Significance of Vowel-Like Regions for Speaker Verification under Degraded Condition

S. R. Mahadeva Prasanna and Gayadhar Pradhan
Department of Electronics and Communication Engineering
Indian Institute of Technology Guwahati, Assam, India
Email: {*prasanna,gayadhar*}*@iitg.ernet.in*

***Abstract*—Vowel-like region (VLR) in speech include vowel, semi-vowel and diphthong. VLR can be identified using vowel-like region onset point (VLROP) event. By production, the VLR has impulse-like excitation and therefore information about the vocal tract system may be better manifested in them. Also the VLR is relatively high signal to noise ratio (SNR) region. Speaker information extracted from such a region may therefore be more speaker discriminative and relatively less affected by the degradations like noise, reverberation and sensor mismatches. Due to this, better speaker modeling and reliable testing may be possible. In this work, VLRs are detected using the knowledge of VLROPs during training and testing. Features from the VLRs are then used for training and testing speaker models. As a result a significant improvement in the performance is reported for speaker verification under degraded condition.**

***Index Terms*—vowel-like region, vowel-like region onset point, speaker information, speaker verification, degraded condition.**

I. Introduction

The state-of-art speaker verification (SV) systems provide good performance when the speech signal is of high quality and free from any mismatch [1]. Such a speech signal is treated as clean speech in the present work. However, in most practical operating conditions, the speech signal is affected by different degradations like background noise, reverberation, sensor mismatch and channel mismatch, resulting in degraded speech. The accuracy of SV system falls significantly under degraded condition [2]. There are many techniques available for dealing with the mismatch between training and testing conditions due to degradation. These techniques may be broadly divided into two groups. In the first group, the mismatch is compensated by removing the degradation effect from both training and testing speech signals. In the second group, the parameters of the speaker model are biased towards the testing environment to match the testing conditions.

In the first group of techniques, the compensation is done at the signal level, feature level, score level or all of them. The methods used for removing the effect of noise and reverberation at the signal level aimed at dealing with high level degradation, involve identifying the high signal to noise ratio (SNR) [3], [4] or signal to reverberation ratio (SRR) regions and enhance them in the time domain [5] or estimate the noise and subtract in frequency domain [6] or estimate reverberation and eliminate the same in the cepstral domain [5], [7]. The popular methods used to remove low level degradation at the feature level include cepstral mean subtraction (CMS) [8], CMS followed by cepstral variance normalization (CVN) and relative spectral (RASTA) filtering [9]. At the score level, the effect of low level degradation can be minimized by suitable score normalization techniques like Hnorm and HTnorm [10] [11]. In the second group of techniques, methods like maximum a posteriori (MAP) [12], maximum-likelihood linear regression (MLLR) [13] and Bayesian maximum a posteriori linear regression (MAPLR) [14] are used for adapting model parameters to the testing environment. Speech from the testing environment is required for adaptation and may not be possible under all practical scenarios.

On top of all these approaches, the performance of the speaker verification system can be further improved by selecting only those speech regions, based on the nature of speech production, that are relatively more speaker discriminative and less affected by various degradations. This can be achieved using the knowledge of vowel-like region onset point (VLROP). VLROP helps in identifying *vowel-like* regions which include vowels, semivowels and diphthongs, that are high SNR regions from the production perspective. Hence they may be more speaker discriminative and exploring this aspect is the focus of this work. The proposed approach is motivated from the earlier studies on using the high SNR or SRR regions from production perspective for speech enhancement [3]–[5], [7].

VLROP is defined as the instant at which the onset of vowel-like region takes place. VLROP corresponds to vowel onset point (VOP) in case of vowels [15], onset of semivowel and onset of diphthong. The typical cases in which VOP occurs include isolated vowel, consonant vowel (CV) and consonant-cluster vowel (C^nV, where $n > 1$). Existing VOP detection methods can be used for the detection of VLROPs. If the VOP detection method is not perfect (i.e., 100% performance), then the errors are manifested in terms missing and spurious VOPs, and also the resolution with which VOPs are detected [16]. Majority of the errors are observed to be due to the cases of semivowels and diphthongs [16]. However, for the speaker verification task we need vowel, semivowel and diphthong regions. Therefore by including the onset of semivowels and diphthongs, the performance of VOP detection can be significantly improved. Hence the motivation for defining VLROP event instead of VOP. With the help of VLROP event, the vowel-like regions can be detected. The main requirement of VLROP detection algorithm is robustness under degraded condition. When it is robust, then similar regions can be selected for both training and testing of SV systems.

The major excitation that provides speaker characteristics to the speech signal is the vibration of vocal folds [17]. Vowel-like regions are produced using the vocal folds vibration and hence may have relatively more speaker information compared to *non-vowel-like* regions from the excitation source perspective. Vowel-like regions are produced by exciting the vocal tract system using impulse-like excitation due to the sudden closure of vocal folds. Due to impulse-like excitation, the impulse response of the vocal tract system may be better manifested and hence more speaker discriminative from vocal tract system perspective. Vowel-like regions are produced by keeping the vocal tract in an open configuration which offers relatively less obstruction for the air flow and hence high SNR regions. Therefore if we have a method for detecting vowel-like regions and use speaker information from such regions, then better speaker modeling as well as reliable testing may be possible. This may help in increasing the robustness of speaker verification system under degraded condition.

In the existing speaker verification systems, speech regions are separated out from the silence regions based on energy threshold, and features from the speech regions are used for modeling and testing. In the proposed approach, vowel-like regions are separated out from the non-vowel-like regions based on the knowledge of VLROP, and features from the vowel-like regions are used for modeling and testing. Suppose if clean speech collected in matched condition is used, then the proposed approach may provide better performance in terms of requirement of data. That is, it may provide nearly same performance using relatively less amount of speech data from the vowel-like regions. Alternatively, the merit of vowel-like regions may be found under degraded condition. If degraded speech collected in mismatched condition is used, then the proposed approach may provide better performance. As mentioned above, this is due to the robustness of vowel-like regions from the production perspective for different degradations.

The rest of the paper is organized as follows: Methods for the detection of VLROPs and VLRs are described in Section II. Proposed speaker verification system using vowel-like regions is described in Section III. The experimental studies are described in Section IV. The experimental results are discussed in Section V. The summary of the present work and scope for the future work are mentioned in Section VI.

II. Detection of VLROP and vowel-like regions in Degraded Speech using excitation source information

In the present work, VLROP refers to the instant at which the onset of vowel-like region takes place. The vowel-like regions are prominent regions in the speech signal due to high amplitude, periodicity, long duration and less zero crossing rate. Considering these distinct properties of vowel-like regions only for the case of vowels, a number of vowel onset point (VOP) detection algorithms have been proposed like locating the rapidly increasing peaks in the amplitude spectrum [18], zero-crossing rate, energy, pitch information [19], training neural network with the trends of energy, zero crossing rate and spectral flatness at the VOP [20] and using excitation source information [21]. More recently, a combined method using the excitation source, spectral peaks and modulation spectrum information is proposed for the detection of VOP [16]. In all these methods, the failing cases are reported for semivowels and diphthongs, due to their similarity in production characteristics with the vowels. Hence attempts are underway to improve VOP detection by devising methods to deal with semivowel and diphthongs. Alternatively, from the speaker verification perspective all the three categories are equally important. Therefore existing VOP detection methods used as it is may provide significant improvement in performance for the case of VLROP detection.

All the VOP detection methods mentioned above are evaluated for clean and broadband speech. For any practical application, the speech signal may be degraded and narrowband in nature. In degraded speech, many features like energy, zero crossing rate and spectral flatness may fail to hypothesize VOPs properly and also increase the number of spurious ones. The same thing will be true for VLROPs also. We therefore need robust features to deal with degradation. The periodicity information present in the Hilbert envelope (HE) of linear prediction (LP) residual is relatively less affected by various degradations and hence a pitch extraction method under adverse conditions is proposed in [22]. Also a VOP detection method is exclusively developed using this information in [21]. The energy associated with HE of the LP residual is proposed to be robust to various degradations compared to signal energy. This is due to the elimination of most of the spectral information due to various degradations and also enhanced periodicity information. Recently a zero frequency filtering (ZFF) approach is proposed for detecting epochs in speech [23]. Since, the zero frequency resonator exploits only signal energy around zero frequency region and attenuates all other information [23], the resonator output signal may provide robustness to various degradations. The strength of excitation derived from ZFF has been demonstrated earlier to have better discriminating ability at the unvoiced-voiced transitions [24]. It is therefore proposed that by combining the features derived from the HE of the LP residual with features from ZFF, it may be possible to provide robustness to the VLROP evidence and also reduce most of the spurious detections in degraded speech. Since both the features contain mainly information about excitation source, the proposed VLROP detection algorithm is termed as *VLROP detection using excitation source information*.

A. VLROP evidence from HE of LP residual

The VLROP evidence using the HE of LP residual is obtained by processing the speech signal through the following steps: The speech signal is processed in blocks of 20 ms with a shift of 10 ms. For each 20 ms block, 10^{th} order LP analysis is performed to estimate the linear prediction coefficients (LPCs) [25]. The time-varying inverse filter is constructed using these LPCs. The speech signal is passed through the inverse filter to extract the LP residual signal. The time varying nature of excitation source characteristic is further enhanced by computing the Hilbert envelope of LP residual [26].

Let $e_a(n)$ be the analytic signal of a given signal $e(n)$. Then,

$$e_a(n) = e(n) + je_h(n) \tag{1}$$

where $e_h(n)$ is the Hilbert transform of $e(n)$. The Hilbert transform is computed as

$$e_h(n) = IDTFT(E_H(\omega)), \tag{2}$$

where

$$E_H(\omega) = \begin{cases} +jE(\omega), -\pi \leq \omega < 0 \\ -jE(\omega), 0 \leq \omega \leq \pi \end{cases} \tag{3}$$

and $E(\omega)$ is the DTFT of $e(n)$. DTFT refers to discrete time Fourier transform and IDTFT refers to inverse of DTFT.

Let $h_e(n)$ be the HE. It is defined as the magnitude of $e_a(n)$ i.e.,

$$h_e(n) = |e_a(n)|\,. \tag{4}$$

$$h_e(n) = \sqrt{e^2(n) + e_h^2(n)}. \tag{5}$$

Let $\phi(n)$ be the phase of $e_a(n)$. It is defined as

$$\phi(n) = \tan^{-1}\big(\frac{e_h(n)}{e(n)}\big). \tag{6}$$

Therefore Hilbert envelope $h_e(n)$ of LP residual $e(n)$ is defined as $h_e(n) = \sqrt{e^2(n) + e_h^2(n)}$, where, $e_h(n)$ is the Hilbert transform of $e(n)$.

For every 5 ms block with one sample shift, the maximum value of the Hilbert envelope of LP residual is noted to construct smoothed excitation contour. The change in the excitation characteristics at the VLROP event is detected by convolving the smoothed excitation contour with a first order Gaussian differentiator (FOGD) of length 100 ms (800 samples for 8 kHz) and standard deviation as one sixth of the window length (134 for 8 kHz) [16], [21]. This convolved output is termed as *VLROP evidence from HE of LP residual*.

B. VLROP evidence using zero frequency filtered signal (ZFFS)

The property of impulse like discontinuity is exploited in ZFF method. The time domain representation of impulse function has an equivalent frequency domain representation of impulses uniformly located at all the frequencies including zero frequency, separated by fundamental frequency, forms the basis for ZFF method [23]. In ZFF method, speech is passed through a resonator located at the zero frequency which preserves the signal energy around the impulse present at zero frequency and removes all other information, mainly due to the vocal tract resonances. The trend in the output of the zero frequency resonator is removed further by considering a window of length one to two pitch periods and the trend removed signal is termed as the zero frequency filtered signal (ZFFS) [23]. The positive zero crossings of the ZFFS give the location of epochs.

The algorithmic steps to estimate the epochs in speech by ZFF are as follows [23]:

- Difference input speech signal $s(n)$

$$x(n) = s(n) - s(n-1) \tag{7}$$

- Compute the output of cascade of two ideal digital resonators at 0 Hz

$$y(n) = -\sum_{k=1}^{4} a_k y(n-k) + x(n) \tag{8}$$

  where a_1 = 4, a_2 = -6, a_3 = 4, a_4 = -1
- Remove the trend i.e.,

$$\hat{y}(n) = y(n) - \bar{y}(n) \tag{9}$$

  where $\bar{y}(n) = \frac{1}{(2N+1)}\sum_{n=-N}^{N} y(n)$ and $2N+1$ corresponds to the average pitch period computed over a longer segment of speech
- The trend removed signal $\hat{y}(n)$ is termed as ZFFS.
- The positive zero crossings of the filtered signal will give the location of the epochs.

The first order difference of ZFFS contains information of slope at epochs that can be treated as strength of excitation at the epochs [24]. The second order difference of the ZFFS is computed to calculate the change in the strength of excitation. For every 5 ms block with one sample shift, the maximum value of the second order difference of ZFFS is noted to construct smoothed excitation contour. The change in the excitation characteristics at the VOP event is detected by convolving the smoothed excitation contour with the FOGD of length 100 ms (800 for 8 kHz) and standard deviation as one sixth of the window length (134 for 8 kHz). The convolved output is termed as *VLROP evidence using ZFFS*.

Finally, the *VLROP evidence using the excitation source information* is obtained by adding the two evidences and normalizing with respect to the maximum value of the sum. The peaks in the combined evidence are selected by finding the maximum value between two successive positive to negative zero crossings with some threshold to eliminate the spurious ones. The peaks location in the combined evidence are hypothesized as the VLROPs.

C. Performance of VLROP detection

The performance of proposed VLROP detection method is evaluated for clean speech using 60 speakers data from the TIMIT database for the two sentences *Don't ask me to carry an oily rag like that* and *She had your dark suit in greasy wash water all year* [27]. The phoneme transcription file originally available in TIMIT database contains the location of phone boundaries. Most of these reference markings correspond to the location of VLROPs, but it may not be true for all. For an example, the 2^{nd} phoneme location of the speech file, TEST/DR7/MPABO/SA2.wav is not the true VLROP location. The true VLROP location is at the sample number 6881 which is 121 sample advanced to the phoneme boundary available in the database. Therefore for the present performance evaluation using the original phoneme marking as the reference, the VLROP instants are marked manually by considering phoneme boundaries as initial candidates for VLROPs and then refining them with the help of waveforms and spectrograms. Using these manually marked references, the performance of the proposed method is measured using the following parameters [28]:

- *Identification rate (IR):* The percentage of reference VLROPs that are matched by the detected VLROPs within the vowel-like regions;
- *Miss rate (MR):* The percentage of reference VLROPs for which no VLROPs detected within the vowel-like regions;
- *Spurious rate (SR):* The percentage of detected VLROPs, which are detected outside the vowel-like regions;
- *Identification accuracy (IA):* For each identified VLROP, the timing error between the identified VLROP and corresponding reference VLROP is measured and finally the standard deviation of the timing error is computed to find the identification accuracy.

The performance of proposed VLROP detection algorithm is given in Table I. For comparison, individual performances of HE of LP residual, ZFFS and method based on most recent VOP detection method [16] are also given in the same table. As it can be observed the proposed method is better both in terms of performance and also resolution. The most recent VOP detection method uses sum of ten largest peaks in the spectrum, smoothed HE of LP residual and modulation spectrum as features. The proposed VLROP detection method based on excitation source information provides the best performance. Even though the HE of LP residual provides slightly poorer performance compared to ZFFS, it combines well to improve the resolution of VLROP. The possible reason for high spurious VLROPs using the most recent VOP detection method is due to the emphasis provided for low energy regions by peak enhancement [16].

TABLE I: Performance of VLROP detection methods using excitation source information and based on recent VOP detection method for speech signals from TIMIT database. The abbreviations Existing, HE, ZFFS and ESI refer to performance due to most recent VOP detection method, HE of LP residual, zero frequency filtered signal and proposed excitation source method.

Method	IR (%)	MR (%)	SR (%)	IA (ms)
Existing	91.86	8.14	17.44	30.9
HE	90.87	9.13	9.47	44.51
ZFFS	95.61	4.39	4.38	24.97
ESI	94.90	5.10	6.90	23.87

In order to evaluate the robustness of proposed algorithm in degraded environment, the same set of TIMIT speech files are reconstructed using the factory-1 noise and white noise of NOISEX-92 database [29]. The energy level of the noise is scaled such that the overall SNR of the noise reconstructed speech is maintained at 3 dB. The performance of the proposed VLROP detection and based on most recent VOP detection methods for noise reconstructed speech are given in the Table II. By comparing the Tables I and II, it can be observed that spurious rate in case of recent VOP detection method and HE envelope of LP residual is relatively more compared to the ZFFS for clean as well as degraded speech, and this difference is more prominent in degraded speech. As mentioned earlier it can be observed from both the tables, the performance of the proposed method is better in every respect compared to each individual feature and the performance is almost same for both clean and degraded speech. Since the performance of VLROP detection is good and robust, for the performance SV system is expected to be better using vowel-like regions.

TABLE II: Performance of VLROP detection methods using excitation source information and based on recent VOP detection method for speech signals from TIMIT database. The abbreviations Existing, HE, ZFFS and ESI refer to performance due to most recent VOP detection method, HE of LP residual, zero frequency filtered signal and proposed excitation source method.

Noise	Method	IR (%)	MR (%)	SR (%)	IA (ms)
factory-1 Noise	Existing	91.94	8.06	24.15	34.15
	HE	89.53	10.47	8.58	51.52
	ZFF	96.86	3.14	9.83	30.09
	ESI	95.25	4.75	5.72	25.19
White Noise	Existing	94.45	5.55	30.67	25.21
	HE	95.97	4.03	16.63	36.18
	ZFF	96.42	3.58	9.12	21.61
	ESI	96.78	3.22	11.53	19.75

D. *Detection of vowel-like regions from degraded speech*

The Fig. 1(b), shows the VLROP evidence for a segment of speech taken from NIST-2003 speaker recognition database. The corresponding evidences for white noise degraded speech are shown in the Fig. 1(c)-(f) for 9 dB, 6 dB, 3dB and 0 dB SNR, respectively. The arrow marks in each evidence correspond to the hypothesized VLROP locations, obtained by the proposed method. The Fig. 1(a) and (b) shows that the hypothesized VLROPs nearly correspond to the starting of vowel-like regions. The small vowel-like regions around the time location 0.042 sec and 1.1 sec are missed due to smaller peak value in the evidence. The Fig. 1(c)-(f) shows that the evidences of noise degraded speech are modified differently compared to the original evidence depending on the level of noise. However, the VLROPs detected and spurious ones remain almost same as in the clean speech. Hence the robustness of the proposed VLROP detection method.

The selection of vowel-like regions using the VLROPs and speech regions using an energy based threshold for the same segment of speech and same noise levels are shown in Fig. 2. The vowel-like regions are selected by taking the 100 ms region right to the hypothesized VLROP locations, which are represented in solid lines. The choice of 100 ms is based on the assumption of average duration of vowel-like region to be of about 100 ms in continuous speech. The speech regions are identified as the speech frames above the energy threshold ($0.1 \times average\ energy$), which are represented in dotted lines. The Fig. 2(a) indicates in a clean speech, since most of the non-speech regions are silence frames, the speech regions selection by energy threshold is accurate. In the proposed method, as discussed earlier by imposing a fixed duration of 100 ms, some of the non-vowel regions get selected for short vowels and some of the vowel-like portions are missed for long vowel-like regions, although the VLROP detection is perfect. But, all the selected regions are mostly vowel-like regions. The Fig. 2(b)-(e) shows that for noise degraded speech also same vowel-like regions are selected, even for severely noise degraded speech. Hence the robustness of detection of vowel-like regions using VLROP. The speech region selection by energy based threshold is affected for 9 dB SNR and completely failed for 6 dB, 3 dB and 0 dB SNR. From the experiments, it is observed that by putting a very high threshold around 70% of speech frames get eliminated for clean speech. In a practical application where clean and noisy speech are equiprobable, by putting a very high threshold eliminates most of the speech frames for clean speech and putting a low threshold selects most of the noise frames. Further, if the noise appears randomly within a particular speech, neither a low threshold nor a high threshold will select the proper speech frames.

The effectiveness of proposed algorithm can be better investigated using clean speech (speech recorded over sensor H01) and degraded speech (speech recorded over sensor D01) of IITG Multi-Variability (MV) speaker recognition database [30]. The speech file shown in Fig. 3(a) is a segment of speech recorded in sensor H01 (clean speech) and the same segment speech recorded over sensor D01 (degraded speech) is shown in Figure 3(c). The VLROP evidences corresponding to clean and degraded speech are shown in Fig. 3(b) and (d), respectively. The Fig. 3 indicates that the vowel-like regions selection by the proposed method is almost same for clean and degraded speech. Alternatively, the speech regions selection by energy based threshold is accurate for clean speech and most of the non-speech frames are selected as speech frames for degraded speech.

The above results indicate that the vowel-like regions can be selected from a degraded speech using VLROP in a robust manner. Therefore, using these relatively less degradation affected and more speaker discriminating regions, a better speaker verification system can be developed under degraded conditions.

III. Speaker Verification using Vowel-like Regions

A. Database

The performance of proposed speaker verification system is evaluated on complete NIST-2003 speaker recognition database at the first level to study the discriminating speaker information present in the vowel-like regions. Then the *factory-1 noise* and *white noise* are taken from NOISEX-92 database to create the noise degraded speech. The energy level of the noise is scaled such that the overall SNR of the noise degraded NIST-2003 test speech is maintained at 9, 6, 3 and 0 dB. Performance of the speaker verification (SV) system is then evaluated on the NIST-2003 for original train speech and noise degraded test speech to study the performance of proposed system on a large speaker recognition database for noise degraded test speech. Performance of the SV system is also evaluated on noise degraded trained speech and original test speech for 3 dB SNR level. Finally, the performance of the SV system is evaluated on IITG multi-variability (MV) speaker recognition database developed in house for evaluating speaker recognition systems for speech data in Indian scenario. The IITG MV database is collected in a setup having five different sensors, two different environments, different Indian languages and two different styles. The five different sensors include headphone microphone (H01) mounted close to the speaker, in built tablet PC microphone, two mobile phones and one digital voice recorder (D01). Except for the headphone microphone, all the other four sensors are placed at a distance of about 2-3 feet from the speaker. Speech was recorded simultaneously

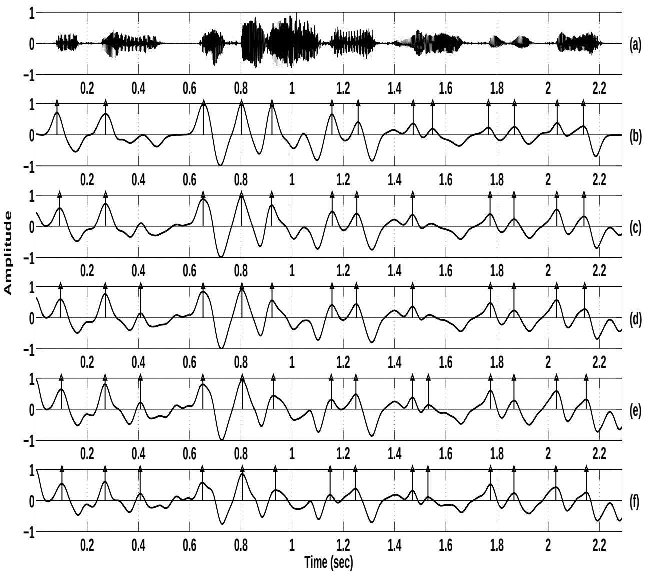

Fig. 1: VLROP evidences for degraded speech. (a) Segment of speech taken from NIST-2003 speaker recognition database, (b) VLROP evidence for clean speech, (c)-(f) VLROP evidences for white noise degraded speech with over all SNR level 9 dB, 6 dB, 3dB and 0 dB, respectively.

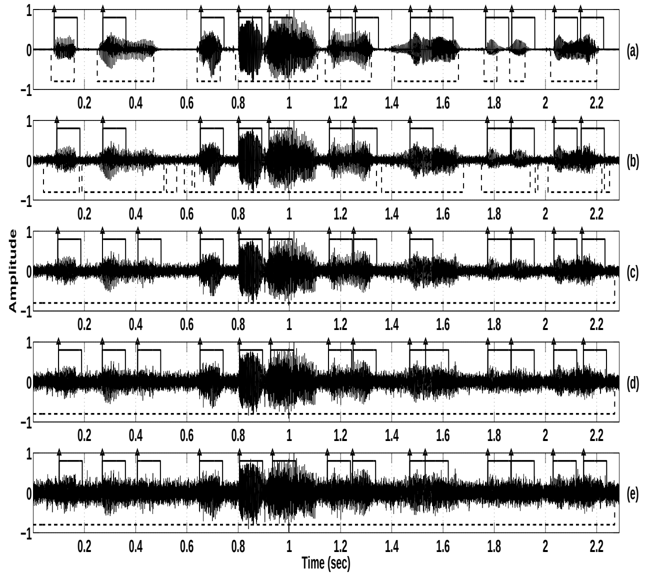

Fig. 2: Vowel-like regions (solid lines) and speech regions (dotted lines) detection in degraded condition. (a) Segment of speech taken from NIST-2003 speaker recognition database, (b)-(e) noise degraded speech with over all SNR level 9 dB, 6 dB, 3dB and 0 dB, respectively.

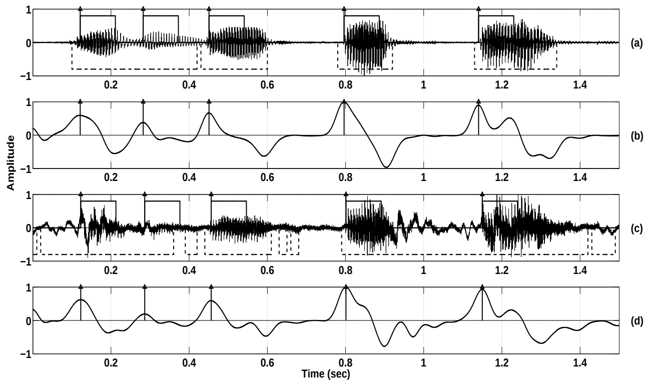

Fig. 3: Vowel-like regions (solid lines) and speech regions (dotted lines) detection for clean and degraded speech of IITG MV speaker recognition database. (a) Clean speech (Speech record over sensor H01), (b) VLROP evidence for clean speech, (c) degraded speech (speech recorded parallely over sensor D01), (d) VLROP evidence for degraded speech.

over these sensors. Speech recorded in headphone microphone and inbuilt tablet PC microphone are at 16 kHz and stored with 16 bits/sample resolution. Speech recorded in digital voice recorder is at 44.4 kHz and stored with 16 bits/sample, which is later resampled to 16 kHz and stored at 16 bits/sample. The speech recorded in two mobile phones are at 8 kHz and sampled at 16 bits/sample. The recording was done in two different environments, namely, office/laboratory and hostel rooms. The recording was done in two languages for each speaker, namely, English and favorite language of the speaker which happens to be one of the Indian languages like Hindi, Kannada, Oriya, Telugu and so on.

B. Detection of Vowel-like regions

As described in Section II, VLROPs are determined using the excitation source information derived from the speech signal. Using each hypothesized VLROP as the anchor point, 100 ms regions right to the VLROPs are marked as vowel-like regions. In case of speaker verification using vowel-like regions, features derived only from these regions are used for training and testing. Alternatively, in case of speaker verification using conventional approach, regions identified based on energy threshold are used.

C. Feature Extraction

In the training and testing process, the speech signal is processed in frames of 20 ms duration at 10 ms frame rate. For each 20 ms Hamming windowed frame, Mel frequency cepstral coefficients (MFCC) are calculated using 22 logarithmically spaced filter banks [31]. The first 13 coefficients excluding zeroth coefficient value are used as a feature vector. Delta (Δ) and delta-delta ($\Delta\Delta$) of MFCC are computed using two preceding and two succeeding feature vectors from the current feature vector [32]. Thus the feature vector will be of 39 dimension with 13 MFCC, 13 ΔMFCC and 13 $\Delta\Delta$MFCC.

D. Parameter normalization

In this work the degradation effect is compensated in the feature domain using CMS followed by CVN. The blind deconvolution like CMS not only subtracts the channel and environmental effect, it also removes some amount of speaker information. Therefore the CMS reduces the performance when there is not much variability in the recording sensor and environment, and it improves the performance when there is variation [33]. In the present experimental setup for the NIST-2003, noise degraded NIST-2003 and for the two mismatched experiments of IITG MV speaker recognition database variation is present from training to testing session. For the clean and sensor matched experiment of IITG MV database, there is no variation in sensor and environment, but for all the three experiments of IITG MV database, the models are built by adapting a sensor mix universal background model (UBM). The speech recorded in digital voice recorder is also severely affected by noise and reverberation. Looking at all these factors, in the present experimental setup the feature vectors are normalized to fit a zero mean and unit variance distribution.

E. Speaker modelling and testing

The main motivation of this work is to study the discriminative information present in the vowel-like regions for speaker modeling and testing in degraded environments. Except for deriving frames from vowel-like regions, there is no difference in the steps of speaker verification system development. Hence, the extensively used Gaussian mixture model (GMM)-UBM based speaker modelling is employed [34]. The UBM is a large GMM which represents the speaker independent distribution of features. The UBM is represented by a weighted sum of C component densities as $U = \{\eta_c, \mu_c, \Sigma_c\}$, $c = 1, ...C$, where μ_c, Σ_c and η_c are the mean vector, covariance matrix, and weight associated with each mixture c, respectively. The speaker dependent models are built by adapting the components of UBM with the speakers training speech using maximum a posteriori (MAP) algorithm [34]. During the testing stage, the log likelihood scores are calculated from both the claimed model and UBM.

F. Baseline SV system

In order to compare the performance obtained using vowel-like regions, we have developed another speaker verification system based on energy threshold ($0.1 \times average\ energy$) which is termed as *baseline system*. The only difference between baseline system and proposed system lies in the selection of speech frames during training and testing process. In the baseline system, the speech frames are selected by using an energy threshold and in the proposed case using vowel-like regions.

G. Performance Comparison

The relative improvement in the performance of the SV system using only vowel-like regions is compared to the baseline system in terms of EER, as follows:

$$EER_R = \frac{(EER_B - EER_V)}{EER_B} \times 100\ \% \tag{10}$$

where EER_R, EER_B and EER_V are the relative improvement in EER, EER of the baseline SV system and the SV system using the vowel-like regions, respectively.

IV. Experimental Studies

In the present experimental set up three experiments are conducted on NIST-2003 speaker recognition database as follows:

1) *Original NIST-2003*: NIST-2003 speaker recognition database is used for the performance evaluation.
2) *Noise degraded NIST-2003 test speech*: Original NIST-2003 train speech is used for training the models and noise degraded speech is used for testing.
3) *Noise degraded NIST-2003 trained speech*: NIST-2003 noise degraded train speech is used for training the models and original test speech is used for testing.

For these three sets of experiment, ten hours of UBM speech were selected from randomly selected 100 male and 100 female speakers of switchboard cellular part 2 Audio database [35]. Using each gender speech, two gender dependent 512 mixture size GMM are built, one for the male speakers and other for the female speakers. Finally, a 1024 mixture size gender independent UBM is built by pooling the two models and normalizing the weights [34]. Two such gender independent UBMs are built one for the baseline SV system using the speech frames selected by energy based threshold and another for the proposed system using the vowel-like regions. During the time of model adaptation and testing, the respective UBM is used.

The performance of both the systems are finally evaluated on IITG MV database for a real environment degraded speech. For this experiment, we consider 100 speakers set of IITG MV database, which includes 75 male speakers and 25 female speakers. The initial 2 minutes of speech data recorded in the first session is used for building the models. For each speaker, 10 speech segments between 30-45 sec duration from the second session are taken as test utterances. Therefore for 100 speakers set there are in total 1000 test trials. In the testing process, each test segment is tested against 11 models, out of which one is genuine model and rest are impostor models. Out of the five sensors, speech recorded over digital voice recorders (D01), due to its high sensitivity and position, is worst affected by environmental noise like air conditioner, fan sound, room reverberation and other surrounding noises present at the time of recording. The speech recorded in the headphone microphone (H01) is more clean compared to other sensors. Accordingly, the speech recorded in D01 is considered as degraded speech and speech recorded in H01 is considered as clean speech.

Keeping the language as English and conversational style, three experiments are conducted on IITG MV database as follows:

1) *Clean and sensor matched*: Speech recorded over sensor H01 is used for training and testing.
2) *Clean trained and degraded test*: Speech recorded over H01 is used for training and speech recorded over D01 is used for testing.
3) *Degraded trained and clean test*: Speech recorded over sensor D01 is used for training and speech recorded over sensor H01 used for testing.

For this experimental setup six hours of UBM speech were selected from 17 male and 17 female speakers those who are not belonging to the present 100 speakers set. This six hours of speech contains three hours of male speech and three hours of female speech. For each speaker, the UBM speech is distributed equally among the two sensors H01 and D01. Using the sensor mixed data, two gender dependent 512 mixture size GMM are built, one for the male and other for the female speech. Finally, a 1024 mixture size gender independent UBM is built by pooling the two models and normalizing the weights. Two such gender independent sensor mixed UBMs are built one for the baseline SV system using the speech frames selected by energy based threshold and another for the proposed system using the knowledge of VOPs. During the time of model adaptation and testing, the respective UBM is used.

V. Results and Discussions

A. NIST-2003 speaker recognition database:

1) NIST-2003 original speaker recognition database: The detection estimation trade-off (DET) plots given in Fig. 4(a) shows that performance of the SV system using vowel-like regions and the baseline system in terms of equal error rate (EER) and is 7.75% and 7.3% , respectively. The relative improvement in EER for the SV system using vowel-like region is -6.16% compared to the baseline system. In NIST-2003 database except the channel effect there is almost no other type of degradation. The sensors used for collecting the speech are mainly electret sensors and same sensor is used for collecting training and testing speech for maximum number of speakers. The non-speech regions in this database are mostly silence regions. For such type of speech, the speech regions can be selected accurately without any difficulty. These speech regions contains vowel-like regions and non-vowel-like regions. For the telephonic speech, the vowel-like regions are less affected by channel compared to non-vowel-like frames due to their high SNR and low frequency. But, the performance of a GMM-UBM based speaker verification system not only depends on the quality of speech feature, but also on the number of feature vectors used for building the UBM, adapting the models and testing the system performance. The vowel-like frames selected by the proposed method is around half in number compared to the speech frames selected by baseline system. This result shows that for such type of speech, the SV system using vowel-like regions with less number of feature vectors gives comparable performance to the baseline system. This implies that the vowel-like regions contain most of the speaker information and slightly improved performance in baseline system may be due to the more number of features used for training and testing.

2) Noise degraded NIST-2003 test speech: For most of the practical implementation of SV system, the training speech can be collected in a clean environment, but at the time of verification, the users may access the system from a remote place.

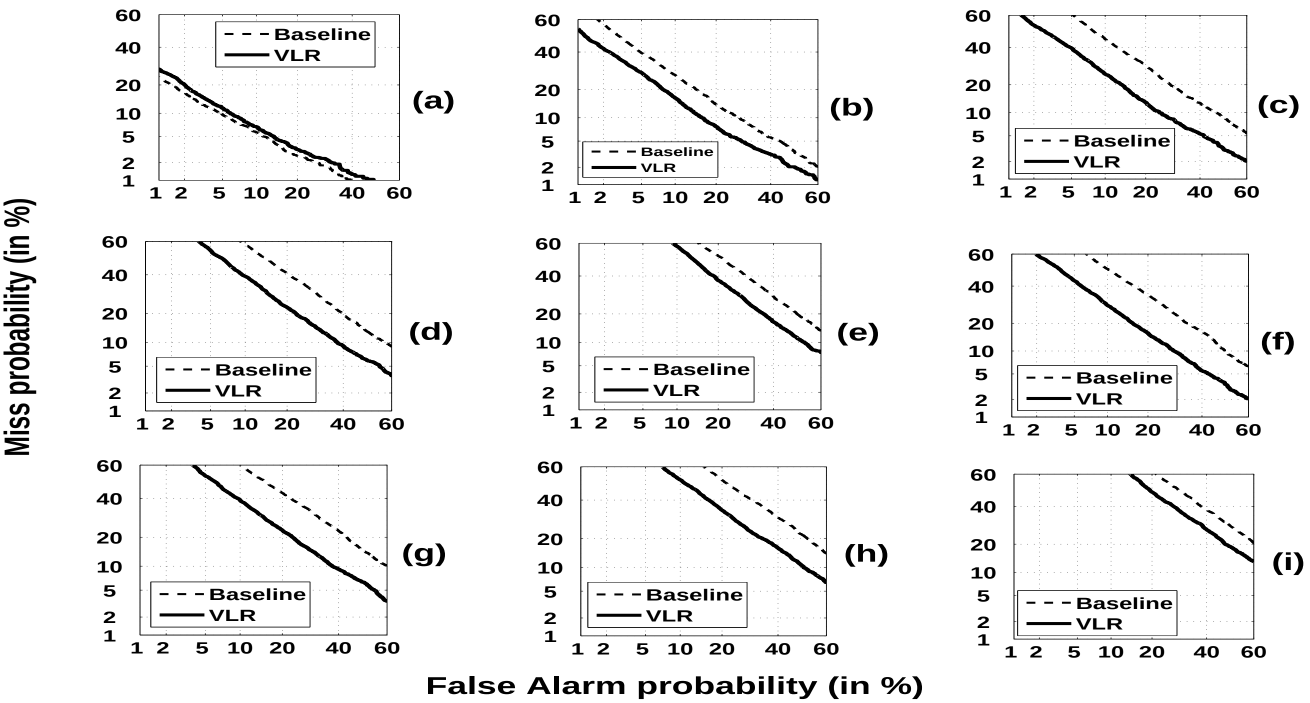

Fig. 4: DET curves showing performance for various experimental setup of NIST-20003 speaker recognition database. (a) Original NIST-2003 speaker recognition database, (b)-(e) factory-1 noise reconstructed test speech for SNR level 9 dB, 6 dB, 3dB and 0 dB, (f)-(i) white noise reconstructed test speech for SNR level 9 dB, 6dB, 3 dB and 0 dB.

This flexibility at the time of verification leads to a situation where the test speech may be degraded by the surrounding environment. To verify the performance of SV system using vowel-like regions for degraded test speech on a large population speaker recognition database, the test speech of the NIST-2003 speaker recognition database are reconstructed using factory-1 noise and white noise of NOISEX-92 database. For each noise, the noise level is scaled such that the overall SNR of noise reconstructed speech is maintained as 9 dB, 6 dB, 3dB and 0 dB, respectively. The DET plots in Fig. 4(b)-(e) shows performance of the SV system using vowel-like regions and the baseline system for factory-1 noise reconstructed test speech under 9 dB, 6 dB, 3 dB and 0 dB SNR, respectively. The EER for the SV system using vowel-like region is 12.73%, 16.21%, 21.27% and 27.91, respectively for 9 dB, 6 dB, 3 dB and 0 dB. The corresponding EER for baseline system is 17.3%, 24.12%, 29.58% and 34.1%, respectively. The relative improvement in EER for the SV using vowel-like regions compared to the baseline is 26.42%, 32.79%, 28.09% and 18.15% for 9db, 6dB, 3dB and 0dB SNR, respectively. The DET plots in the Fig. 4(f)-(i) shows performance of the SV using vowel-like region and the baseline system for white noise reconstructed test speech with SNR level of 9 dB, 6 dB, 3 dB and 0 dB, respectively. The EER for the SV system using vowel-like region is 17.75%, 24.49%, 26.33% and 33.06, respectively for 9 dB, 6 dB, 3 dB and 0 dB. The corresponding EER for baseline system is 26.92%, 30.85%, 35.05% and 38.94%, respectively. The relative improvement in EER for the SV using vowel-like regions compared to the baseline is 34.06%, 30.34%, 24.88% and 15.10% for 9db, 6dB, 3dB and 0dB SNR, respectively.

3) Noise degraded NIST-2003 trained speech: The second important application of SV system is the forensic use. In this type of application the person under check can talk from any environment depending on his own choice and this leads to a situation where the training speech may be degraded. To verify the performance of the proposed SV system for noise degraded train speech, the training speech of NIST-2003 is reconstructed using factory-1 noise and white noise of NOISEX-92 database. For this experiment the over all SNR level is kept at 3 dB in each case. The DET plots given in Fig. 5(a)-(b) shows the performance of vowel-like regions and baseline for two noise reconstructed NIST-2003 trained speech. The EER of the SV system using vowel-like region is 23.75% and 30.84% for factory-1 and white noise reconstructed speech, respectively. The corresponding EER for baseline SV are 28.32% and 35.5%. The relative improvement in EER for the SV system using vowel-like region compared to the baseline is 16.14% and 14.1% for factory-1 noise and white noise reconstructed train speech, respectively. This experiment shows that better speaker modeling is possible in degraded environments by selecting the vowel-like regions.

As discussed earlier, these set of results show that EER of the baseline system and the SV system using vowel-like regions increased differently from their clean (original NIST-2003 speech) performance depending on the type of noise and SNR. But, the relative increase in EER for the SV using vowel-like region is much less compared to the baseline SV system. This may be due to two different factors, (1) The vowel-like regions are selected with very less error for noise degraded speech. (2) The speaker information in vowel-like regions is relatively more robust to degradation compared to the nonvowel-like regions. This better selection of more speaker discriminative vowel-like regions reduced the mismatch between the training and testing speech compared to the baseline. Due this better modeling and reliable testing, the SV using vowel-like regions gives significantly improved performance compared to the baseline system for noise degraded speech.

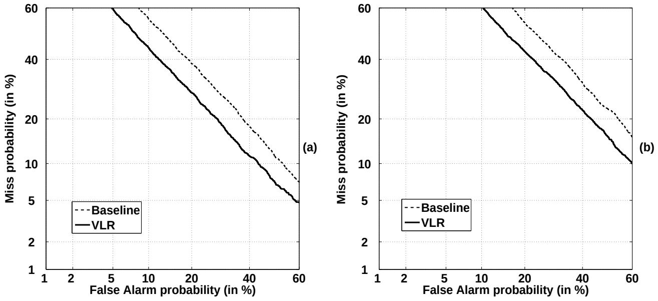

Fig. 5: DET curves showing performance for noise degraded NIST-2003 trained speech (a) Factory-1 noise reconstructed train speech for SNR level 3 dB, (b) white noise reconstructed trained speech for SNR level 3 dB.

B. IITG MV speaker recognition database

1) Clean and sensor matched: The clean speech of IITG MV speaker recognition database is collected using a headphone microphone mounted close to the speaker. The training and testing speech used for this experiment are wideband speech and collected through the same sensor. Therefore, the speech used for this experiment does not contain any degradation like noise, reverberation, channel and sensor variation. This is the most favoring condition for the baseline system. The DET plots in Fig. 6(a) shows that for clean and sensor matched condition, performance of the vowel-like regions and baseline system in terms of EER are 3.4% and 2.9%, respectively. The relative performance improvement in the SV using vowel-like region over the baseline system in terms of EER is -17.24%. Number of test files and speakers used for this experiment is less compared to NIST-2003 speaker recognition database, but the relative performance of the systems can be compared for two databases. As discussed earlier, the speech frames selected for the baseline system in clean speech experiments of IITG MV speaker recognition database and NIST-2003 speaker recognition database are perfect. The only degradation present in NIST-2003 database is the channel effect and sensor mismatch for some speakers. Due to this degradation the relative performance of the SV system using vowel-like region to the baseline system is 11.08% better compared to the clean experiment of IITG MV database. As discussed earlier, these two results indicate that vowel-like regions are less affected by channel and sensors compared to the non-vowel-like regions.

2) Clean trained and degraded test: This experiment is conducted to better investigate performance of the SV system using vowel-like region for real environment degraded speech. The degraded test speech recorded in sensor D01 contains noise and reverberation. This degradation varies differently within the same speech and for different speech files, depending on the recording environment. Further, for this experiment the training and testing speech is collected over different sensors. The DET plots in Fig. 6(b) shows performance of vowel-like regions and baseline system in terms of EER and is 14.1% and 20.2%, respectively. The relative performance improvement in the SV using vowel-like region over the baseline system in terms of EER is 30.2%. This results shows that for most of the practical uses a better speaker verification system can be developed using vowel-like regions.

3) Degraded trained and clean test: This experiment is conducted to verify the significance of vowel-like regions for modeling the speaker information in a more practical environment degraded speech. The DET plots in Fig. 6(c) shows performance of the SV system using vowel-like regions and the baseline system in terms of EER and is 15.3% and 19.8%, respectively. The relative performance improvement in the SV using vowel-like region over the baseline system in terms of EER is 22.73%. This result shows that even in severely degraded speech signal using the vowel-like regions better speaker modeling is possible. These results infer that, if data is not a constraint, a better speaker verification system can be developed using vowel-like regions.

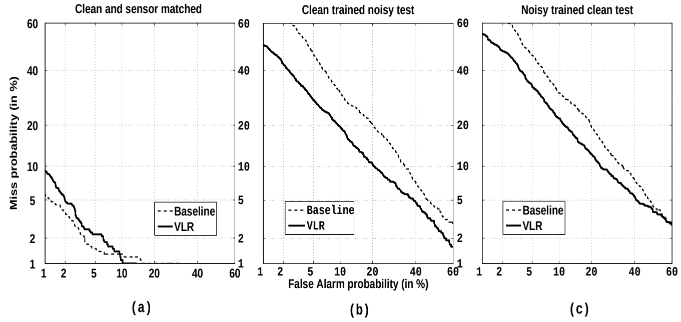

Fig. 6: DET curves showing performance for various experimental setup of IITG MV speaker recognition database. (a) Clean and sensor matched, (b) clean trained noisy test, (c) noisy trained and clean test.

VI. Summary and Conclusions

In this work we proposed a new VLROP detection method for clean and degraded speech by utilizing the advantages of Hilbert envelope of linear prediction residual and zero frequency filtered output. The performance of proposed method is evaluated using 60 speaker subset of TIMIT database for clean as well as noise degraded speech. Using the knowledge of VLROPs, the 100 ms regions right to each VLROP are identified as vowel-like regions. A speaker verification system is developed using MFCC as the speaker feature and GMM-UBM as modeling technique. The environmental effect is compensated in the cepstral domain using CMS followed by CVN. The performance of the proposed speaker verification system is evaluated on NIST-2003 speaker recognition database. In the second level, two different noises are taken from NOISEX-92 database to create noise degraded NIST-2003 speech. Performance of the proposed system is evaluated for different degraded conditions on noise reconstructed NIST-2003 speaker recognition database. Finally, performance of the SV system is evaluated on the IITG MV speaker recognition database for clean and real degraded speech.

This work shows that for clean speech, with less number of vowel-like frames, the proposed SV system gives slightly poor performance compared to the baseline system. Alternatively, for degraded conditions, the proposed system provides significantly improved performance. In the practical scenarios a robust speaker verification system can be developed by selecting the vowel-like regions. The future work may focus on applying different compensation techniques to suppress the degradation further in the detected vowel-like regions.

REFERENCES

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

[2] J. Ming, T. J. Hazen, J. R. Glass, and D. A. Reynolds, "Robust speaker recognition in noisy conditions," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 15, no. 5, pp. 1711–1723, July 2007.

[3] P. Krishnamoorthy and S. R. M. Prasanna, "Enhancement of noisy speech by temporal and spectral processing," *Speech Communication (in press)*, 2010.

[4] B. Yegnanarayana, C. Avendano, H. Hermansky, and P. S. Murthy, "Speech enhancement using linear prediction residual," *Speech Communication*, vol. 28, pp. 25–42, May 1999.

[5] B. Yegnanarayana and P. S. Murthy, "Enhancement of reverberant speech using lp residual signal," *IEEE Trans. Speech Audio Processing*, vol. 8, pp. 267–281, May 2000.

[6] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-27, pp. 113–120, April 1979.

[7] P. Krishnamoorthy and S. R. M. Prasanna, "Reverberant speech enhancement by temporal and spectral processing," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 17, no. 2, pp. 253–266, February 2009.

[8] H. K. Kim and R. C. Rose, "Cepstrum-domain acoustic feature compensation based on decomposition of speech and noise for ASR in noisy environments," *IEEE Trans. on Speech and Audio Processing*, vol. 11, no. 5, pp. 435–446, September 2003.

[9] H. Hermansky, N. Morgan, A. Bayya, and P. Kohn, "RASTA-PLP speech analysis technique," in *ICASSP*, march 1992, pp. I.121–I.124.

[10] D. A. Reynolds, "The effect of handset variability on speaker recognition performance: experiments on the switchboard corpus," in *ICASSP*, vol. 1, Atlanta, Ga, USA, 1996, pp. 113–116.

[11] R. Auckenthaler, M. Carey, and H. L. Thomas, "Score normalization for text-independent speaker verification systems," *Digital Signal Processing*, vol. 10, no. 1, pp. 42–54, January 2000.

[12] J. L. Gauvain and C. H. Lee, "Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains," *IEEE Trans. on Speech and Audio Processing*, vol. 2, no. 2, pp. 291–298, April 1994.

[13] M. Gales, D. Pye, and P. C. Woodland, "Variance compensation within the mllr framework for robust speech recognition and speaker adaptation," in *ICSCP*, vol. 3, October 1996, pp. 1832–1835.

[14] X. Zhang, H. Wang, X. Xiao, J. Zhang, and Y. Yan, "Maximum a posteriori linear regression for speaker recognition," in *ICASSP*, March 2010, pp. 4542–4545.

[15] A. N. Khan and B.Yegnanarayana, "Vowel onset point based variable frame rate analysis for speech recognition," in *Proceedings of ICISIP*, 2005, pp. 392–394.

[16] S. R. M. Prasanna, B. V. S. Reddy, and P. Krishnamoorthy, "Vowel onset point detection using source, spectral peaks, and modulation spectrum energies," *IEEE Trans. audio, speech, and lanuage processing*, vol. 17, no. 4, pp. 556–565, May 2009.

[17] K. N. Stevens, *Acoustic Phonetics*. The MIT Press Cambridge, Massachusetts, London, England, 2000.

[18] D. J. Hermes, "Vowel onset detection," *J. Acoust. Soc. Amer.*, vol. 87, pp. 866–873, 1990.

[19] J. Wang, C. Hu, S. Hung, and J. Lee, "A heirarchical neural network based c/v segmentation algorithm for mandarin speech recognition," *IEEE Trans. Signal Processing*, vol. 39, no. 9, pp. 2141–2146, September 1991.

[20] J. Y. S. R. K. Rao, C. C. Sekhar, and B. Yegnanarayana, "Neural network based appraoch for detection of vowel onset points," in *Int. Conf. Adv. Pattern Recognition Digital Tech.*, vol. 1, December 1999, pp. 316–320.

[21] S. R. M. Prasanna and B. Yegnanarayana, "Detection of vowel onset point events using excitation source information," in *INTERSPEECH*, September 2005, pp. 1133–1136.

[22] ——, "Extraction of pitch in adverse conditions," in *ICASSP*, vol. 1, August 2004, pp. 109–112.

[23] K. S. R. Murthy and B. Yegnanarayana, "Epoch extraction from speech signals," *IEEE Trans. on Audio,Speech, And Language Processing*, vol. 16, pp. 1602–1613, November 2008.

[24] 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, June 2009.

[25] J. Makhoul, "Linear prediction:a tutorial review," *Proc. IEEE*, vol. 63, no. 04, pp. 561–580, April 1975.

[26] B. Yegnanarayana, S. R. M. Prasanna, J. M. Zachariah, and S. Gupta, "Combining evidence from source, suprasegmental and spectral features for a fixed text speaker verification system," *IEEE Trans. speech and audio Processing*, vol. 13, no. 4, pp. 575 – 582, July 2005.

[27] *"TIMIT acoustic-phonetic continuous speech corpus, " NTIS Order PB91-505065, National Institute of Standards and Technology, Gaithersburg, Md, USA, 1990, Speech Disc 1-1.1.*

[28] P. A. Naylor, A. Kounoudes, J. Gudnason, and M. Brookes, "Estimation of glottal closure instants in voiced speech using the DYPSA algorithm," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 34–43, January 2007.

[29] A. Varga and H. J. M. Steeneken, "Assessment for automatic speech recognition: II. noisex-92: A database and an experiment to study the effct of additive noise on speech recognition systems," *Speech Communication*, vol. 12, no. 3, pp. 247–251, July 1993.

[30] B. C. Haris, G. Pradhan, A. Misra, S. Shukla, R. Sinha, and S. R. M. Prasanna, "Multi-variability speech database for robust speaker recognition," in *National conf. on communication (NCC) (Under Review)*, 2011.

[31] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans on Acoust., Speech and Signal Processing*, vol. ASSP-28, no. 4, pp. 357–366, August 1980.

[32] F. K. Soong and A. E. Rosenberg, "On the use of instantaneous and transitional spectral information in speaker recognition," *IEEE Trans. on Acoustics Speech and Signal Processing*, vol. 36, no. 6, pp. 871–879, June 1988.

[33] D. A. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models," *Speech Communication*, vol. 17, pp. 91–108, March 1995.

[34] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital Signal Processing*, vol. 10, pp. 19–41, January 2000.

[35] *Linguistic Data Consortium,"Switchboard cellular part 2 audio," in http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2004S07 ,2004.*