



# Robust Features for Sonorant Segmentation in Continuous Speech

*Sri Harsha Dumpala, Bhanu Teja Nellore, Raghu Ram Nevali,  
Suryakanth V. Gangashetty and B. Yegnanarayana*

International Institute of Information Technology, Hyderabad, India

{sriharsha.dumpala, bhanu.nellore, raghuram.nevali}@research.iiit.ac.in,  
svg@iiit.ac.in, yegna@iiit.ac.in

## Abstract

Sonorant segmentation of speech signals is critical in developing Automatic Speech Recognition (ASR) systems, audio search systems and for automatic segmentation of speech corpora. In this work, acoustic features based on excitation source and vocal tract system characteristics of sonorant sounds are proposed for segmentation of sonorant regions in continuous speech. The features are based on zero frequency resonator signal energy, strength of excitation and dominant resonance frequency around epochs. An algorithm is developed to relate these features in hierarchical manner using knowledge-based approach. The performance of the proposed algorithm is studied on three different datasets, at varying levels of degradation. TIMIT database is used to test the validity and AMI meeting corpus and Telugu (an Indian language) dataset are considered to test the utility of the proposed features.

**Index Terms:** Sonorant, non-sonorant, zero frequency resonator, epochs, zero time windowing.

## 1. Introduction

Sonorant refers to the sound that is produced with no sufficiently strong constriction so as to produce turbulent noise or stoppage of airflow [1]. The broad classes namely vowels, nasals and approximants are categorized as sonorants whereas fricatives, stops and non-speech regions are considered as non-sonorants [2]. Sonorant is an essential component of the speech signal as it corresponds to the robust, high signal-to-noise ratio (SNR) regions of speech. Even if the speech is highly corrupted by noise, most sonorant regions can still be clearly perceived by humans. Also, most of the information can be retrieved from sonorant regions [2]. Hence, robust segmentation of sonorant regions leads to improved Automatic Speech Recognition (ASR). Applications of sonorant segmentation are not limited to ASR but extend to tasks like automatic segmentation and labeling of speech corpora for speech synthesis, audio search and analysis of paralinguistic elements of speech like laughter, speech-laugh [3, 4] etc.

Sonorant segmentation forms the root node in the hierarchical tree for broad class segmentation of speech, where a broad class refers to a group of speech sounds sharing similarity in their speech production mechanism [5-9]. It is essential to segment the speech into sonorant regions with high accuracy to build reliable systems based on broad class segmentation of speech. In literature, sonorant segmentation was performed by using features such as Mel Frequency Cepstral Coefficients (MFCCs), knowledge-based acoustic features or a combination of both [2, 7-11]. Support Vector Machines (SVMs) trained using MFCCs were used to detect the sonorant segments [7].

Also SVMs, trained separately using MFCCs obtained in various noise environments, were connected adaptively to perform robust segmentation of sonorants [2]. Acoustic features extracted from the speech signal, which explicitly represent the knowledge about speech, were also used [8-11]. Various acoustic features like energies and ratio of energies in different frequency bands and integration of cues in different narrow bands were considered for the purpose of sonorant segmentation [8-10]. Combination of MFCCs and acoustic features was also considered for this task [11].

In most of the previous studies, spectral features are well exploited. But excitation source features also carry important cues to discriminate sonorants from other speech sounds. The main objective of this study is to define an acoustic feature set based on combination of both excitation source and vocal tract system characteristics for robust segmentation of sonorants in continuous speech. In this study, both excitation source and spectral features are extracted around epochs but not at the frame level. Epochs refer to the instants of significant excitation of the vocal tract system. Regions around epochs have high SNR values, and are relatively more robust to external degradations than other regions, making proposed features more reliable in noisy environments [12,13].

The paper is organized as follows: Section 2 describes the datasets used for analysis. Section 3 explains the method for epoch extraction and the acoustic features proposed. Algorithm used for automatic segmentation of sonorant regions is described in Section 4. Results are discussed in Section 5. Summary and conclusions are provided in Section 6.

## 2. Database

The datasets used to evaluate the performance of the proposed algorithm are TIMIT database [14], AMI meeting corpus [15] and Telugu (an Indian language) database.

A subset of the standard TIMIT database is used for analysis. To set the thresholds on the acoustic features, 100 'si' and 'sx' utterances from TIMIT training set, spoken by 35 speakers (25 male and 10 female) are used. All the 1344 'si' and 'sx' utterances in the TIMIT test set, spoken by 168 speakers (112 male and 56 female) are used to evaluate the proposed algorithm.

To evaluate the performance of the proposed approach on spontaneous speech, AMI meeting corpus is used. This corpus consists of 100 hours of elicited meeting recordings, where each meeting involves 4 speakers discussing on the project of designing a new TV remote controller. Each meeting session is of 30 - 35 minutes duration. All meetings are in English, but a large portion of the speakers are non-native English speakers. Audio is collected using individual lapel microphones and

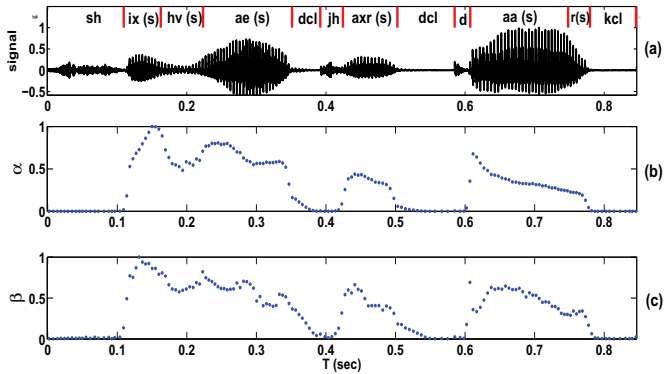


Figure 1: *Acoustic features from ZFR signal (a) Speech waveform for an utterance “ she had your dark”. Manually marked phoneme labels are given above the signal, where phones followed by (s) represents sonorant segments. (b) ZFR signal energy ( $\alpha$ ), and (c) strength of excitation ( $\beta$ ) values around epochs.*

headset condenser microphones and also a series of array microphones. In this analysis, we used the headset condenser microphone data collected from each speaker. 200 utterances spoken by 50 speakers (35 male and 15 female), each of 1 - 3 seconds duration are manually sliced from the meeting recordings for ease of analysis. For evaluation purpose, each utterance is manually labeled at phoneme level based on the annotations provided with the AMI corpus. All non-verbal segments like laughter, cough etc., are discarded, but events like speech-laugh where speech co-occurs with laughter are retained.

A dataset consisting of Telugu (an Indian language) utterances is considered to evaluate the language dependency of the approach. This dataset consists of 100 predefined Telugu utterances spoken by 3 male and 2 female native speakers. All speakers are students of IIIT-Hyderabad. Each utterance is of 4 - 6 seconds duration, and is recorded in a quiet environment at a sampling rate of 16 kHz, using a standard headset microphone connected to a zoom handy recorder. Sonorant and non-sonorant boundaries are manually marked for all utterances.

In order to study the robustness of the proposed approach, white noise from Noisex database [16] is added to the above mentioned databases at various SNRs.

### 3. Epoch extraction method and acoustic features for sonorant segmentation

Zero frequency resonator (ZFR) output for the speech signal is used for accurate estimation of the epochs [12]. Zero frequency filtering (ZFF) approach involves passing the speech signal twice through a digital resonator having poles at zero frequency. The trend in the output of the ZFR is removed by local mean subtraction using a window of length equal to 1 - 2 pitch periods, to highlight the small fluctuations. The negative to positive zero crossings in the trend removed output are observed to be synchronized with the instants of glottal closure, called epochs. This method of epoch extraction was shown to be robust against different types of degradations even at very low SNRs. All the proposed acoustic features are extracted around epoch locations, as explained below.

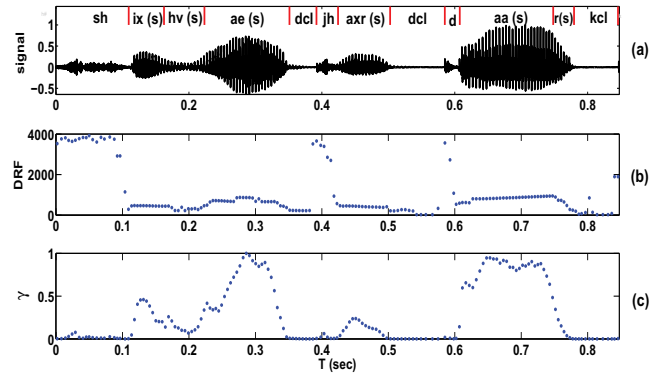


Figure 2: *Dominant resonance frequency based acoustic features. (a) Speech signal, (b) dominant resonance frequency (DRF) values, and (c) dominant resonance strength ( $\gamma$ ) values at epochs.*

#### 3.1. Zero frequency resonator signal energy ( $\alpha$ )

ZFR signal energy ( $\alpha$ ) is computed as the energy of the ZFR output within a window length of 2 msec, centered at every epoch location (1 msec on each side of the epoch). The absence of an increase in intra-oral pressure during the production of sonorants allows free vibration of the vocal folds [6]. This free vocal fold vibration produces higher energy concentration around epochs, resulting in higher  $\alpha$  values in sonorant regions compared to non-sonorants as can be seen in Fig. 1(b).

#### 3.2. Strength of excitation source ( $\beta$ )

The slope of the ZFR signal around epochs is proportional to the strength of excitation source ( $\beta$ ).  $\beta$  values correspond to the rate of glottal closure [17]. Sonorants have sharper glottal closure due to lack of buildup in intra-oral pressure, resulting in stronger excitation of the vocal tract system [6]. Hence,  $\beta$  values are higher in sonorant regions compared to non-sonorants as shown in Fig. 1(c).

#### 3.3. Dominant resonance frequency (DRF)

Spectral characteristics of the vocal tract system are extracted using zero time windowing (ZTW) approach, which involves multiplying the speech signal with a highly decaying impulse-like window to ensure high temporal resolution [18]. Spectral features derived using ZTW are further highlighted by exploiting the high resolution and additive properties of the Hilbert envelope of numerator group-delay (HNGD) function to generate the HNGD spectrum [19]. Using ZTW, spectral information can be obtained with high spectral and temporal resolutions at any instant of time, even for speech segments less than 5 msec duration. The frequency of dominant peaks in the HNGD spectrum represents the dominant resonance of the vocal tract system, thus is called as dominant resonance frequency (DRF) [19]. In this paper, the HNGD spectrum of the speech signal computed using a window length of 10 msec placed at epoch locations is used to extract the DRF values at epochs. The DRF values at epochs are lower and consistent for sonorants compared to non-sonorants as shown in Fig. 2(b). This is due to the lack of strong constriction in the production of sonorants [1]. Hence, the DRF values can be used as a reliable feature to discriminate sonorants from non-sonorants.

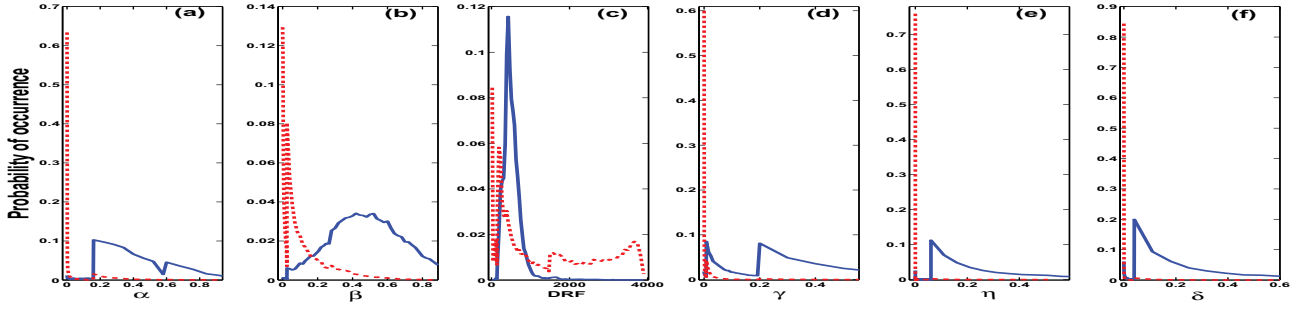


Figure 3: Normalized distributions of 100 TIMIT utterances for 25 male and 10 female speakers for (a)  $\alpha$ , (b)  $\beta$ , (c) DRF, (d)  $\gamma$ , (e)  $\eta$ , and (f)  $\delta$ . Solid lines represent sonorant and dotted lines represent non-sonorant distributions.

### 3.4. Dominant resonance strength ( $\gamma$ )

Dominant resonance strength ( $\gamma$ ) is measured as the magnitude of HNGD spectrum at the DRF. There is no strong constriction formed during the production of sonorants [1]. This results in higher energy concentration around DRFs in sonorants, which is evident from the higher values of  $\gamma$  in sonorant regions compared to non-sonorants as shown in Fig. 2(c).

### 3.5. Product of ZFR signal energy and dominant resonance strength ( $\eta$ )

The product of ZFR signal energy and dominant resonance strength ( $\eta$ ) is computed at every epoch location by multiplying the value of ZFR signal strength ( $\alpha$ ) at each epoch to the corresponding value of dominant resonance strength ( $\gamma$ ) at that epoch as given in (1).

$$\eta[i] = \alpha[i]\gamma[i], \quad i = 1, 2, \dots, N \quad (1)$$

where  $\eta[i]$ ,  $\alpha[i]$  and  $\gamma[i]$  refer to the values of  $\eta$ ,  $\alpha$  and  $\gamma$  at the  $i^{\text{th}}$  epoch location, and  $N$  refers to the total number of epochs.

### 3.6. Product of strength of excitation source and dominant resonance strength ( $\delta$ )

The product of strength of excitation source and dominant resonance strength ( $\delta$ ) is computed at every epoch location as given in (2).

$$\delta[i] = \beta[i]\gamma[i], \quad i = 1, 2, \dots, N \quad (2)$$

where  $\delta[i]$ ,  $\beta[i]$  and  $\gamma[i]$  refer to the values of  $\delta$ ,  $\beta$  and  $\gamma$  at the  $i^{\text{th}}$  epoch location, and  $N$  refers to the total number of epochs.

Distributions of features ( $\alpha$ ,  $\beta$ , DRF,  $\gamma$ ,  $\eta$  and  $\delta$ ) for sonorant and non-sonorant segments for 100 TIMIT train utterances, spoken by 25 male and 10 female speakers are shown in Fig. 3. The distribution of features is represented in terms of probability of occurrence, which is obtained by dividing the number of feature values for each class (i.e., sonorants and non-sonorants) with the total number of samples considered for that class. Distributions of feature values for sonorants are concentrated more in one region and that for non-sonorants are concentrated more in a different region. This shows the ability of these features to discriminate sonorants from other regions of speech.

Table 1 gives the mean and variance values of the acoustic features for the 100 TIMIT train utterances. These values show that the acoustic features can classify sonorants from non-sonorants with a large margin. It is also evident from Fig. 3 and Table 1 that  $\eta$  and  $\delta$  values provide better discrimination

between sonorant and non-sonorant classes compared to  $\gamma$  values. Hence  $\eta$  and  $\delta$  are considered for sonorant segmentation, instead of directly using  $\gamma$  values.

Table 1: Mean and variance values of the acoustic features in sonorant and non-sonorant regions.

	Sonorants		Non-sonorants	
	Mean	Variance	Mean	Variance
$\alpha$	0.3229	0.2346	0.0227	0.0715
$\beta$	0.4771	0.2213	0.0658	0.1122
$\gamma$	0.1501	0.2092	0.0065	0.0265
$\eta$	0.1402	0.1883	0.0013	0.0106
$\delta$	0.1429	0.1902	0.0017	0.0115

## 4. Algorithm for automatic sonorant segmentation

A hierarchy based algorithm is employed for automatic sonorant segmentation in continuous speech. The thresholds on the features are decided based on the distribution plots shown in Fig. 3. Except DRF, all other feature values are normalized between 0 and 1. The algorithm for sonorant segmentation is explained as follows:

*Step 1:* The presence of significant glottal activity in sonorants compared to non-sonorant is captured by ZFR signal energy ( $\alpha$ ) and strength of excitation source ( $\beta$ ) features. The binary decision based on  $\alpha$  and  $\beta$  is computed as follows:

$$d_1[i] = \begin{cases} 1, & \text{if } \alpha[i] > 0.02 \quad \& \quad \beta[i] > 0.05 \\ 0, & \text{otherwise,} \end{cases} \quad (3)$$

where  $\alpha[i]$ ,  $\beta[i]$  and  $d_1[i]$  refer to the values of  $\alpha$ ,  $\beta$  and the first sonorant evidence at the  $i^{\text{th}}$  epoch location, respectively.

*Step 2:* The second level of evidence for discriminating sonorants from other regions in continuous speech is computed using DRF feature combined with the evidence  $d_1$  obtained in step 1 as follows:

$$d_2[i] = \begin{cases} 1, & \text{if } d_1[i] = 1 \quad \& \quad 150 < DRF[i] < 1300 \\ 0, & \text{otherwise,} \end{cases} \quad (4)$$

where  $DRF[i]$ ,  $d_1[i]$  and  $d_2[i]$  refer to the values of DRF, the first level evidence and the second level evidence at the  $i^{\text{th}}$  epoch location, respectively.

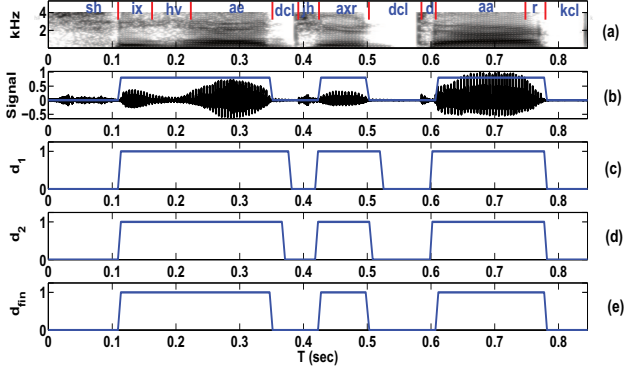


Figure 4: Binary decisions obtained at each step of the proposed algorithm for sonorant segmentation. (a) Spectrogram of speech signal with manually marked phones, (b) speech signal with manual sonorant labels. Binary decision obtained after (c) Step 1, (d) Step 2, and (e) Step 3.

Step 3: The final decision on sonorant regions is obtained by validating the hypothesized decision obtained in previous step with  $\eta$  and  $\delta$  values. Final decision is computed as follows:

$$d_{fin}[i] = \begin{cases} 1, & \text{if } d_2[i] = 1 \ \& \ \eta[i] > 0.0005 \ \& \ \delta[i] > 0.0002 \\ 0, & \text{otherwise,} \end{cases} \quad (5)$$

where  $\eta[i]$ ,  $\delta[i]$ ,  $d_2[i]$  and  $d_{fin}[i]$  refer to the values of  $\eta$ ,  $\delta$ , the second level evidence and the final sonorant decision at the  $i^{th}$  epoch location, respectively. Fig. 4 shows the decisions obtained by the proposed algorithm at every step for a speech utterance.

## 5. Results and discussion

The performance of the proposed approach for sonorant detection in continuous speech is evaluated on the datasets described in Sec. 2. Performance is measured in terms of number of epochs/frames correctly detected in sonorant regions (true positive rate (TPR)), number of spurious epochs/frames hypothesized in non-sonorant regions (false alarm rate (FAR)) and total number of correctly detected epochs/frames in both sonorant and non-sonorant regions (accuracy (Acc)). Epochs derived using ZFR [11] explained in Sec. 3 and the sonorant/non-sonorant decision obtained from manual labeling are used to generate the reference epochs in sonorant and non-sonorant regions. In order to have a comparison with previous works [2, 7, 8], frame based results with frame length 10 msec and frame shift of 5 msec are also provided. Frames occurring after sonorant epochs are considered as sonorant frames, whereas those occurring after non-sonorant epochs as non-sonorant.

Tables 2 - 4 give the performance values obtained on the 3 datasets, i.e., TIMIT, AMI meeting corpus and Telugu dataset, respectively, added with white noise across various SNRs. It is noteworthy that the thresholds on the acoustic features are kept constant for all the 3 datasets at all SNRs. Results given in Table 2 for TIMIT database show that the proposed approach when compared to previous systems [2,7,8,11] performs equally well on clean speech but has a better performance at lower SNR levels. As given in Table 3, performance of the proposed algorithm on AMI meeting corpus, which contains conversational speech, is similar to that of the TIMIT database, which is read speech

(Table 2), even at various SNRs. Also, the performance values given in Table 4 for Telugu language are similar to the results obtained for TIMIT and AMI meeting corpus. This shows that the proposed acoustic features are robust to external degradations and are also language independent.

Table 2: Performance of proposed algorithm for sonorant segmentation on TIMIT database, added with white noise across various SNR levels.

SNR	Epoch based results			Frame based results		
	Acc (%)	TPR (%)	FAR (%)	Acc (%)	TPR (%)	FAR (%)
clean	93.95	94.47	7.53	92.82	93.61	8.00
30 dB	93.93	94.45	7.60	92.79	93.58	8.03
20 dB	93.87	94.39	7.68	92.72	93.53	8.13
10 dB	93.43	94.02	8.49	92.14	92.99	9.02
5 dB	92.37	93.07	8.98	91.01	91.92	9.45
0 dB	90.66	91.04	9.95	89.62	89.85	10.59

Table 3: Performance of proposed algorithm for sonorant segmentation on AMI meeting corpus, added with white noise across various SNR levels.

SNR	Epoch based results			Frame based results		
	Acc (%)	TPR (%)	FAR (%)	Acc (%)	TPR (%)	FAR (%)
clean	90.84	91.53	9.58	90.02	90.43	9.80
30 dB	90.80	91.48	9.66	89.97	90.37	9.84
20 dB	90.64	91.30	9.74	89.70	90.19	9.92
10 dB	90.49	91.16	9.79	89.54	90.01	9.96
5 dB	90.19	90.82	9.94	89.21	89.66	10.12
0 dB	89.81	89.95	10.18	88.42	88.69	11.75

Table 4: Performance of proposed algorithm for sonorant segmentation on Telugu (an Indian language) dataset, added with white noise across various SNR levels.

SNR	Epoch based results			Frame based results		
	Acc (%)	TPR (%)	FAR (%)	Acc (%)	TPR (%)	FAR (%)
clean	94.04	96.16	8.33	93.39	95.42	8.75
30 dB	93.99	96.09	8.39	93.32	95.34	8.82
20 dB	93.87	95.98	8.48	93.21	95.25	8.90
10 dB	93.75	95.86	8.60	93.08	95.11	9.03
5 dB	93.22	95.30	8.97	92.49	94.52	9.56
0 dB	91.93	94.06	9.36	91.07	93.25	9.97

Main source of error is in manual markings of the datasets used, particularly at the vowel boundaries. Most of the missed regions of sonorants are because of poorly articulated nasals, and those of false alarms are due to voice bars /b/ and /d/, which have similarities with nasals and approximants in a few situations.

## 6. Summary and conclusions

In this paper, acoustic features based on excitation source and system characteristics extracted around epochs, are proposed for robust segmentation of sonorant regions in continuous speech. An algorithm is developed to relate the proposed acoustic features in a hierarchical manner for sonorant segmentation. The algorithm has been validated on 3 different datasets i.e., TIMIT, AMI meeting corpus and Telugu language at various SNRs. Evaluation results show that the proposed method is robust to speaker, environment, style of speech (read or conversational) and language variations. Results obtained for sonorant segmentation are encouraging and motivated us to define robust acoustic feature for other broad classes like vowels, nasals etc., which forms our future work.

## 7. References

- [1] A. Juneja, "Speech recognition based on phonetic features and acoustic landmarks," Ph.D. thesis, University of Maryland, College Park, USA, December 2004.
- [2] Ken Schutte and James R. Glass, "Robust detection of sonorant landmarks," in *Proc. Interspeech*, Lisbon, Portugal, pp. 1005-1008, September 2005.
- [3] A. Jansen and P. Niyogi, "Point process models for spotting keywords in continuous speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, pp. 1457-1470, November 2009.
- [4] Sri Harsha Dumpala, Karthik Venkat Sridaran, Suryakanth V. Gangashetty and B. Yegnanarayana, "Analysis of laughter and speech-laugh signals using excitation source information," in *Proc. ICASSP*, Florence, Italy, pp. 975-979, May 2014.
- [5] Sharlene A. Liu, "Landmark detection for distinctive feature based speech recognition," *The Journal of the Acoustical Society of America*, vol. 100, no. 5, pp. 3417-3430, May 1996.
- [6] Kenneth N. Stevens, "Toward a model for lexical access based on acoustic landmarks and distinctive features," *The Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1872-1891, April 2002.
- [7] A. Jansen and P. Niyogi, "Modeling the temporal dynamics of distinctive feature landmark detectors for speech recognition," *The Journal of the Acoustical Society of America*, vol. 124, no. 3, pp. 1739-1758, September 2008.
- [8] Juneja Amit and Carol Espy-Wilson, "A probabilistic framework for landmark detection based on phonetic features for automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 123, no. 2, pp. 1154-1168, February 2008.
- [9] N. Bitar, "Acoustic analysis and modeling of speech based on phonetic features," Ph.D. thesis, Boston University, USA, 1997.
- [10] L. K. Saul, M. G. Rahim and J. B. Allen, "A statistical model for robust integration of narrowband cues in speech," *Computer, Speech and Language*, vol. 15, no. 2, pp. 175-194, April 2001.
- [11] Z. Yessenbayev, "Robust segmentation of speech signal processing using MFCC and acoustic parameters," in sixth Asia modelling symposium, Bali, Indonesia, pp. 103-108, May 2012.
- [12] K. Sri Rama Murthy and B. Yegnanarayana, "Epoch extraction from speech signals," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 8, pp. 1602-1613, November 2008.
- [13] B. Yegnanarayana and P. Satyanarayana Murthy, "Enhancement of reverberant speech using LP residual signal," *IEEE Trans. Speech and Audio Process.*, vol. 8, no. 3, pp. 267-281, May 2000.
- [14] John Garofolo, Lori Lamel, William Fisher, Jonathan Fiscus, David Pallett, Nancy Dahlgren and Victor Zue, "TIMIT Acoustic-Phonetic Continuous Speech Corpus, Linguistic Data Consortium," Philadelphia, USA, 1993.
- [15] I. McCowan, J. Carletta, W. Kraaij, S. Ashby, S. Bourban, M. Flynn, M. Guillelot, T. Hain, J. Kadlec, V. Karaiskos, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, W. Post, D. Reidsma and P. Wellner, "The AMI meeting corpus," in *International Conference on Methods and Techniques in Behavioral Research*, vol. 88, pp. 137-140, 2005.
- [16] Andrew Varga and Herman J. M. Steeneken, "Assessment for automatic speech recognition: ii. noisx-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, July 1993.
- [17] k. Sri Rama Murthy, B. Yegnanarayana and M. Anand Joseph Xavier, "Characterization of glottal activity from speech signals," *IEEE Signal Processing Letters*, vol. 16, no. 6, pp. 469-472, June 2009.
- [18] B. Yegnanarayana and Dhananjaya 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, July 2013.
- [19] Anand Joseph M., Guruprasad S. and B. Yegnanarayana, "Extracting formants from short segments of speech using group delay functions," in *Proc. Interspeech*, Pittsburgh, Pennsylvania, USA, pp. 1009-1012, September 2006.