INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Excitation Source and Vocal Tract System based Acoustic Features for Detection of Nasals in Continuous Speech

Bhanu Teja Nellore[1], *Sri Harsha Dumpala*[2], *Karan Nathwani*[3], *Suryakanth V Gangashetty*[1]

[1]Speech Processing Laboratory, International Institute of Information Technology, Hyderabad, India
[2]TCS Research and Innovation-Mumbai, India
[3] Indian Institute of Technology Jammu, India

bhanu.nellore@research.iiit.ac.in, d.harsha@tcs.com,
karan.nathwani@iitjammu.ac.in, svg@iiit.ac.in

Abstract

The aim of the current study is to propose acoustic features for detection of nasals in continuous speech. Acoustic features that represent certain characteristics of speech production are extracted. Features representing excitation source characteristcs are extracted using zero frequency filtering method. Features representing vocal tract system characteristics are extracted using zero time windowing method.

Feature sets are formed by combining certain subsets of the features mentioned above. These feature sets are evaluated for their representativeness of nasals in continuous speech in three different languages, namely, English, Hindi and Telugu. Results show that nasal detection is reliable and consistent across all the languages mentioned above.

Index Terms: nasal, voice bar, zero frequency filter, epochs, zero time windowing.

1. Introduction

During speech production, nasals refer to the sounds produced by lowering the velum in the vocal tract in order to allow passage of airflow through nasal cavity, while simultaneously obstructing the airflow to the oral cavity. The sounds /n/, /m/ and /ŋ/ are classified under nasals [1]. During the production of nasals, relatively significant energy is radiated through the nasal cavity compared to the oral cavity. Hence, analysis of nasals find application in the analysis of velopharyngeal dysfunctions leading to hyper-nasality. Nasal detection helps in speech segmentation for automatic speech recognition [2]. Stable spectral characteristics of nasals can also be used for speaker recognition [3].

Previous studies related to automatic detection of nasals include [4–11]. In this study, acoustic features are proposed for detection of nasals in continuous speech. Acoustic features used in this study exploit production characteristics of nasals related to both excitation source and vocal tract system (our previous studies that exploit production characteristics of other sounds such as sonorants, vowels and bursts include [12], [13] and [14]). All acoustic features are extracted around epochs as opposed to frame based analysis used in previous studies [9–11]. The proposed method also doesn't require prior training for nasal detection.

This paper is organized as follows. Section 2 explains the methods used for extraction of epochs and acoustic features for detection of nasals. Experimental details and results are mentioned in Section 3. Discussion on results is presented in Section 4. The final section gives the summary and conclusion of this study.

2. Features proposed for analysis of nasals

In this work, all features values are estimated around the glottal closure instants called epochs. Hence, accurate estimation of epoch location is vital in the analysis.

2.1. Extraction of excitation source based acoustic features

Epoch locations are extracted using zero frequency filtering method [15]. This method involves passing the differenced speech signal through a cascade of two zero frequency resonators (ZFR). This results in an output that grows/decays as a polynomial function of time. The trend in this output of ZFR is removed by local mean subtraction, using a window of length one to two pitch periods to highlight the small fluctuations of the output of ZFR. The resulting mean subtracted signal is called zero frequency filtered (ZFF) signal. The filtered signal clearly shows sharper zero crossings around the epoch locations. Hence the negative to positive zero crossing instants in ZFF signal are called epochs. A sample ZFF signal along with the marked epoch locations are shown in Figure 1(b). This method of epoch extraction was shown to be robust against different types of degradations, even at low signal-to-noise ratios [15]. The features of the glottal source of excitation derived from ZFF signal are as follows:

2.1.1. Strength of excitation (α)

Slope of ZFF signal around epochs gives a measure of the strength of impulse-like excitation (α). α corresponds to the rate of glottal closure [16]. Sharper the glottal closure, higher is the value α and vice-versa. α values are lower for nasals compared to vowels and approximants but nasals have higher α values compared to stops [17, 18]. This can also be seen in Figure 1(c). The lower values of α in nasals compared to vowels is due to complete closure formed in the oral cavity and narrow constricted path of nasal tract [18]. As also shown in Figure 1(c), α values are relatively lower in stops compared to nasals. This is due to the complete closure formed in both oral and nasal cavity for stops, whereas for nasals, complete closure occurs only in oral cavity.

2.1.2. Energy of excitation (β)

Energy of excitation (β) is computed as the energy of the ZFF signal within a window length of 3 msec, centered at every epoch location (1.5 msec on each side of epoch). Window length of 3 msec is considered around each epoch to capture the predominant excitation source information around the epoch locations. As shown in Figure 1(d), β values are lower for nasals

http://dx.doi.org/10.21437/Interspeech.2019-2785

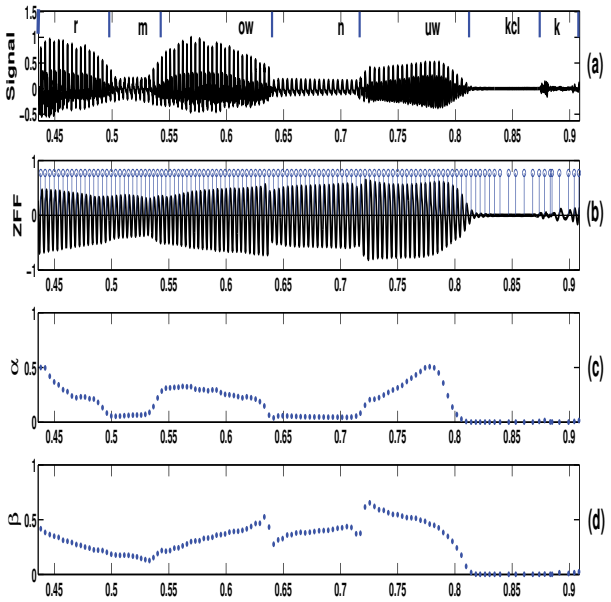

Figure 1: *Acoustic features from ZFF signal (a) Speech waveform for an utterance " ..ma/ nucl...". Manually marked phoneme labels are given above the signal. (b) ZFF signal along with hypothesized epochs at the positive zero crossings of the ZFF signal, (c) strength of excitation (α) values around epochs, and (d) energy of excitation values (β) around epochs. X-axes represent time in seconds.*

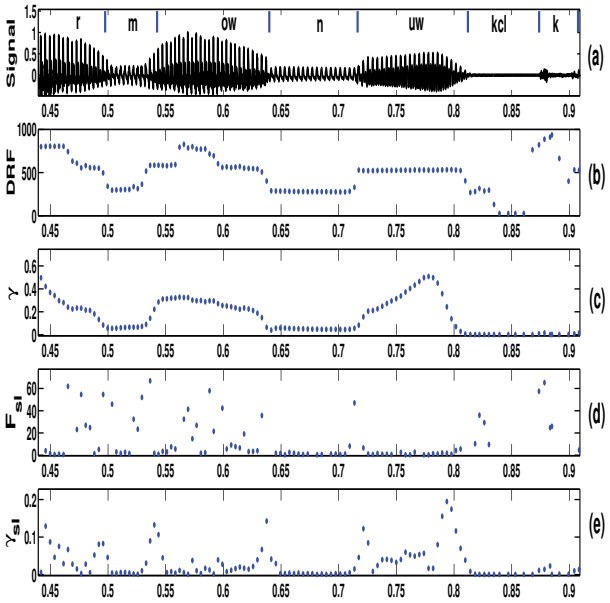

Figure 2: *Dominant resonance frequency based acoustic features. (a) Speech signal with manually marked phoneme labels, (b) dominant resonance frequency (DRF) values, (c) dominant resonance strength (γ) values, (d) slope of DRF (F_{sl}) values, and (e) Slope of dominant resonance strength (γ_{sl}) values at epochs. X-axes represent time in seconds.*

compared to vowels and approximants but β values are higher for nasals when compared to most voiceless sounds. This is due to narrow constricted path of nasal tract and complete closure formed in oral cavity [18].

2.2. Extraction of vocal tract system based acoustic features

Spectral characteristics of the vocal tract system are extracted using Zero-time windowing (ZTW) approach [19]. Brief description of the ZTW approach is as follows: Speech signal s[n] is multiplied with a highly decaying impulse-like window, which is defined as:

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

Where N is the window length.

$$x[n] = w[n] \times s[n], \qquad n = 1, 2, ...N \quad (2)$$

This windowing operation results in a impulse-like signal x[n], which has most of the energy at beginning of the window i.e., near zero time. Hence, the components of s[n] gets more emphasis at the beginning of the window, which is required to find the instantaneous spectral characteristics of the speech signal at any desired instant of time, such as at the epochs. Discrete Fourier transform (DFT) of the windowed signal x[n] results in an output that grows/decays polynomially. The hidden spectral features can be highlighted by exploiting the additive and high resolution properties of the group-delay function. In ZTW, numerator of the group-delay function (NGD) is used to avoid the problem of division by zero and also NGD provides higher resolution compared to group-delay function. NGD function is computed as:

$$g(\omega) = X_R(\omega)Y_R(\omega) + X_I(\omega)Y_I(\omega) \quad (3)$$

Where $X(\omega) = X_R(\omega) + jX_I(\omega)$ is the discrete-time Fourier transform (DTFT) of x[n] and $Y(\omega) = Y_R(\omega) + jY_I(\omega)$ is the DTFT of y[n] = nx[n].

The hidden spectral peaks are highlighted by successively differencing the NGD spectrum (DNGD). Hilbert envelope (HE) of DNGD (HNGD) spectrum is computed to account for the several integration and differencing operations performed and to highlight the peaks better [20]. In this work, the HNGD spectrum is obtained at every epoch location by considering a window W[n] of length 10 msec. Using ZTW, spectral information can be obtained with high spectral and temporal resolution at any instant of time, even for speech segments less than 5 msec. Spectral features derived from the HNGD spectrum, considered for analysis in this work are as follows:

2.2.1. Dominant resonance frequency (DRF)

Dominant resonance frequency (DRF) refers to the frequency of the dominant peaks in the obtained HNGD spectrum, as they represent the dominant resonances of the vocal tract system [21]. As shown in Figure 2(b), DRF values are lower for nasals compared to most of the other speech sounds. This is because of nasal tract coupling in nasals as opposed to oral tract coupling in other sounds.

2.2.2. Dominant resonance strength (γ)

Dominant resonance strength (γ) is measured as the magnitude of the HNGD spectrum at DRF. Large surface area and longer length of nasal cavity results in relatively lower γ values in nasal segments compared to vowels and approximants. But γ values of stop sounds are lower compared to that of nasals [17]. This order of γ values for different speech sounds is shown in Figure

2(c).

Spectral characteristics of nasals are relatively stable, atleast at low frequency regions of the spectrum as there are no moving parts in the nasal cavity [5, 10]. This spectral stability in nasals can be captured by computing slope of DRF and γ values as follows:

2.2.3. Slope of dominant resonance frequency (F_{sl})

Slope of dominant resonance frequency (F_{sl}) refers to the first order difference of the DRF values at epochs. Here absolute value of slope is considered. F_{sl} values for nasals are relatively low compared to other speech sounds as shown in Figure 2(d). Hence, F_{sl} values can be used as a feature to discriminate nasals from other segments of speech.

2.2.4. Slope of dominant resonance strength (γ_{sl})

Slope of dominant resonance strength (γ_{sl}) refers to the first order difference of the γ values at epochs. Here absolute value of slope is considered. As shown in Figure 2(e), γ_{sl} values are nearly zero in nasal segments and are also relatively lower for nasals compared to other sounds.

2.3. Acoustic features based on combination of excitation source and vocal tract system features

Most of the spurious decisions in detection of nasals, using excitation source and vocal tract system features arise because of the similarities in feature values between nasals and voice bars. Voice bars refer to the voiced region of the acoustic waveform corresponding to the phonation, when both oral and nasal cavities are completely closed [22]. The complete closure formed in oral cavity for both voice bars and nasals, results in the similarity in feature values. But the alternative path for airflow through nasal cavity in nasals results in acoustic characteristics that can be exploited to discriminate nasals from voice bars. To capture this variation, features based on combination of both excitation source and vocal tract system are used, which are explained as follows:

2.3.1. Product of strength of excitation source and dominant resonance strength (δ)

The product of strength of excitation source and dominant resonance strength (δ) is computed at every epoch location as given in (4).

$$\delta[i] = \alpha[i] \times \gamma[i], \qquad i = 1, 2, ...N \tag{4}$$

Where δ[i], α[i] and γ[i] refers to the values of δ, α and γ at ith epoch location and N refers to the total number of epochs.

Values of δ are relatively higher for nasals compared to voice bars as shown in Figure 3(b) and can be used to distinguish nasals from voice bars.

2.3.2. Product of energy of excitation and dominant resonance strength (η)

The product of energy of excitation and dominant resonance strength (η) is computed at every epoch location by multiplying the value of energy of excitation(β) at each epoch to the corresponding value of dominant resonance strength (γ) at that epoch as given in equation below.

$$\eta[i] = \beta[i] \times \gamma[i], \qquad i = 1, 2, ...N \tag{5}$$

Where η[i], β[i] and γ[i] refers to the values of η, β and γ at ith epoch location and N refers to the total number of epochs.

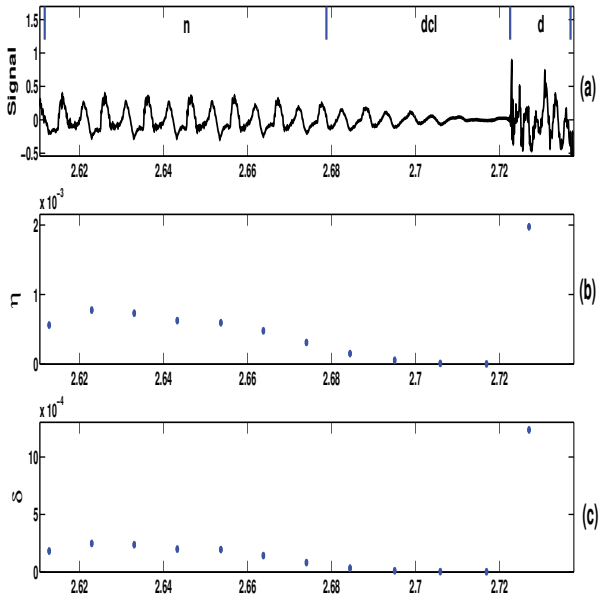

Figure 3: *Combination of excitation source and vocal tract system based features. (a) Speech signal, (b) η values, and (c) δ values, computed around epochs. X-axes represent time in seconds.*

Values of η are relatively higher for nasals compared to voice bars as shown in Figure 3(c) and can be used to distinguish nasals from voice bars.

Since, features δ and η are formed as a combination of excitation source and spectral features, they are referred to as combination features in this work.

3. Experimental details and results

In this section experimental details such as thresholds laid on acoustic features mentioned in Section 2, feature sets considered for experiments, database and evaluation procedure are explained.

3.1. Acoustic feature thresholds

Experiments are performed to study the effectiveness of different combinations of (i) excitation source features (α and β), (ii) vocal tract system features (DRF, γ, F_{sl} and γ_{sl}) and (iii) combined features (η and δ) for nasal detection. In this analysis, except DRF and F_{sl}, all other features values are normalized between 0 and 1.

For all the acoustic features discussed in Section 2, thresholds and bounds are laid emperically as shown in Table 1. In Table 1 Min. Val., and Max. Val., refer to minimum and maximum values of the corresponding features respectively. L bound and U bound refer to the lower bound and upper bound of the thresholds laid on those features respectively.

3.2. Feature sets

Based on the thresholds given in Table 1, particular epoch is decided as either nasal or not. The following experiments are performed to analyze the importance of different set of features for nasal detection. Five feature sets are formed by considering different combination of features discussed in Section 2. The feature sets are as follows:

Feature Set 1: In this set, only excitation source features (α

Table 1: *Empirically laid thresholds on the proposed features for nasal detection.*

Feature	Min. Val.	Max. Val.	L bound	U bound
α	0	1	0.08	0.80
β	0	1	0.002	0.18
DRF	0	4000	150	550
γ	0	1	0.0002	0.18
F_{sl}	0	4000	0	140
γ_{sl}	0	1	0	0.05
η	0	1	0.00035	0.075
δ	0	1	0.00012	0.13

and β) are considered.

Feature Set 2: In this set, only spectral features (DRF, γ, F_{sl} and γ_{sl}) are considered.

Feature Set 3: This set contains combination of both excitation source and spectral features i.e., α, β, DRF, γ, F_{sl} and γ_{sl}.

Feature Set 4: In this set, along with excitation source and spectral features, combined features (η and δ) are also used. Hence, this set contains the following features: α, β, DRF, γ, F_{sl}, γ_{sl}, η and δ.

Feature Set 5: In this set, combined features (η and δ), along with other features, which are independent of the combined features (DRF, F_{sl} and γ_{sl}) are considered for analysis.

3.3. Database for evaluation

The performance of proposed feature sets for nasal detection in continuous speech is evaluated on TIMIT database [23]. A subset of the TIMIT testing database, consisting of 150 randomly selected utterances, each of 3-4 seconds duration, spoken by 40 speakers (25 male and 15 female) is considered for analysis. Also, a dataset consisting of Indian languages typically, Hindi and Telugu, is considered to test the language dependency of the feature set. This dataset consists of 40 predefined Hindi utterances spoken by 1 male and 1 female native speakers and 35 predefined Telugu utterances spoken by 1 male and 2 female native speakers. All speakers are students of IIIT-Hyderabad. Each utterance is of 5 to 8 seconds duration and is recorded in a quiet environment at a sampling frequency of 16 kHz, using a standard headset microphone connected to a zoom handy recorder. Nasal and non-nasal (all sounds other than nasals), boundaries are manually marked for all utterances.

3.4. Evaluation procedure and results

Performance is measured in terms of number of epochs correctly detected in nasal regions (true positive rate) and number of spurious epochs hypothesized in non-nasal regions (false alarm rate). Epochs derived using ZFR [10] explained in Section 2 and the nasal decision obtained from manual labeling is used to generate the reference epochs in nasal and non-nasal regions. True positive rate is computed as: $P_{tpr} = N_{cn}/N_{ne} * 100\%$, where N_{cn} is number of epochs correctly detected as nasals of the total number of N_{ne} epochs in manually labeled nasal regions. False alarm rate is computed as: $P_f = N_f/N_{nn} * 100\%$, where N_f is number of non-nasal epochs detected as nasals, out of a total N_{nn} non-nasal epochs.

The performance of nasal detection using the feature sets on TIMIT dataset is shown in Table 2. The performance on Hindi and Telugu datasets using Feature Set 4 is shown in Table 3.

Table 2: *Performance of proposed features for detection of nasals in continuous speech on TIMIT dataset.*

Feature Set	P_{tpr} (%)	P_f (%)
Feature Set 1	97.88	38.02
Feature Set 2	93.21	23.98
Feature Set 3	91.86	19.77
Feature Set 4	91.04	15.92
Feature Set 5	91.49	16.58

Table 3: *Performance of Feature Set 4 on Indian Languages.*

Language	P_{tpr} (%)	P_f (%)
Hindi	90.54	16.88
Telugu	90.31	16.74

4. Discussion

The performance on TIMIT dataset for different feature sets defined for nasal detection in continuous speech is given in Table 2. Performance values in Table 2, show that, considering both excitation source and vocal tract system features instead of considering either excitation source or vocal tract system features, represents the nasal properties better. It can also be observed that eliminating the redundant features as in feature set 5, does not affect the accuracy much. Results also show that, the features are robust to inter-speaker and intra-speaker variations. The performance of the proposed features (Feature set 4) evaluated on Indian languages i.e., Telugu and Hindi, is given in Table 3. Performance measures for Telugu and Hindi are similar to the measures obtained for TIMIT database, showing the language invariance of the proposed features. Main source of error is in manual marking of the nasal boundaries. Most of the false alarms are caused because of nasalized vowels, especially closed vowels like /u/, which have properties similar to nasals. Some false alarms are caused by voice bars.

4.1. Comparison with results in literature

To the best of our knowledge, there is no unsupervised signal processing based method for detection of nasals in continuous speech in literature. Most of the studies [9–11] use machine learning techniques to build classification models for nasal detection. [11] reports an equal error rate (EER) of 4.1%. However, it must be noted that it uses 460 hours of training data from Librispeech corpus [24] for training connectionist temporal classification model.

The system proposed in [11] is relatively more accurate than the method proposed in this study for nasal detection. However, the method proposed in this study doesn't require prior training and was shown to produce reliable and consistent performance across different languages. Hence it can be used to detect nasals for under resourced languages.

5. Summary and conclusion

In this study, acoustic features based on both excitation source and vocal tract system characteristics are proposed for detection of nasals in continuous speech. All features are extracted around epochs, using ZFF and ZTW techniques. Results obtained show that, the proposed features are efficient for nasal detection and are robust across languages. Since the proposed method does not require prior training, it can be used for nasal detection in under resourced languages.

6. References

[1] K. N. Stevens, *Acoustic Phonetics*. Massachusetts: MIT press, 2000.

[2] D. Yu, S. M. Siniscalchi, L. Deng, and C. Lee, "Boosting attribute and phone estimation accuracies with deep neural networks for detection-based speech recognition," in *Proc. of International Conference on Acoustics, Speech and Signal Processing*, 2012, pp. 4169–4172.

[3] K. Amino, H. Makinae, and T. Kitamura, "Nasality in speech and its contribution to speaker individuality," in *Proc. of Interspeech*, Singapore, 2014, pp. 1688–1692.

[4] O. Fujimura, "Analysis of nasal consonants," *The Journal of the Acoustical Society of America*, vol. 34, no. 12, pp. 1865–1875, 1962.

[5] O. Fujimura, "Formant-antiformant structure of nasal murmurs," in *Proc. of the speech communication seminar*, vol. 1, 1962, pp. 1–9.

[6] A. S. House and K. N. Stevens, "Analog studies of the nasalization of vowels," *Journal of Speech and Hearing Disorders*, vol. 21, no. 2, pp. 218–232, 1956.

[7] C. Weinstein, S. S. McCandless, L. Mondshein, and V. Zue, "A system for acoustic-phonetic analysis of continuous speech," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 23, no. 1, pp. 54–67, February 1975.

[8] P. Mermelstein, "On detecting nasals in continuous speech," *The Journal of the Acoustical Society of America*, vol. 61, no. 2, pp. 581–587, 1977.

[9] M. Y. Chen, "Nasal detection module for a knowledge-based speech recognition system," in *Proc. of International Conference on Spoken Language Processing*, vol. 4, 2000, pp. 636–639.

[10] T. Pruthi and C. Y. Espy-Wilson, "Acoustic parameters for automatic detection of nasal manner," *Speech Communication*, vol. 43, no. 3, pp. 225 – 239, 2004.

[11] M. Cernak and S. Tong, "Nasal speech sounds detection using connectionist temporal classification," in *Proc. of International Conference on Acoustics, Speech and Signal Processing*, 2018, pp. 5574–5578.

[12] S. H. Dumpala, B. T. Nellore, R. Nevali, and B. Yegnanarayana, "Robust features for sonorant segmentation in continous speech," in *Proceedings of Interspeech*, Dresden, Germany, May 2015, pp. 1987–1991.

[13] S. H. Dumpala, B. T. Nellore, R. Nevali, S. V. Gangashetty, and B. Yegnanarayana, "Robust vowel landmark detection using epoch-based features," in *Proceedings of Interspeech*, San Francisco, USA, Sep. 2016, pp. 160–164.

[14] B. T. Nellore, R. Prasad, S. R. Kadiri, S. V. Gangashetty, and B. Yegnanarayana, "Locating burst onsets using sff envelope and phase information," in *Proc. Interspeech 2017*, Hyderabad, India, 2017, pp. 3023–3027.

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

[16] K. S. R. Murty, B. Yegnanarayana, and M. A. Joseph, "Characterization of glottal activity from speech signals," *IEEE signal processing letters*, vol. 16, no. 6, pp. 469–472, 2009.

[17] Dhananjaya N., "Signal processing for excitation-based analysis of acoustic events in speech," Ph.D. dissertation, Indian Intitute of Technology Madras, Chennai, 2011.

[18] V. K. Mittal, B. Yegnanarayana, and P. Bhaskararao, "Study of the effects of vocal tract constriction on glottal vibration," *The Journal of the Acoustical Society of America*, vol. 136, no. 4, pp. 1932–1941, 2014.

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

[20] M. A. Joseph, S. Guruprasad, and B. Yegnanarayana, "Extracting formants from short segments of speech using group delay functions," in *Proc. of Interspeech*, 2006, pp. 1009–1012.

[21] R. Prasad and B. Yegnanarayana, "Acoustic segmentation of speech using zero time liftering (ZTL)," in *Proc. of Interspeech*, 2013, pp. 2292–2296.

[22] N. Dhananjaya, S. Rajendran, and B. Yegnanarayana, "Features for automatic detection of voice bars in continuous speech." in *Proc. of Interspeech*, 2008, pp. 1321–1324.

[23] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "DARPA TIMIT acoustic-phonetic continous speech corpus," *NASA STI/Recon Technical Report N*, vol. 93, Feb. 1993.

[24] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2015, pp. 5206–5210.