Significance of Glottal Activity Detection for Speaker Verification in Degraded and Limited Data Condition

Ashutosh Pandey, Rohan Kumar Das, Nagaraj Adiga, Naresh Gupta and S R Mahadeva Prasanna
Department of Electronics and Electrical Engineering
Indian Institute of Technology Guwahati, Guwahati, Assam 781039
Email:{ashutosh.pandey, rohankd, nagaraj, naresh.gupta, prasanna}@iitg.ernet.in

***Abstract*—The objective of this work is to establish the importance of speaker information present in the glottal regions of speech signal. In addition, its robustness for degraded data and significance for limited data is sought for the task of speaker verification. An adaptive threshold method is proposed to use on zero frequency filtered signal to get the glottal activity regions. Feature vectors are extracted from regions having significant glottal activity. An i-vector based speaker verification system is developed using NIST SRE 2003 database and the performance of proposed method is evaluated in degraded and limited data condition. Robustness of proposed method is tested for white and babble noise. Further, short utterances of test data are considered to evaluate the performance in limited data condition. The proposed method based on the selection of glottal regions is found to perform better than the baseline energy based voice activity detection method in degraded and limited data conditions.**

Index Terms- Glottal activity detection (GAD), limited data, speaker verification, degraded condition.

I. Introduction

In speaker verification (SV) voice of a speaker is used to verify his/her claimed identity [1]. I-vector based SV has been considered as the state-of-the-art method for SV because of its low complexity, high performance and easy channel/session compensation [2]. Different existing applications of SV impose the constraints of limited and degraded data. Performance of i-vector based SV system drops significantly for degraded and limited data constraint [3], [4].

Several methods has been proposed to improve the performance of i-vector based SV system in limited and degraded data constraints. In [3], the authors propose the use of vowel and non-vowel like segmentation at analysis stage to improve the performance under degraded condition. Using only vowel like regions has been shown to be more robust to noise [5]. In [4], the authors have explored the performance of i-vector system for short utterance using different channel/session compensation techniques. Correct choice of speech duration for training and testing phase can also improve the performance of i-vector based SV [6]. A minimax strategy can also be used to estimate the sufficient statistics, in order to increase the robustness of the extracted i-vectors [7]. It has been shown that source features can also be combined with conventional vocal tract based feature to boost the performance under limited data conditions [8]- [10].

In this work, an improvement in the performance of i-vector based text independent SV system is sought under degraded and limited data condition by a robust feature selection technique. Use of glottal activity detection (GAD) is proposed to select the regions of interest for feature extraction [11]. Glottal activity is the process of opening and closing of vocal folds which causes most of the significant excitation of vocal tract system (VTS). GAD requires suppressing the effect of non stationary vocal-tract system in the speech sample. A common approach is to first estimate the impulse response of VTS and then use the inverse of it to suppress the effects of VTS on input signal [12], [13]. Linear prediction (LP) analysis is most commonly used method to estimate the characteristics of VTS [14]. In this method linear prediction coefficients are calculated to inverse filter the VTS and LP residual signal is used for glottal region selection. LP residual shows regions of varying energy in glottal regions and has noise like behaviour in non-glottal regions. Efficacy of LP analysis method depends on energy of speech signal and it is not robust to noise.

Zero frequency filter method has been proposed to extract glottal regions [15]. Performance of this method does not depend on the energy of speech signal. Also, it has been observed to be more robust to noise [16]. This is the motivation to use it for degraded condition in this work. In zero frequency based method speech signal is filtered using cascade of two zero frequency resonators and local mean subtraction is performed on filtered signal to get zero frequency filtered signal (ZFFS). Energy of ZFFS can be used for extracting glottal activity regions [15]. In this work use of adaptive threshold on energy of ZFFS is proposed to detect the glottal activity regions. An algorithm is developed to decide the threshold which is applied on the average energy of ZFFS. The selected regions are used for feature extraction. Extracted feature vectors are used in the i-vector based SV system and performance is evaluated

978-1-4799-8641-5/15/$31.00

for limited and degraded data. In this work, it is shown that glottal regions are more robust to noise and also gives better performance in limited data condition.

The rest of the paper is organised as follows: Section II describes the proposed method to extract glottal activity regions. Section III explains the steps involved in the development of i-vector based SV system. Section IV gives the result and analysis. Section V discusses the conclusion of work.

II. Proposed method for Glottal Activity Detection

Opening and closing of vocal folds creates an air flow of variable pressure. This pressure is maximum at closing phase of glottal cycle. Instant opening and closing of folds causes a sudden change in air pressure. Therefore, excitation input can be approximated by the sequence of impulses of varying amplitude [15]. Fourier transform of an impulse function is a constant at all frequencies. Therefore, information about the discontinuity due to impulse can be extracted around any frequency range. Filtering of speech signal around zero frequency can be used to extract information about discontinuity due impulses in excitation signal as, VTS has all its resonances at higher frequencies. Zero frequency resonator exploits the same concept and extracts information about glottal regions. There are two steps involved in proposed method for GAD. First, use the zero frequency filter to get ZFFS and then apply adaptive threshold on ZFFS to get glottal regions. Figure 1 shows the complete flow of proposed method to get the glottal activity regions.

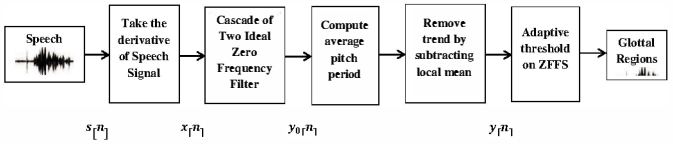

Figure 1: Block diagram showing steps involved in extraction of glottal activity regions

Following steps are used to get ZFFS [16]:

a) Pre-emphasis of speech signal is performed for spectral flattening using first order FIR filter. This is done by taking the derivative of speech signal $s[n]$

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

b) Two ideal zero-frequency resonators are used in cascade to filter the derivative signal. This cascade of two filters provides a roll-off of 24 dB per octave so that higher frequency components are adequately dampened [15]. Resulting signal is given by

$$y_0[n] = \sum_{k=1}^{4} a_k y_0[n-k] + x[n] \tag{2}$$

Algorithm 1 Threshold Calculation

Task: Decide the threshold to use on ZFFS.

Parameters: Number of frames selected by VAD: n_v, Number of frames selected by GAD: n_g, Thresholds T_v and T_g

Initialization: Initialize $k = 0$, set $T_g = T_{g0}$, $T_v = T_{v0}$, $\Delta T_g = \Delta T_{g0}$ and get the value of n_v and n_g

Main Iteration:

while $n_g > n_v$ **do**

- Update $T_g = T_g + \Delta T_g$
- Use T_g to update n_g

Output: T_g

where $a_1 = -4, a_2 = 6, a_3 = -4$ and $a_4 = 1$

c) Average pitch period id required to decide the size of window for local mean subtraction in the next step. Hilbert envelop of LP residual for speech is computed and, periodicity of auto-correlation of Hilbert envelop is analyzed for average pitch period computation.

d) Local mean at each sample is subtracted from $y_0[n]$. This removes the trend in signal accumulated because of zero-frequency resonator. Any window size $(2N+1)$ between one to two pitch period is adequate for mean subtraction [16]. The resulting signal

$$y[n] = y_0[n] - \frac{1}{2N+1} \sum_{m=-N}^{N} y_0[n+m] \tag{3}$$

is the ZFFS. Here $2N+1$ corresponds to the number of local samples considered for mean subtraction. In this work, average pitch period is used as window size for local mean subtraction.

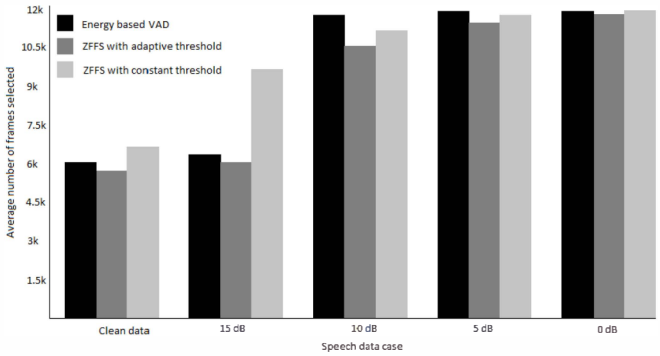

Figure 2: Average number of frames selected by three different methods for different level of white noise, on all training data. Constant threshold equal to 7% of average energy of signal is applied in energy based VAD and ZFFS with constant threshold

Energy of ZFFS can be used for GAD [15]. Therefore, there is a need to use proper threshold on the energy of ZFFS to decide between glottal and non glottal frames.

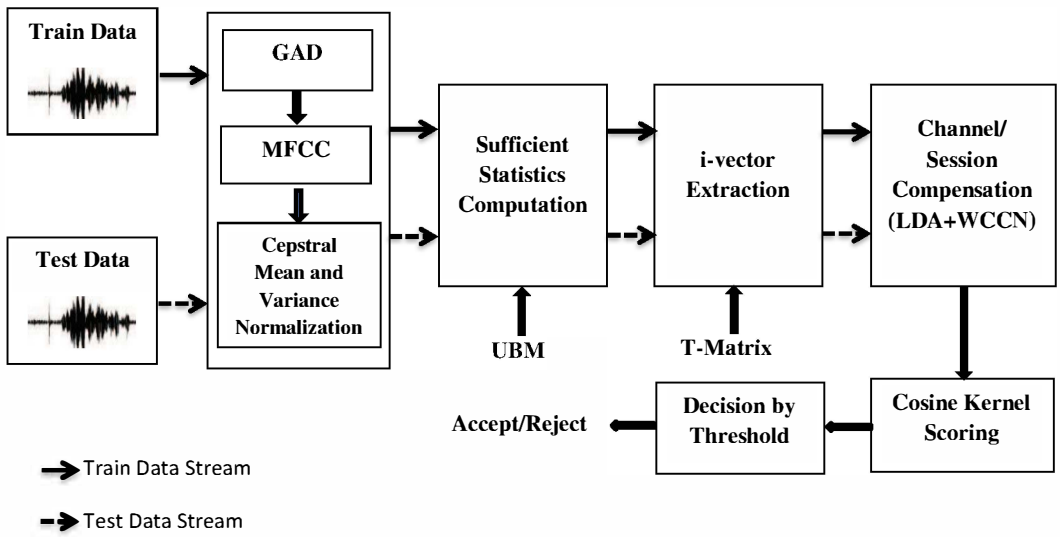

Figure 4: Block diagram showing the steps involved in development of proposed framework for i-vector based SV system

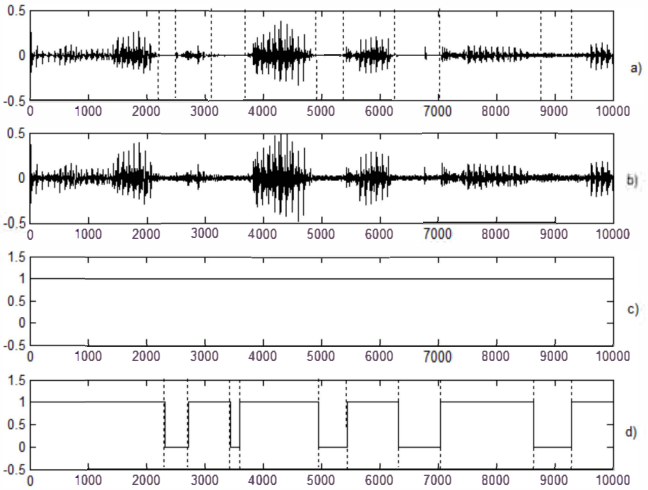

Figure 3: a) Speech signal without noise and marked glottal regions in the speech. b) 10 dB white noise added speech. Proposed method is applied on this speech. c) Region of interest selected by energy based VAD method. d) Glottal regions selected by proposed method in noise added speech

Energy based voice activity detection (VAD) method selects the regions having energy above certain threshold. Many non-glottal frames which are of comparable energy also gets selected as glottal regions that affects the performance of VAD method in degraded condition. Using proper threshold on energy of ZFFS will select only glottal frames. As discussed, VAD method also selects many non glottal-regions, therefore, it is assumed that the number of regions selected by VAD should be greater than the glottal regions. It is observed that using constant threshold on ZFFS does not always select less number of frames than VAD method and varies with the environment condition. This can be observed from the Figure 2. In proposed algorithm information from VAD method is utilized to decide a proper threshold. Algorithm 1 describes the proposed method.

The proposed algorithm for GAD starts from an initial threshold of low value and keeps increasing the threshold in finite steps until the number of regions selected by GAD (n_g) is less than that selected by VAD (n_v). Final threshold is applied on the energy of frames in ZFFS and a decision is made between glottal and non glottal regions. The proposed algorithm always selects lesser number of frames than VAD. Figure 2 illustrates the number of frames selected by three different methods. As can be observed, behaviour of constant threshold on ZFFS does not tally with the argument that number of frames selected by VAD is greater than that selected by GAD.

Figure 3 illustrates the selection of glottal regions in 10 dB white noise. It can be observed that VAD method selects all regions as regions of interest, but many of them are only noise. GAD regions selected by proposed method closely matches with marked glottal regions in original speech signal.

III. Experimental Setup

NIST SRE 2003 database is used to develop an i-vector based SV system [17], [2]. Baseline SV system is developed using energy based VAD method. Figure 4 shows the block diagram for the i-vector based SV system. The GAD module shown in Figure 4 of proposed framework is replaced with VAD for development of baseline SV system. Short term processing is done on the speech signal using fixed window size of 20 ms with shift of 10 ms. In VAD based baseline system, frames having energy greater than 7% of average energy of utterance are selected as speech frames. In proposed algorithm $T_{g0} = 0.1, T_{v0} = 0.07$ and $\Delta T_{g0} = 0.05$ is used. The output of proposed algorithm T_g times the average energy of speech frame in speech signal is used as threshold on energy of frames

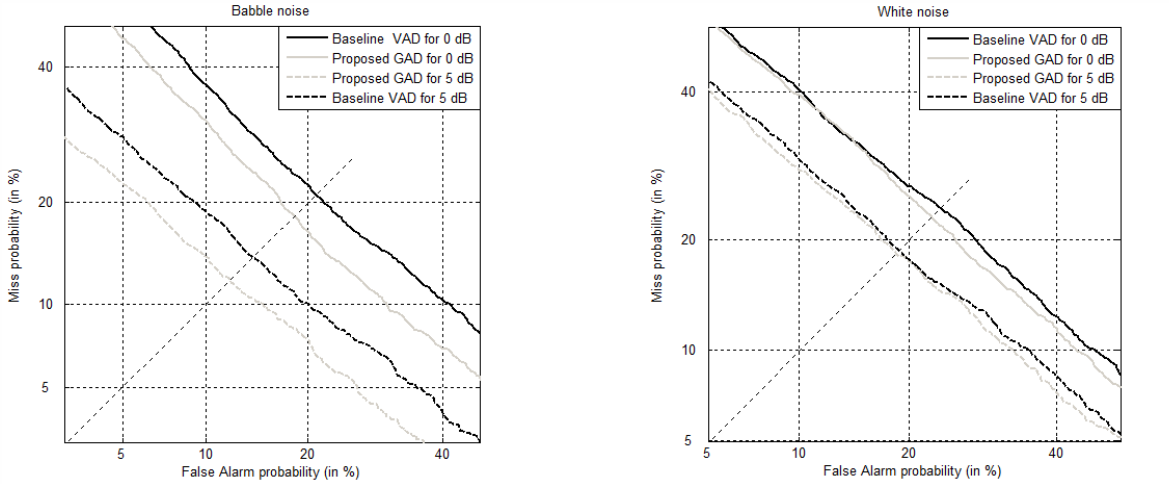

Figure 5: DET curves plotted for proposed vs. baseline SV system in the presence of white and babble noise for sufficient data condition

in ZFFS to decide between glottal and non-glottal regions. Selected frames are further processed for feature extraction. 13-dimensional MFCC features concatenated with their first and second order derivatives are extracted for each of the selected frames in analysis stage. This gives a 39-dimensional feature vector. Cepstral mean subtraction (CMS) and cepstral variance normalization (CVN) are applied for normalization on the extracted features. Switchboard Corpus II cellular data is used as development data. A gender independent universal background model (UBM) is developed using a sub-part of development data [18]. A total variability matrix (T- matrix) of 400 columns is trained using the entire development data [2]. 150-dimensional linear discriminant analysis (LDA) and full dimensional within class covariance normalization (WCCN) is applied for channel/session compensation [2]. Cosine kernel scoring is done between trained i-vectors and test i-vector to get the verification result. For addition of noise in the clean speech data Noisex-92 data is used and performance is evaluated for different SNR condition of white and babble noise [19].

IV. Results and Analysis

Performance evaluation is done over baseline SV system based on VAD method. This system is developed using conventional energy based VAD in the signal analysis stage. Performance is measured in terms of equal error rate (EER) and decision cost function (DCF). Speech data without noise has been termed as *clean speech*. Performance of system for sufficient data and degraded condition is listed in Table I. It can be inferred from the table that proposed GAD method over-performs baseline system in degraded condition. For babble noise condition, absolute improvement in EER and (DCF) values are 0.41 (0.0075), 2.13 (0.051) and 2.58 (0.0538) for 10 dB, 5 dB and 0 dB respectively. Improvement in performance over baseline system increases with decrease in signal to noise ratio (SNR). This demonstrates that the significance of GAD increases with decreasing SNR of babble noise. For white noise, the developed system over-performs the baseline system for all the four SNR of noise used in the experiment. This shows the significance of GAD for white noise. Maximum improvement is observed for 0 dB and 10 dB. Detection error tradeoff (DET) curves for white and babble noise is plotted in Figure 5. It also demonstrate the improvement of developed system over baseline system.

Noise	SNR (dB)	Baseline VAD	Proposed GAD
Clean Speech		2.4 (0.0474)	2.57 **(0.0460)**
White	**15**	7.09 (0.1294)	**6.68 (0.1261)**
	10	15.22 (0.2876)	**14.32 (0.2706)**
	5	18.79 (0.3559)	**18.52 (0.3500)**
	0	23.67 (0.4421)	**22.62 (0.4257)**
Babble	**15**	4.16 (0.0783)	4.83 (0.0902)
	10	8.36 (0.1547)	**7.95 (0.1472)**
	5	14.05 (0.2747)	**11.92 (0.2237)**
	0	20.96 (0.3974)	**18.38 (0.3436)**

Table I: Performance of developed i-vector based SV system using NIST SRE 2003 database for different noisy condition

For limited data constraint, test data of limited duration is considered. The experiments are conducted under sufficient training data and limited test data. Limited test data is created by truncating the test data of NIST SRE 2003 database. Four different cases of limited test data duration, 10s, 5s, 3s and 2s are considered. Performance for limited data for clean speech is listed in Table II.

An absolute improvement of 0.09 (0.0027), 0.45 (0.0065) and 0 (0.0049) is observed in EER and (DCF) over baseline system for test data duration of 5s, 3s and 2s respectively. This observation demonstrates that if test data duration is

Table II: Performance of developed i-vector based SV system using NIST SRE 2003 database for degraded and limited data condition

Noise	SNR (db)	Baseline VAD				Proposed GAD			
Speech Duration		10s	5s	3s	2s	10s	5s	3s	2s
Clean Speech		5.87 (0.1090)	10.52 (0.1977)	16.94 (0.3100)	22.31 (0.4128)	5.96 (0.1081)	**10.43 (0.1950)**	**16.49 (0.3035)**	**22.31 (0.4079)**
White	15	12.42 (0.2299)	19.11 (0.3579)	25.29 (0.4727)	29.45 (0.5562)	**12.29 (0.2306)**	**18.52 (0.3520)**	**24.98 (0.4670)**	**29.45 (0.5558)**
	10	22.58 (0.4266)	28.55 (0.5414)	33.42 (0.6304)	36.40 (0.6852)	**19.87 (0.3764)**	**26.11 (0.4916)**	**30.53 (0.5739)**	**34.33 (0.6501)**
	5	24.71 (0.4666)	30.40 (0.5674)	34.51 (0.6479)	38.30 (0.7248)	**24.62 (0.4653)**	**29.86 (0.5646)**	**34.28 (0.6438)**	**37.62 (0.7066)**
	0	29.95 (0.5668)	35.32 (0.6667)	39.93 (0.7463)	42.82 (0.7930)	**28.86 (0.5436)**	**34.78 (0.6592)**	**38.26 (0.7159)**	**41.15 (0.7710)**
Babble	15	9.08 (0.1698)	16.26 (0.3044)	21.32 (0.4007)	26.29 (0.4951)	10.12 (0.1917)	**16.03 (0.3025)**	22.49 (0.4212)	26.96 (0.5052)
	10	13.87 (0.2626)	21.09 (0.3936)	26.11 (0.4937)	30.67 (0.5765)	**13.46 (0.2537)**	**19.33 (0.3664)**	**25.38 (0.4820)**	**30.04 (0.5639)**
	5	20.60 (0.3851)	27.60 (0.5166)	31.62 (0.5906)	35.05 (0.6643)	**18.02 (0.3380)**	**25.02 0.4697**	**29.49 (0.5537)**	**33.15 (0.6237)**
	0	28.46 (0.5381)	34.28 (0.6471)	37.53 (0.6988)	40.20 (0.7521)	**26.15 (0.4925)**	**32.43 (0.6118)**	**35.00 (0.6608)**	**37.71 (0.7065)**

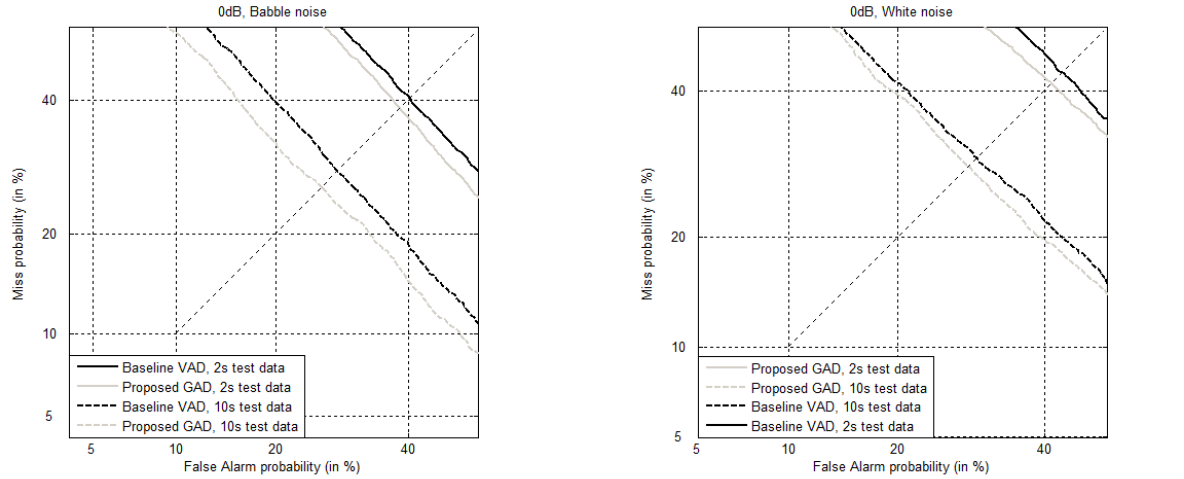

Figure 6: DET curves plotted for 0 dB White and Babble noise in limited data condition

less than or equal to 5s then speaker specific information present in regions other than GAD may not contain significant information for discrimination among speakers. Imposed constraint of both limited and degraded data is evaluated and significant improvement over baseline system is observed. Performance for all the combination of limited and degraded condition is listed in Table II. This demonstrates the significance of GAD for practical system based applications as these application impose combined constraint of limited and degraded data. DET curves for the developed system for limited data condition in presence of white and babble noise are shown in Figure 5.

V. Conclusion

In this work a novel adaptive threshold method is proposed using ZFFS to get glottal activity regions. Robustness of glottal activity regions in degraded data and its significance for limited data condition in SV task is demonstrated. It is concluded that significance of GAD increases with decreasing SNR in speech signal. For limited data condition, it is concluded that, for test data duration less than or equal to 5s, information present in

other than glottal regions may not give much discriminative speaker specific information. The proposed method gives significant improvement in combination of limited and degraded data condition. This signifies the importance of GAD for practical system applications as they impose both limited and degraded data constraint. In future, an optimisation strategy can be employed to get better thresholds to use on zero frequency filtered signal. Also, source information can be combined with proposed method to improve performance of i-vector based SV system.

ACKNOWLEDGMENT

This work is supported by an ongoing project on "Development of speech based multi-level person authentication system" funded by the e-security division of Department of Electronics & Information Technology (DeitY), Govt. of India.

REFERENCES

[1] T. Kinnunen and H. Li, "An overview of text independent speaker recognition: From features to supervectors," *Speech Commun.*, vol. 52, pp. 12 40, Jan. 2010.

[2] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-End Factor Analysis for Speaker Verification", *IEEE Trans. on Audio, Speech, Lang. Process.*, vol. 19, No. 4, pp. 778-798, May 2011.

[3] G. Pradhan and S. R. M. Prasanna, "Speaker Verification by Vowel and Nonvowel Like Segmentatio", *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 4, pp. 854-867, Apr. 2013.

[4] A. Kanagasundaram, R. Vogt, D. Dean, S. Sridharan, and M. Mason, "I-Vector Based Speaker Recognition on Short Utterances", in *Interspeech 2011*, 2011.

[5] S. R. M. Prasanna and G. Pradhan, "Significance of vowel-like regions for speaker verification under degraded condition", *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 8, pp. 2552 2565, May 2011.

[6] A. Sarkar, D. Matrouf, P. Bousquet and J. Bonastre, "Study of the Effect of I-vector Modeling on Short and Mismatch Utterance Duration for Speaker Verification", in *Interspeech*, 2012 , ISCA.

[7] V. Hautamaki, Y. Cheng, R. Padmanabhan, C. Lee, "Minimax i-vector extractor for short duration speaker verification", *Proc. Interspeech 2013*, 2013, Lyon, France.

[8] K. Murthy and B. Yegananarayana, "Combining evidence from residual phase and MFCC features for speaker recognition", *IEEE Signal Processing Letters*, vol. 13(1), pp. 5255, 2006.

[9] R K Das, Abhiram B, S R M Prasanna and A G Ramakrishnan, "Combining Source and System Information for Limited Data Speaker Verification", *Proc. Interspeech 2014*, 2014, Singapore, pp. 1836-1840.

[10] R K Das, D. Pati and S R M Prasanna, "Different aspects of source information for limited data speaker verification", *Proc. National Conference on Communication (NCC) 2015*, 2015, IIT Bombay, pp. 1-6.

[11] N.Adiga and S R M Prasanna, "Detection of Glottal Activity using Different Attributes of Source Information", *IEEE Signal Processing Letters*, Vol. 22, No. 11, November 2015.

[12] P. Alku, J. Vintturi, and E. Vilkman, "On the linearity of the relationship between the sound pressure level and the negative peak amplitude of the differentiated glottal flow in vowel production", *Speech Commun.,* vol. 28, pp. 269-281, Aug. 1999.

[13] R. Smits and B. Yegnanarayana, "Determination of Instants of Significant Excitation in Speech Using Group Delay Function", *IEEE Trans. Speech Audio Process.*, vol. 3, no. 5, pp. 325-333, Sep. 1995.

[14] J. Makhoul, "Linear prediction: A tutorial review," *Proc. IEEE*, vol. 63, pp. 561-580, Apr. 1975.

[15] K. Sri Rama Murty, B. Yegnanarayana, , and M. Anand Joseph, "Characterization of Glottal Activity From Speech Signals", *IEEE Signal Processing Letters*, Vol. 16, No. 6, June 2009.

[16] K. Sri Rama Murty and B. Yegnarayana, "Epoch Extraction From Speech Signals," *IEEE trans. audio, speech, lang. process.*, vol. 16, no. 8, November 2008.

[17] "The NIST Year 2003 Speaker Recognition Evaluation Plan", *NIST*, Feb 2003.

[18] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models", *Digital Signal Processing*, vol. 10, no. 1- 3, pp. 19 41, 2000.

[19] A. Varga and H. J. M. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems", Speech Commun., vol. 12, no. 3, pp. 247 251, Jul. 1993.