants

Figure 1 shows the normalized HNGD spectra ($|H(\omega)|$) for nasal and approximant segments obtained using the ZTW method ($l = 5$ ms). The spectra (Figs. 1(a–d)) are obtained at glottal closure instant (GCI) locations in the nasal (/m/) and approximant (/l/, /w/, /r/) segments. The nasal segment exhibits significant low frequency spectral prominence as compared to approximants, where the spread results in prominent peaks appearing in the frequency range of 500–600 Hz, and beyond. The low frequency nasal resonance (300–450 Hz) and the higher frequency characteristic resonances for approximants are resolved well in the HNGD spectra. The differences in the spectral structure reflect the changes in the production characteristics.

3.1. Features for discriminating between nasal and approximant segments

The following features are derived from the HNGD spectrum to discriminate among nasals and approximants.

- 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,

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

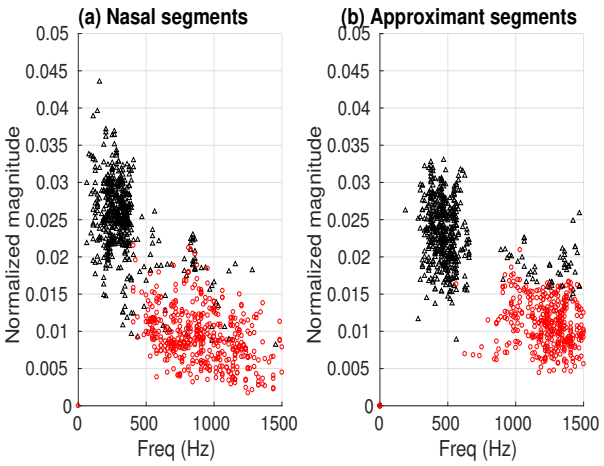

Figure 2: *ρ_{D_1} vs. ρ_{S_1} ($\triangle$ black) and ρ_{D_2} vs. ρ_{S_2} (o red) for (a) nasal and (b) approximant segments.*

where ω_i is the frequency location of i^{th} peak in the HGND spectrum $H(\omega)$. The peak locations are identified at the zero–crossings of the differenced HNGD spectrum. The ρ_{D_2} and ρ_{S_2} values correspond to the second dominant peak.

- Cumulative sum ($\rho_H(\omega)$): The cumulative sum of the normalized HNGD spectrum is used to parameterize the gradient of the low frequency spectral energy.
- Spectral center of gravity (ρ_C): The center of gravity for a low frequency (ω = 0–1200 Hz) range is computed to capture the concentration of spectral energy. It is given by

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

- Ratio of spectral energies (ρ_E): The ratio of energies in the frequency ranges 0–500 Hz and 500–1200 Hz is used to highlight the presence of the spectral null in case of nasals.

These parameters are obtained around GCI locations in each segment, for utterances of different male and female speakers of English in the TIMIT dataset [20]. The scatter plots of (ρ_{D_1},ρ_{S_1}) and (ρ_{D_2},ρ_{S_2}) for nasal segments are shown in Fig. 2(a), and for approximant segments in Fig. 2(b). The ρ_{D_1} values for nasals appear in a cluster in the frequency range of 200–350 Hz. For approximants, these appear in a cluster in the higher frequency range of 450–600 Hz. The ρ_{D_2} values for these classes also appear in different frequency ranges of 500–1000 and 1000–1500 Hz, respectively.

Figure 3 illustrates the cumulative sum contour normalized with respect to total energy content till 500 Hz. The high gradient in the low frequency range result in a convex contour for nasals, as can be seen in Figs. 3(a) and 3(b) for /m/ and /n/ segments, respectively. The contour appears relatively concave in Figs. 3(c) and 3(d), for /r/ and /l/ segments, respectively. This difference is highlighted by observing $\rho_H(\omega)$ at 300 Hz. The nasal spectral energy rises beyond half of its net energy at this frequency, whereas the approximant spectral energy stays below 0.6, as can be seen in Fig. 3(e). The parameters ρ_C and ρ_E also help in discriminating nasal and approximant segments. The ρ_C values usually lie in the 300–400 Hz range for nasal segments, and in the 450–550 Hz range for approximant segments.

4. Classification algorithms

The parameter set $S = \{\rho_{D_1}, \rho_{D_2}, \rho_{S_1}, \rho_{S_2}, \rho_H, \rho_C, \rho_E\}$ is derived at GCIs for segments of nasals (/m/ and /n/) and approximants (/l/, /r/, /j/ and /w/). The HNGD spectra are computed for l = 5 ms, for 50 samples locations around each GCI. An analysis window duration comparable to the average human pitch period helps in obtaining the spectral characteristic with an instantaneous nature. The GCIs are obtained using the ZFF method [16]. Two methods, namely the knowledge–based approach and classification with SVM, are used to study the effectiveness of the proposed parameter set in discriminating the two classes. The knowledge–based approach gives better insight into the changes in production mechanism of these utterances. But the problem with this approach is setting suitable thresholds. The SVM classifier avoids this problem, but requires good labeled data.

Table 1: *Parameters and the corresponding threshold values.*

ρ_{D_1}(Hz)	ρ_{D_2}(Hz)	$1-(\rho_{S_1}/\rho_{S_2})$
$\theta_1 = 400$	$\theta_2 = 600$	$\theta_3 = 0.4$
ρ_C(Hz)	$\rho_{H_{\omega_1}}$	ρ_E
$\theta_4 = 400$	$\theta_5 = 0.5$	$\theta_6 = 1$

The knowledge–based method utilizes the average behavior of the parameters to derive thresholds for the task. The parameters and their thresholds are given in Table 1. The thresholds are obtained by observing these parameter values across 300 instances of these classes.

A segment is classified as nasal if it satisfies majority of the following conditions.

$\rho_{D_1} \leq \theta_1$; $\rho_{D_2} \leq \theta_2$; $1-(\rho_{S_2}/\rho_{S_1}) > \theta_3$;
$\rho_C \leq \theta_4$; $\rho_{H_{\omega_1}} \geq \theta_5$; $\rho_E \geq \theta_6$

The parameters are tested for 500 different segments of nasal and approximant classes. The results obtained using the proposed algorithm are shown in Table 2. The unsupervised algorithm provides a good classification between the two classes. To obtain a more generalization to the classification algorithm, we use a supervised classification method. A support vector machine (SVM) is trained for the purpose of classification between the two classes using the parameter set S [22]. The design and complexity of an SVM classifier is determined by a subset of the data to be classified. Given a set of training vectors $\{\mathbf{x}_i\}_{i=1}^N$ and the corresponding class labels $\{y_i\}_{i=1}^N$, the SVM results in a classification hyperplane. Natural data usually leads in classification problems that can only be solved using a nonlinear decision surface. Kernel–based transformations allow a dot product to be computed in a higher dimensional space without explicitly mapping the data into these spaces [22]. The present classification method uses a radial basis function kernel, which is a form of a Gaussian kernel.

Table 2: *Discrimination results for nasals and approximants.*

Detection (%)	Nasals	Approximants
Nasals	95	5
Approximants	20	80

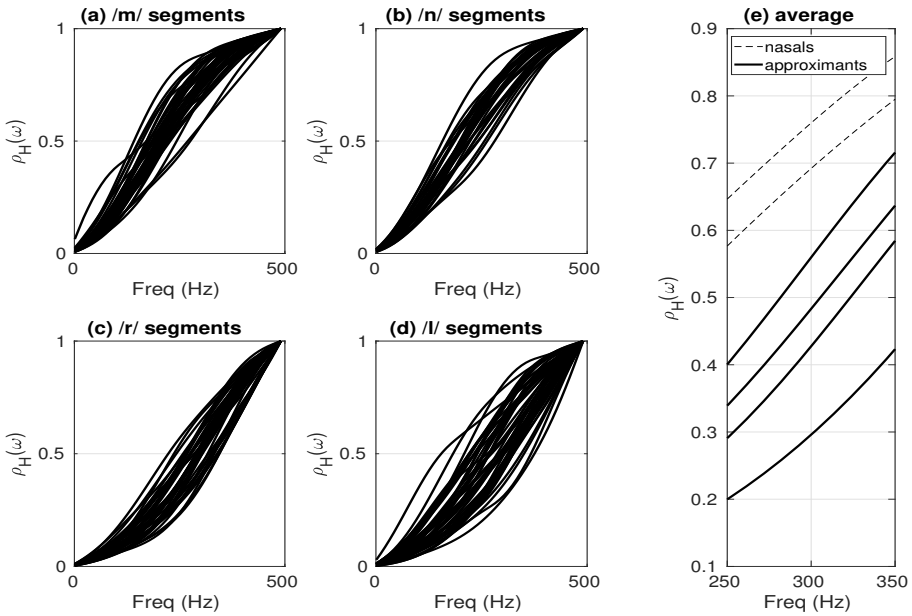

Figure 3: *Cumulative sum contours for the HNGD spectra in the frequency region* 0–500 *Hz for nasals (a) /m/ and (b) /n/ and approximants (c) /r/ and (d) /l/. (e) Average cumulative sum contours for nasals (- -) and approximants (–).*

The classification procedure trains a SVM across 1500 instances of nasal and approximant segments from TIMIT. The classifier is tested across the TIMIT dataset, which results in an average classification rate of 92% across the test data after 100 runs. A comparison of the classification results obtained using the proposed parameter set with other parameters is given in Table 3 The SVM is trained on the MFCC parameters obtained from the same instances of the two classes. Another parameter set, spectral band peak amplitudes, is derived based on the peak amplitudes in the frequency bands 0–788 Hz, 788 Hz–2 kHz, 2–3 kHz, 3–4 kHz and 4–5 kHz, which are used for nasal segment identification [7]. Table 3 gives the classification rate (α) obtained using these parameters. The table shows that the proposed parameters outperform other parameters. The misclassification of 8% of approximants as nasals and vice versa using the parameter set S can be attributed to the factors related to the duration of the segments, as well as a high amount of phonation noise for some speakers and segments. The proposed parameter set explores the distinction in the acoustic system response for the two classes and is effective in the task of discrimination. The results prove to be beneficial towards improving ASR systems, audio search engines and related applications. A further study is planned to explore the utility of these parameters towards discrimination among different nasal sounds, and approximant sounds.

Table 3: *Comparison of the proposed parameter set S, with other parameters to discriminate nasals vs. approximants.*

Methods	MFCC	SBPE	S
α (in %)	82.5	67.2	91.6

5. Summary and conclusion

The paper examined the discrimination of nasal and approximant segments in continuous speech, using parameters derived from the instantaneous HNGD spectra. The HNGD spectrum is derived using the ZTW method, which has the ability to resolve the spectral peaks from small analysis window. A smaller window helps to resolve the spectral characteristics for the glottal open and closed phases with minimum overlap. The spectra obtained around the high SNR GCI region is utilized to parameterize the spectra. A set of seven parameters is used to derive an unsupervised algorithm to distinguish between the two classes. The parameters are further used to train a SVM classifier. A RBF kernel results in a classification accuracy of 92% for speakers in the TIMIT dataset. The proposed method helps improving the recognition of nasals and approximants in continuous speech.

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. The second author would like to thank Tata Consultancy Services (TCS), India for supporting his PhD program.

7. References

[1] Stevens, Kenneth N. “Acoustic phonetics.” Vol. 30. MIT press, 2000.

[2] Fujimura, Osamu. “Analysis of nasal consonants.” The Journal of the Acoustical Society of America, 34.12 (1962), pp. 1865–1875.

[3] Fujimura, Osamu. “Formant–antiformant structure of nasal murmurs.” Proceedings of the speech communication, seminar. Vol. 1. (1962), pp. 1–9

[4] Weinstein, C., McCandless, S. S., Mondshein, L., and Zue, V., “A system for acoustic-phonetic analysis of continuous speech.” IEEE Transactions on Acoustics, Speech, and Signal Processing, 23.1 (1975), pp. 54–67.

[5] Mermelstein, Paul. “On detecting nasals in continuous speech.”

The Journal of the Acoustical Society of America, 61.2 (1977), pp. 581–587.

[6] Glass, J., and V. Zue. "Detection of nasalized vowels in American English." in Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing, (ICASSP'85). 10. (1985), pp. 1569–1572.

[7] Chen, Marilyn Y. "Nasal detection module for a knowledge-based speech recognition system." in Proc. of Int. Conf. on Spoken Language Processing, (Interspeech, 2000) (Beijing, China), (2000), pp. 636–639.

[8] Pruthi, Tarun, and Carol Y. Espy–Wilson. "Acoustic parameters for automatic detection of nasal manner." Speech Communication, 43.3 (2004), pp. 225–239.

[9] Chen, Marilyn Y. "Acoustic parameters of nasalized vowels in hearing-impaired and normal-hearing speakers." The Journal of the Acoustical Society of America, 98.5 (1995), pp. 2443–2453.

[10] Chen, Marilyn Y. "Acoustic correlates of English and French nasalized vowels." The Journal of the Acoustical Society of America, 102.4 (1997), pp. 2360–2370.

[11] Pruthi, T., and Carol Y. Espy–Wilson. "Acoustic parameters for the automatic detection of vowel nasalization." in Proc. of Int. Conf. on Spoken Language Processing, (Interspeech, 2000) (Beijing, China), (2000), pp. 1925–1928.

[12] Pruthi, Tarun. "Analysis, vocal–tract modeling, and automatic detection of vowel nasalization." Doctoral dissertation, University of Maryland. 2007.

[13] Vijayalakshmi, P., M. Ramasubba Reddy, and Douglas O'Shaughnessy. "Acoustic analysis and detection of hypernasality using a group delay function." IEEE transactions on biomedical engineering, 54.4 (2007), pp. 621–629.

[14] Carol Y. Espy-Wilson, "Acoustic measures for linguistic features distinguishing the semivowels /wjrl/ in American English." The Journal of the Acoustical Society of America, 92.2, (1992), pp. 736–757.

[15] Dhananjaya, N., and B. Yegnanarayana. "Voiced/nonvoiced detection based on robustness of voiced epochs." IEEE Signal Processing Letters, 17.3 (2010), pp. 273–276.

[16] K. S. R. Murty, B. Yegnanarayana, "Epoch extraction from speech signals." IEEE Transactions on Audio, Speech, and Language Processing, 16.8 (2008), pp. 1602–1613.

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

[18] Anand Joseph, M., S. Guruprasad, and B. Yegnanarayana, "Extracting formants from short segments of speech using group delay functions." in Proc. of Int. Conf. on Spoken Language Processing, (Interspeech, 06) (Pittsburgh, Pennsylvania), (2006), pp. 1009–1012.

[19] N. Dhananjaya. "Signal processing for excitation–based analysis of acoustic events in speech." PhD thesis, Dept. of Computer Science and Engineering, IIT Madras, Chennai, Oct. 2011.

[20] Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., and Pallett, D. S.. "DARPA TIMIT acoustic-phonetic continuous speech corpus." NASA STI/Recon Technical Report N, 1993.

[21] 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, 140.1, (2016), pp. 666–677.

[22] Vapnik, Vladimir., "The nature of statistical learning theory." , New York, Springer-Verlag, 1995.