Significance of Vowel-Like Regions for Speaker Verification Under Degraded Conditions

S. R. Mahadeva Prasanna, *Member, IEEE*, and Gayadhar Pradhan

***Abstract*—Vowel-like regions (VLRs) in speech includes vowels, semi-vowels, and diphthong sound units. VLR can be identified using a 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 a 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 paper, VLRs are detected using the knowledge of VLROPs during training and testing. Features from the VLRs are then used for training and testing the speaker models. As a result, significant improvement in the performance is reported for speaker verification under degraded conditions.**

***Index Terms*—Degraded condition, speaker information, speaker verification (SV), vowel-like region (VLR), vowel-like region onset point.**

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 speaker model synthesis (SMS) [12], parallel model combinations (PMC) [13], and microphone arrays [14] are used for adapting model parameters to the testing environment. Speaker modeling methods like maximum *a posteriori* (MAP) [15], maximum-likelihood linear regression (MLLR) [16], and Bayesian maximum *a posteriori* linear regression (MAPLR) [17] may also be used for adapting model parameters to the testing environment. Such type of techniques require *a priori* information about the testing environment and may not be possible under all practical scenarios.

On top of all these approaches, the performance of the SV 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 (VLR) onset points (VLROPs). VLROP helps in identifying VLRs 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 the production perspective for speech enhancement [3]–[5], [7].

VLROP is defined as the instant at which the onset of VLR takes place. VLROP corresponds to vowel onset point (VOP) in case of vowels [18], 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 of missing and spurious VOPs, and also the resolution with which VOPs are detected [19]. The majority of the errors are observed to be due to the cases of semivowels and diphthongs [19]. However, for the SV task we need vowel,

Manuscript received November 01, 2010; revised February 06, 2011, April 15, 2011; accepted April 25, 2011. Date of publication May 16, 2011; date of current version September 30, 2011. This work was supported by the project titled *"Development of Person Authentication System based on Speaker Verification in Uncontrolled Environment"*, funded by the Department of Information Technology (DIT), New Delhi, India. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Engin Erzin.

The authors are with the Department of Electronics and Electrical Engineering, Indian Institute of Technology Guwahati, Guwahati-781039, Assam, India (e-mail: prasanna@iitg.ernet.in; gayadhar@iitg.ernet.in).

Digital Object Identifier 10.1109/TASL.2011.2155061

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 the VLROP event instead of VOP. With the help of VLROP event, the VLRs 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 [20]. VLRs 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. VLRs are produced by exciting the vocal tract system using impulse-like excitation due to the sudden closure of vocal folds. Due to the 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. VLRs 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 VLRs and use speaker information from such regions, then better speaker modeling as well as more reliable testing may be possible. This may help in increasing the robustness of SV systems under degraded conditions.

In the existing SV 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, VLRs are separated out from the non-VLRs based on the knowledge of VLROP, and features from the VLRs are used for modeling and testing. Suppose if the 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 VLRs. Alternatively, the merit of VLRs may be found under degraded conditions. 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 VLRs 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. The proposed speaker verification system using VLRs 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 VLROPs and VLRS in Degraded Speech Using Excitation Source Information

In the present work, the VLROP refers to the instant at which the onset of VLR takes place. The VLRs are prominent regions in the speech signal due to high amplitude, periodicity, long duration, and lower zero crossing rate. Considering these distinct properties of VLRs 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 [21], zero-crossing rate, energy and pitch information [22], training neural network with the trends in energy, zero crossing rate and spectral flatness at the VOP [23], and using excitation source information [24]. More recently, a combined method using the excitation source, spectral peaks and modulation spectrum information is proposed for the detection of VOP [19]. In all these methods, the failing cases are reported mostly 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 semivowels 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 the performance for the case of VLROP detection.

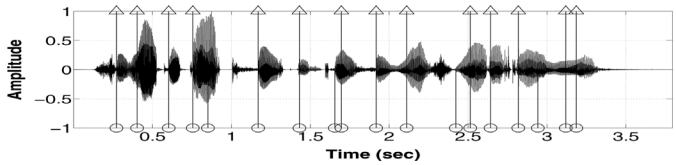

Fig. 1. Speech signal of the text *she had your dark suit in greasy wash water all year* taken from TIMIT database with reference VOPs (arrows) and reference VLROPs (circles).

Fig. 1 shows a speech signal of the text *she had your dark suit in greasy wash water all year* taken from TIMIT database. The figure also gives the markings of VOPs and VLROPs taken from the manual labeling available. The number of VLROPs are 18, more compared to VOPs which are only 14. Hence, more regions compared to the case of VOPs alone. Further, these regions are identified based on some speech production knowledge and hence their detection is robust. All the VOP detection methods mentioned above are evaluated for clean and wideband 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 is 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 [25]. Also a VOP detection method is exclusively developed using this information in [24]. 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 [26]. Since, the zero-frequency resonator exploits only signal energy around the zero-frequency region and attenuates all other information [26], the resonator output signal may provide robustness to various degradations. The strength of excitation derived from ZFF signals (ZFFS) has been demonstrated earlier to have better discriminating ability at the unvoiced–voiced transitions [27]. It is therefore proposed that by combining the features derived from the HE of the LP residual with features from ZFFS, 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 Using 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, tenth-order LP analysis is performed to estimate the linear prediction coefficients (LPCs) [28]. 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 [24].

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

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

where

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

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

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

The HE of LP residual shows instantaneous variations in the residual signal and for VLROP detection we need to preserve only variations at pitch period level. Therefore, to construct a smoothed excitation contour, the maximum value of the HE of LP residual for every 5-ms block with one sample shift is noted. 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) [19], [24]. This convolved output is termed as *VLROP evidence from HE of LP residual*. Fig. 2 shows a portion of speech signal shown in Fig. 1, its LP residual, HE of LP residual, smoothed excitation contour and VLROP evidence from HE of LP residual. The smoothed excitation contour considers the envelope of the HE of LP residual as shown in Fig. 2(d) and sufficient for finding the VLROP evidence.

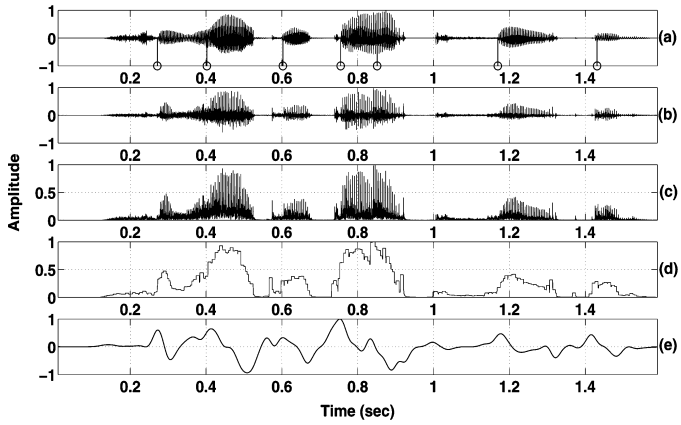

Fig. 2. Steps involved in VLROP evidence using HE of LP residual. (a) A portion of speech signal of the text *she had your dark suit in greasy wash water all year* taken from TIMIT database with reference VLROPs (circles). (b) LP residual. (c) HE of LP residual. (d) Smoothed excitation contour. (e) VLROP evidence using 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 [26]. In the 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) [26]. 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 [26]:

- 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 $\bar{y}(n) = 1/(2N+1)\sum_{n=-N}^{N} y(n)$ 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.

Fig. 3 illustrates the various steps involved in zero-frequency filtering. The speech signal is passed through the zero-frequency resonator and the trend is removed to obtain the ZFFS shown in Fig. 3(b). The zero crossings give the locations of epochs and are shown in Fig. 3(c). The slope around the zero crossings are

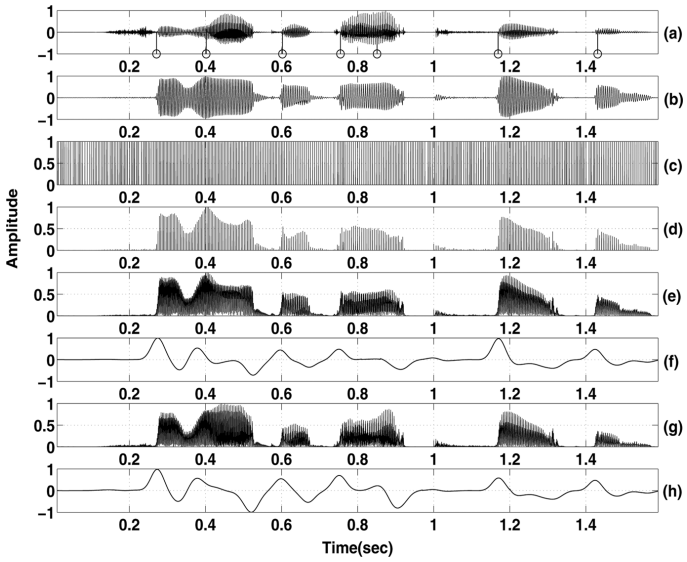

Fig. 3. Steps involved in zero-frequency filtering. (a) A portion of speech signal of the text *she had your dark suit in greasy wash water all year* taken from TIMIT database with reference VLROPs (circles). (b) Zero-frequency filtered signal (ZFFS). (c) Epoch locations. (d) Strength of excitation. (e) Absolute value of first-order difference of ZFFS. (f) VLROP evidences using first-order difference of ZFFS. (g) Absolute value of second-order difference of ZFFS. (h) VLROP evidences using second-order difference of ZFFS.

proposed earlier as strength information [27]. The epochs with their strength are given in Fig. 3(d). Thus, first-order difference of ZFFS [Fig. 3(b)] given in Fig. 3(e) contains strength of excitation information. The second-order difference of ZFFS therefore contains change in the strength of excitation and is given in Fig. 3(g). The main event is the change in strength of excitation at VLROP. Hence, second-order difference of ZFFS is hypothesized to give improved and robust detection of VLROP. The change in the excitation characteristics at the VLROP 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*. Figs. 3(f) and (h) show VLROP evidences using first-order difference of ZFFS and second-order difference of ZFFS, respectively. The VLROP evidence using second-order difference of ZFFS better discriminates the VLROP compared to the evidence using the first-order difference of ZFFS (refer to region around 0.8 s).

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* [29]. 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 second 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 samples 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 [30]:

- *Identification rate (IR):* The percentage of reference VLROPs that are matched by the detected VLROPs within the VLRs;
- *Miss rate (MR):* The percentage of reference VLROPs for which no VLROPs detected within the VLRs;
- *Spurious rate (SR):* The percentage of detected VLROPs, which are detected outside the VLRs;
- *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.

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

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 [19] are also given in the same table. 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 the 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 [19].

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 [31]. 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 the spurious rate in case of recent VOP detection method and HE

TABLE II
PERFORMANCE OF VLROP DETECTION METHODS USING EXCITATION SOURCE INFORMATION AND BASED ON RECENT VOP DETECTION METHOD FOR NOISE RECONSTRUCTED 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

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, the performance of SV system is expected to be better using VLRs.*

D. Detection of VLRs From Degraded Speech

Fig. 4(b) shows the VLROP evidence for a segment of speech taken from NIST-2003 speaker recognition database [32]. The corresponding evidences for *white noise* degraded speech are shown in the Fig. 4(c)–(f) for 9 dB, 6 dB, 3 dB, and 0 dB SNR, respectively. The arrow marks in each evidence correspond to the hypothesized VLROP locations, obtained by the proposed method. Fig. 4(a) and (b) shows that the hypothesized VLROPs nearly correspond to the starting of VLRs. Fig. 4(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 VLRs 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. 5. The VLRs are selected by considering 100-ms regions 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 VLR to be about 100 ms in continuous speech. The speech regions are identified as the speech frames above the energy threshold ($0.06 \times average\ energy$), which are represented in dotted lines. In case of clean speech shown in Fig. 5(a), 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 VLRs, although the VLROP detection is perfect, but all the selected regions are mostly VLRs. Fig. 5(b)–(e) shows that for noise degraded speech also same VLRs are selected, even for severely degraded speech. Hence, the robustness of detection of VLRs using VLROP. The speech region selection by energy based threshold is affected for 9-dB SNR and completely fails for 6 dB, 3 dB, and 0 dB SNR. From the experiments, it is observed that by using 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 using a very high threshold eliminates most of the speech frames for clean speech and using a low threshold selects most of the noise frames. Further, if the noise appears randomly within a particular speech signal, neither low nor high threshold will select the proper speech frames.

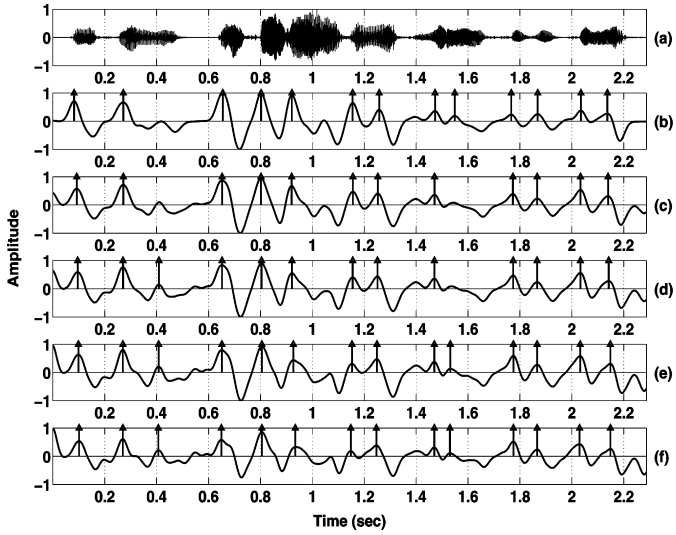

Fig. 4. VLROP evidences for degraded speech. (a) Segment of speech taken from NIST-2003 speaker recognition database. (b) VLROP evidence for clean speech. (c)–(f) VLROP evidences for *white noise* degraded speech with overall SNR level 9 dB, 6 dB, 3 dB, and 0 dB, respectively.

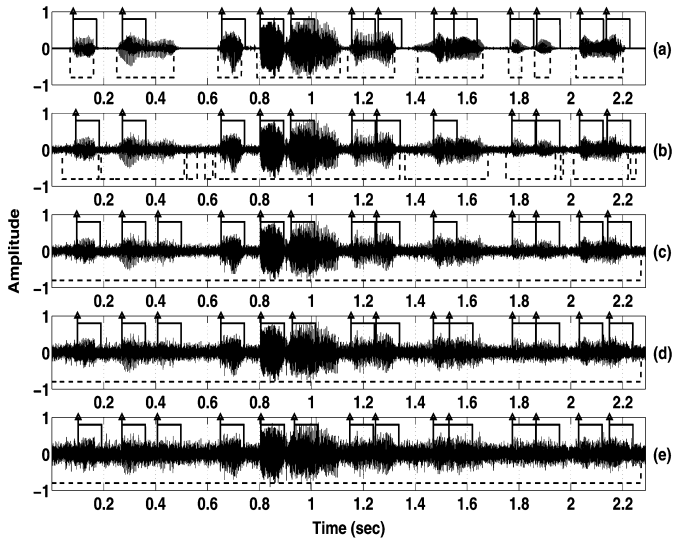

Fig. 5. VLRs (solid lines) and speech regions (dotted lines) detection in degraded condition. (a) Segment of speech taken from NIST-2003 speaker recognition database. (b)–(e) noise degraded speech with overall SNR level 9 dB, 6 dB, 3 dB, and 0 dB, respectively.

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 [33]. The speech file shown in Fig. 6(a) is a segment of speech recorded in sensor H01 (clean speech) and the same segment

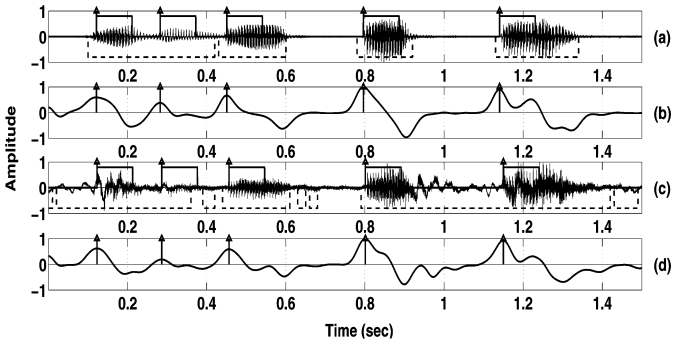

Fig. 6. VLRs (solid lines) and speech regions (dotted lines) detection for clean and degraded speech of IITG MV speaker recognition database. (a) Clean speech (speech record over sensor H01). (b) VLROP evidence for clean speech. (c) Degraded speech (speech recorded in parallel over sensor D01). (d) VLROP evidence for degraded speech.

speech recorded over sensor D01 (degraded speech) is shown in Fig. 6(c). The VLROP evidences corresponding to clean and degraded speech are shown in Fig. 6(b) and (d), respectively. The Fig. 6 indicates that the VLRs 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 VLRs can be selected from degraded speech using VLROP in a robust manner. Therefore, using these relatively less degradation affected and more speaker discriminating regions, a better SV system can be developed under degraded conditions.

III. Speaker Verification Using VLRs

A. Database

The performance of proposed SV system is evaluated on complete NIST-2003 speaker recognition database at the first level to study the discriminating speaker information present in the VLRs. The study includes development of SV system using VLRs, only vowel regions and speech regions based on energy threshold. Then the *factory-1 noise* and *white noise* are considered from NOISEX-92 database to create the noise degraded speech. Usually for the noisy speech the SNR level around 3 dB is mostly considered. Hence, the energy level of the noise is scaled such that the overall SNR of the noise degraded NIST-2003 speech is maintained at 9, 6, 3, and 0 dB (multiples of 3). Performance of the 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 then evaluated on noise degraded train speech and original test speech for 9, 6, 3, and 0-dB SNR level. The performance is evaluated for both noise degraded train and test speech for 9, 6, 3, and 0 dB.

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, inbuilt 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 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 44.1 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 VLRs

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 VLRs. In case of speaker verification using VLRs, features derived only from these regions are used for training and testing. In case of only vowel regions 80% highest evidence VLRs are used. Finally, 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 (MFCCs) are calculated using 22 logarithmically spaced filter banks [34]. The first 13 coefficients excluding zeroth coefficient value are used as a feature vector. Delta (Δ) and delta-delta $(\Delta\Delta)$ of MFCCs are computed using two preceding and two succeeding feature vectors from the current feature vector [35]. Thus, the feature vector will be of 39 dimensions with 13 MFCCs, 13Δ MFCCs, and 13$\Delta\Delta$ MFCCs.

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 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 [36]. 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. For all the four 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 Modeling and Testing

The main motivation of this work is to study the discriminative information present in the VLRs for speaker modeling and testing in degraded environments. Except for deriving frames from VLRs, there is no difference in the steps of SV system development. Hence, the extensively used Gaussian mixture model (GMM)-UBM-based speaker modeling is employed [10]. 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 $U = \{\mu_c, \Sigma_c, \eta_c\}, c = 1, \ldots C$, 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 [10]. During the testing stage, the log likelihood scores are calculated from both the claimed model and UBM.

For a GMM-UBM SV system the score normalization technique such as test score normalization (T-norm) provides performance improvement [11], [37]. Hence, in the present work the T-norm is employed as the score normalization technique [11].

F. Baseline SV System

In order to compare the performance obtained using VLRs, we have developed SV system based on energy threshold $(0.06 \times average\ energy)$ which is termed as *baseline system*. The energy threshold is based on several SV experiments with different thresholds and using the one that gives best performance. 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 VLRs. Further to compare the performance of VLROPs and VOPs, 80% of highest evidence VLROPs are used as VOPs and the SV system is developed.

G. Performance Comparison

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

$$\mathrm{EER}_R = \frac{(\mathrm{EER}_B - \mathrm{EER}_V)}{\mathrm{EER}_B} \times 100\% \tag{10}$$

where EER_R, EER_B, and EER_V are the relative improvement in EER, EER of the baseline SV system, and the EER of SV system using the VLRs, respectively.

IV. Experimental Studies

In the present experimental setup, the following four experiments are conducted on NIST-2003 speaker recognition database:

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 train speech*: NIST-2003 noise degraded train speech is used for training the models and original test speech is used for testing;
4) *Noise degraded NIST-2003 train and test speech*: NIST-2003 noise degraded train speech is used for training the models and noise degraded test speech is used for testing.

For these four sets of experiment, 30 hours of UBM speech was selected from randomly selected 250 male and 250 female speech of switchboard cellular part 2 Audio database [38]. Using each gender speech, two gender-dependent 512 mixture size GMMs 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 [10]. Three such gender-independent UBMs are built, first one for the baseline SV system using the speech frames selected by energy based threshold, second for only vowel regions and third for the proposed system using the VLRs. During the time of model adaptation and testing, the respective UBM is used. The SV system using only vowel regions is used for initial study for comparison with VLRs. Later studies will be with respect to VLRs and speech regions based on energy based threshold. The T-norm is applied by using a set of 100 speakers (50 males and 50 females) randomly selected from NIST-2002 speaker recognition database. During the noise degraded experiments, T-norm speech is maintained at same noise and SNR level as the training speech. For the speaker verification using VLRs the T-norm models are built using the VLRs. In case of baseline system the models are built using the speech frames selected by energy based threshold.

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, ten speech segments between 30–45 s 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 recorder (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, four 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;

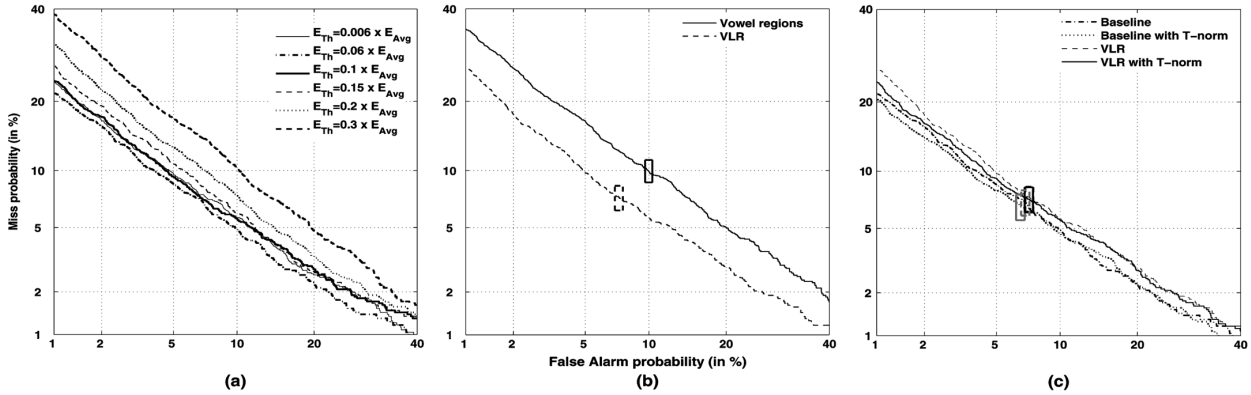

Fig. 7. DET curves showing performance for various experimental setup of NIST-20003 speaker recognition database. (a) Effect of energy threshold on speaker verification performance. (b) Performance using vowel regions and VLRs. (c) Performance of baseline system and SV system using VLRs. The boxes indicate the 95% confidence intervals at EER operating points.

TABLE III
SUMMARY OF SV PERFORMANCE FOR VARIOUS EXPERIMENTAL SETUP OF NIST-20003 SPEAKER RECOGNITION DATABASE WITHOUT (W/O) AND WITH T-NORM

Score normalization	Equal error rate (%)								
	Energy threshold (E_{Avg})						Baseline	Vowel	
	0.006	0.06	0.1	0.15	0.2	0.3	$0.06 \times E_{Avg}$	regions	VLR
w/o T-norm	7.12	6.91	7.04	7.6	8.67	10.07	6.91	9.89	7.28
T-norm	-	6.54	-	-	-	-	6.54	-	7.14

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

For this experimental setup, six hours of UBM speech was 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 GMMs are built, one for the male and the 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 VLROPs. During the time of model adaptation and testing, the respective UBM is used. Due to unavailability of same type of data for these set of experiments, the T-norm is applied using the registered speakers excluding the speakers which will be tested against the current test segment.

V. RESULTS AND DISCUSSIONS

A. NIST-2003 Speaker Recognition Database

1) NIST-2003 Original Speaker Recognition Database: At the first level, to select the optimal value of energy threshold for the baseline SV system, the SV performance is evaluated for six different thresholds starting from a very low threshold value ($0.006\times$ *average energy* ($E_{\rm Avg}$)) to a comparable higher threshold ($0.3\times E_{\rm Avg}$). The detection estimation tradeoff (DET) plots given in Fig. 7(a) shows the performance of the SV system in terms of equal error rate (EER) for threshold values $0.006 \times E_{\rm Avg}$, $0.06 \times E_{\rm Avg}$, $0.1 \times E_{\rm Avg}$, $0.15 \times E_{\rm Avg}$, $0.2 \times E_{\rm Avg}$, and $0.3 \times E_{\rm Avg}$ is 7.12%, 6.91%, 7.04%, 7.6%, 8.67%, and 10.07%, respectively. These results show that performance of SV system highly depends on the value of energy threshold used for selecting the speech frames. As discussed earlier use of higher threshold eliminates maximum portion of the speech regions and as a result, the SV performance reduces. For the present experimental setup, the best performing threshold value ($0.06 \times E_{\rm Avg}$) is fixed as the energy threshold for the baseline system for further experiments.

The DET plots in Fig. 7(b) shows performance of the SV system using only vowel regions and VLRs in terms of EER, and is 9.89% and 7.28%, respectively. Boxes on the DET curves indicate the 95% confidence interval at the EER operating points [39]. The 95% confidence intervals at EER operating points do not overlap. The relative improvement in EER for the SV system using VLRs is 35.85% compared to the SV system using only vowel regions. Thus, VLROPs are preferable over VOPs. The DET plots in Fig. 7(c) shows performance of the SV system for the baseline system and SV system using VLRs. The performance of baseline system and SV system using VLRs are given in the Table III. From the table it can be observed that the EER of baseline system and the SV system using VLRs are reduced to 6.54% and 7.14% with the application of T-norm. For the same database and similar complexity system, the EER of baseline system is significantly better compared to systems reported in literature [40], [41].

It is also observed that the T-norm provides more improvement to the baseline compared to the SV system using VLRs. The additional score normalization like T-norm is generally

TABLE IV
NUMBER OF FRAMES USED FOR TRAINING AND TESTING IN BASELINE SYSTEM AND SV SYSTEM USING VLRS

Data set	Baseline			VLR		
	Avg.	Min	Max	Avg.	Min	Max
NIST-2003 Train	6070	2085	8151	3091	844	4532
NIST-2003 Test	1621	144	2984	836	72	1705

used to reduce channel, handset and other degradations effect on the verification scores. As discussed earlier, the VLRs are less affected by various degradations compared to the non-VLRs. Therefore, the verification scores obtained from the VLRs are relatively less affected by such degradations. Hence, the additional score normalization provides relatively less performance improvement to the SV system using VLRs compared to the baseline.

The relative improvement in EER for the SV system using VLRs is -9.17% compared to the baseline system. This is expected since 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 contain VLRs and non-VLRs. For the telephonic speech, the VLRs 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 SV 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 average number of frames used for training and testing of baseline system and SV system using VLRs are given in the Table IV. The table also contains the minimum and maximum number of frames used for training and testing. The average number of frames used for training and testing of baseline system is around two times more than VLRs. From Fig. 7(c) it can be seen that the 95% confidence intervals at EER operating point of the baseline system overlaps with that of the SV system using VLRs. This result shows that for such type of speech, the SV system using VLRs with nearly half the number of feature vectors gives comparable performance to the baseline system. This implies that the VLRs contain most of the speaker information and slightly improved performance in baseline system may be due to the significantly 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. 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 VLRs 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, 3 dB, and 0 dB, respectively. The DET plots in Fig. 8(a)–(d) shows performance of the SV system using VLRs and the baseline system for *factory-1 noise* reconstructed test speech under 9 dB, 6 dB, 3 dB, and 0 dB SNR, respectively. The DET plots in Fig. 8(e)–(h) shows performance of the SV using VLRs 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 performance of baseline system and SV system using VLRs are given in the Table V. From the table it can be observed that the VLRs provide significantly better performance compared to the baseline system. For the *factory-1 noise* degraded speech the relative improvement in EER for the SV using VLRs compared to the baseline is 26.32%, 35.56%, 30.25% and 17.36% for 9 dB, 6 dB, 3 dB, and 0 dB SNR, respectively. Similarly, for *white noise* degraded speech the relative improvement in EER for the SV using VLRs compared to the baseline is 34.91%, 29.9%, 24.9% and 16.86% for 9 dB, 6 dB, 3 dB, and 0 dB SNR, respectively. This set of experiment shows that without the knowledge of training and testing environment condition T-norm cannot help to improve the verification performance. This type of situation is expected in most of the practical application of SV system. In such a situation, a better SV system can be built using the VLRs.

3) Noise Degraded NIST-2003 Train 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 overall SNR level is kept at 9 dB, 6 dB, 3 dB, and 0 dB in each case. The DET plots given in Fig. 9(a)–(d) shows the performance of SV system using VLRs and baseline for *factory-1 noise* reconstructed train speech under 9 dB, 6 dB, 3 dB, and 0 dB SNR, respectively. The DET plots in Fig. 9(e)–(h) shows performance of the SV using VLRs and the baseline system for *white noise* reconstructed train speech with SNR level of 9 dB, 6 dB, 3 dB, and 0 dB, respectively. The performance of baseline system and SV system using VLRs are given in the Table VI. For the *factory-1 noise* degraded speech the relative improvement in EER for the SV using VLRs compared to the baseline is 17.88%, 10.50%, 12.73% and 14.46% for 9 dB, 6 dB, 3 dB, and 0 dB SNR, respectively. Similarly, for *white noise* degraded speech the relative improvement in EER for the SV using VLRs compared to the baseline is 30%, 24.07%, 13.66% and 8.88% for 9 dB, 6 dB, 3 dB, and 0 dB SNR, respectively. This experiment shows that better speaker modeling is possible in degraded environments by selecting the VLRs.

4) Noise Degraded Train and Test Speech: This set of experiments is conducted by assuming the situation where the type of noise and SNR level are known *a priori*. In such type of situation the mismatch between the training and testing speech can be reduced to some extent by corrupting the training speech to satisfy the testing condition. For this set of experiments, the

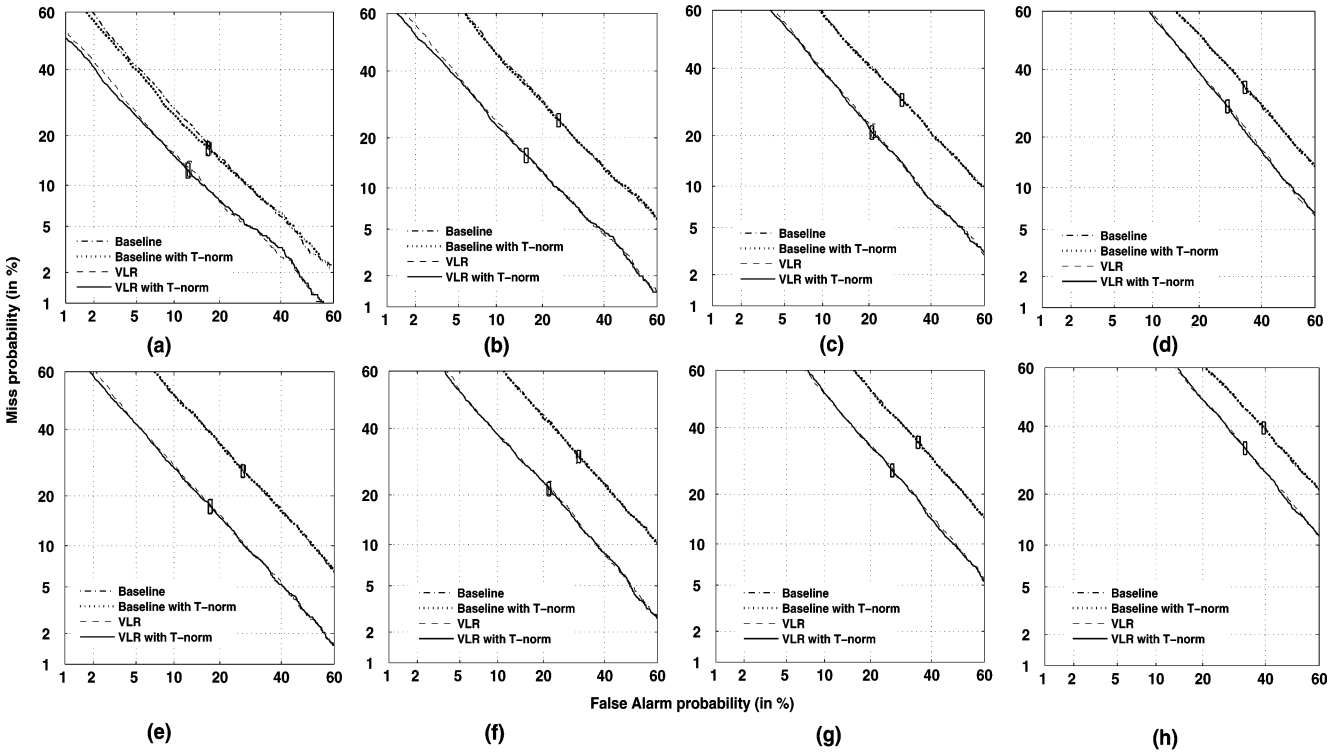

Fig. 8. DET curves showing performance for noise degraded NIST-2003 test speech. (a)–(d) *factory-1 noise* reconstructed test speech for SNR level 9 dB, 6 dB, 3 dB, and 0 dB, (e)–(h) *white noise* reconstructed test speech for SNR level 9 dB, 6 dB, 3 dB, and 0 dB. The boxes indicate the 95% confidence intervals at EER operating points.

training and testing speech of NIST-2003 database is reconstructed using *factory-1 noise* and *white noise* of NOISEX-92 database and the overall SNR level is kept at 9 dB, 6 dB, 3 dB, and 0 dB in each case. For each case, the training speech, test speech and T-norm speech are reconstructed using same noise and SNR level. The DET plots given in Fig. 10(a)–(d) shows the performance of SV system using VLRs and baseline for *factory-1 noise* reconstructed train and test speech under 9 dB, 6 dB, 3 dB, and 0 dB SNR, respectively. The DET plots in Fig. 10(e)–(h) shows performance of the SV using VLRs and the baseline system for *white noise* reconstructed train speech with SNR level of 9 dB, 6 dB, 3 dB, and 0 dB, respectively. The performance of baseline system and SV system using VLRs are given in the Table VII. For the *factory-1 noise* degraded speech the relative improvement in EER for the SV using VLRs compared to the baseline is -2.31%, -3.63%, -2.47%, and -6.21% for 9 dB, 6 dB, 3 dB, and 0 dB SNR, respectively. Similarly, for *white noise* degraded speech the relative improvement in EER for the SV using VLRs compared to the baseline is -2.13%, -6.51%, -8.08%, and -14.01% for 9 dB, 6 dB, 3 dB, and 0 dB SNR, respectively. The good performance in case of baseline system is due to the matching in the noise condition during training and testing. The slight degradation in performance for VLRs compared to baseline may be due to the less number of vowel-like frames used during training and testing. Since this is noise matching condition, the main merit of VLRs is providing nearly same performance with significantly less number of frames.

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

Finally, the gender-dependent performance of the proposed SV system using the VLRs is compared with the baseline system. For the clean speech condition the relative performance improvement (RPI) in case of male speakers is more compared to female speakers. Alternatively, in case of degraded speech, the RPI for female speakers is more compared to male speakers. For instance in case of NIST-2003 original speaker recognition database the RPI for male speakers is 2.14% and for female speakers is -16.5%. Alternatively, for 0 dB *factory-1* degraded NIST-2003 test speech, the RPI in case of male speakers is only 13.08% as compared to 20.41% in case of female speakers. This observation infers that the VLRs in case of female speakers seem to be more robust.

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 mi-

TABLE V
SUMMARY OF SV PERFORMANCE FOR NOISE DEGRADED NIST-2003 TEST SPEECH WITHOUT (W/O) AND WITH T-NORM

Noise	Score normalization	Equal error rate (%)							
		SNR: 9 dB		SNR: 6 dB		SNR: 3 dB		SNR: 0 dB	
		Baseline	VLR	Baseline	VLR	Baseline	VLR	Baseline	VLR
Factory-1	w/o T-norm	17.25	12.87	24.61	15.94	29.85	21.27	34.15	28.36
	T-norm	16.98	12.51	24.66	15.89	29.85	20.82	34.15	28.22
White	w/o T-norm	26.91	17.61	30.53	21.77	35.18	26.51	39.52	32.83
	T-norm	26.84	17.47	30.80	21.59	35.14	26.39	39.49	32.83

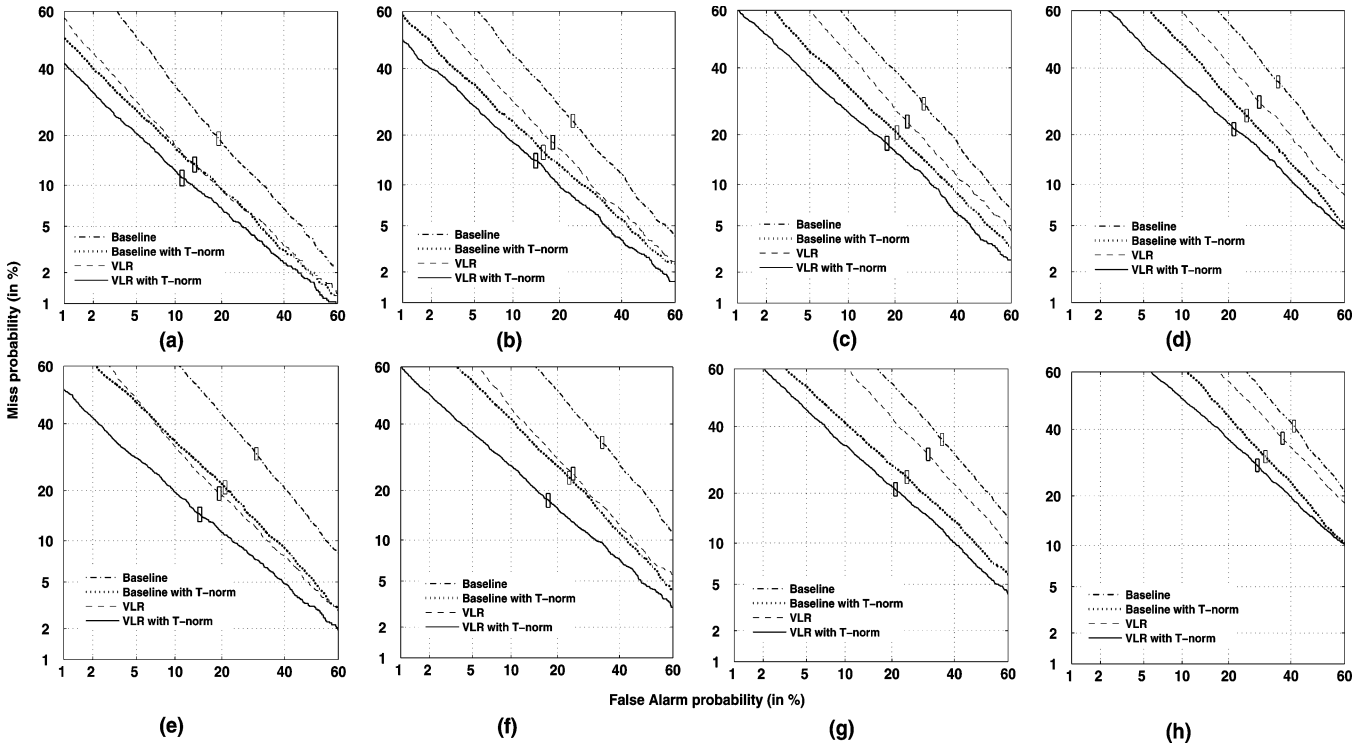

Fig. 9. DET curves showing performance for noise degraded NIST-2003 train speech. (a)–(d) *factory-1 noise* reconstructed train speech for SNR level 9 dB, 6 dB, 3 dB, and 0 dB, (e)–(h) *white noise* reconstructed train speech for SNR level 9 dB, 6 dB, 3 dB, and 0 dB. The boxes indicate the 95% confidence intervals at EER operating points.

crophone 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 performance of SV system using VLRs and baseline system for various experimental setup of IITG MV speaker recognition database is given in Table VIII. The DET plots in Fig. 11(a) shows that for clean and sensor matched condition, performance of the VLRs and baseline system in terms of EER are 2.25% and 1.95%, respectively. The relative performance improvement in the SV using VLRs over the baseline system in terms of EER is −15.38%. 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 the 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 VLRs to the baseline system is 6.21% better compared to the clean experiment of IITG MV database. As discussed earlier, these two results indicate that VLRs are less affected by channel and sensors compared to the non-VLRs.

2) Clean Train and Degraded Test: This experiment is conducted to better investigate performance of the SV system using VLRs 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. 11(b) shows performance of VLRs and baseline system in terms of EER and is 12.7% and 18.6%, respectively. The relative performance improvement in the SV using VLRs over the baseline system in terms of EER is 31.72%. This result shows that for most of the practical uses a better SV system can be developed using VLRs.

3) Degraded Train and Clean Test: This experiment is conducted to verify the significance of VLRs for modeling the speaker information in a more practical environment degraded speech. The DET plots in Fig. 11(c) shows performance of the

TABLE VI
SUMMARY OF SV PERFORMANCE FOR NOISE DEGRADED NIST-2003 TRAIN SPEECH WITHOUT (W/O) AND WITH T-NORM

Noise	Score normalization	Equal error rate (%)							
		SNR: 9 dB		SNR: 6 dB		SNR: 3 dB		SNR: 0 dB	
		Baseline	VLR	Baseline	VLR	Baseline	VLR	Baseline	VLR
Factory-1	w/o T-norm	19.15	13.59	23.70	18.24	28.39	23.35	35.36	28.95
	T-norm	13.64	11.2	15.99	14.31	20.49	17.88	25.02	21.4
White	w/o T-norm	30.21	19.28	33.78	24.39	35.59	30.89	41.10	36.98
	T-norm	20.90	14.63	23.26	17.66	24.22	20.91	31.07	28.31

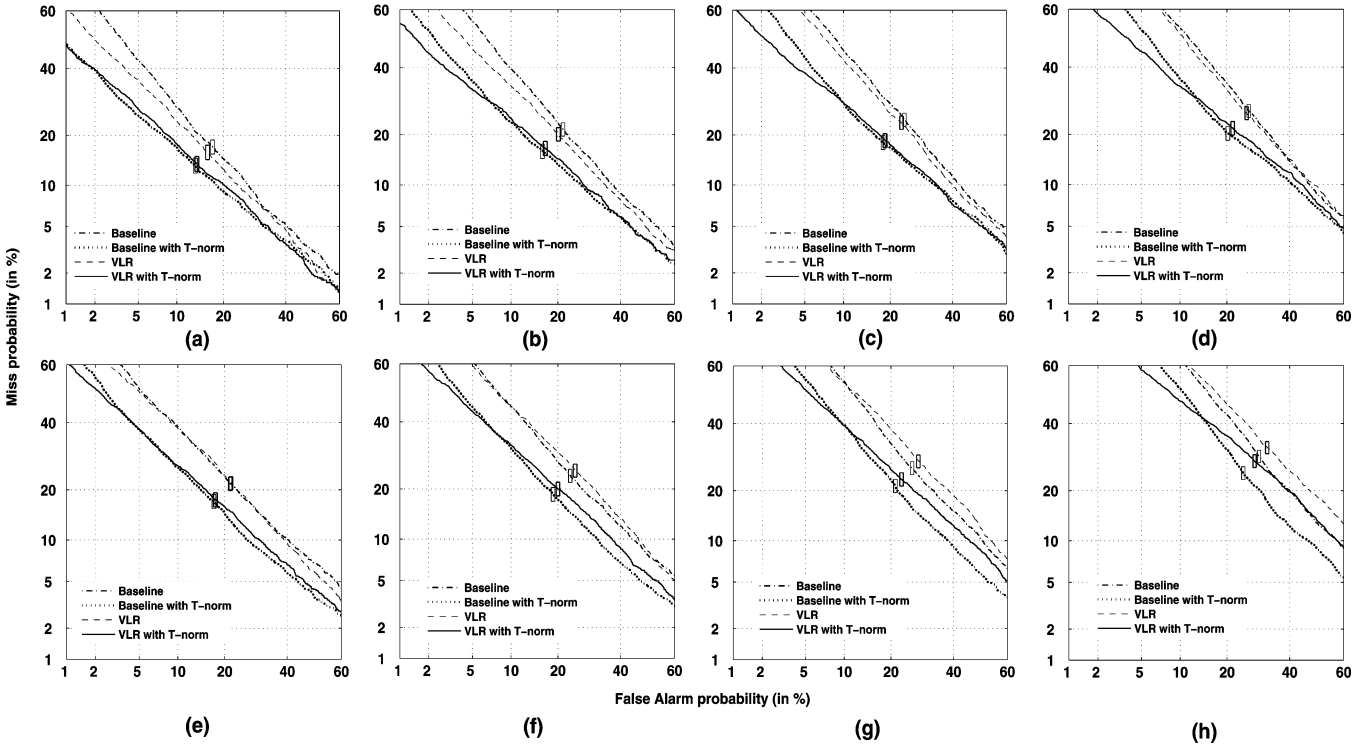

Fig. 10. DET curves showing performance for noise degraded NIST-2003 train and test speech. (a)–(d) *factory-1 noise* reconstructed train and test speech for SNR level 9 dB, 6 dB, 3 dB, and 0 dB, (e)–(h) *white noise* reconstructed train and test speech for SNR level 9 dB, 6 dB, 3 dB, and 0 dB. The boxes indicate the 95% confidence intervals at EER operating points.

TABLE VII
SUMMARY OF SV PERFORMANCE FOR NOISE DEGRADED NIST-2003 TRAIN AND TEST SPEECH WITHOUT (W/O) AND WITH T-NORM

Noise	Score normalization	Equal error rate (%)							
		SNR: 9 dB		SNR: 6 dB		SNR: 3 dB		SNR: 0 dB	
		Baseline	VLR	Baseline	VLR	Baseline	VLR	Baseline	VLR
Factory-1	w/o T-norm	17.25	15.98	21.31	20.05	23.98	23.08	26.15	25.51
	T-norm	13.37	13.68	16.21	16.8	18.15	18.6	20.28	21.54
White	w/o T-norm	21.68	21.49	23.35	24.79	25.97	27.95	29.35	32.11
	T-norm	17.29	17.66	18.74	19.96	21.14	22.85	24.48	27.91

SV system using VLRs and the baseline system in terms of EER and is 11% and 12.2%, respectively. The relative performance improvement in the SV using VLRs over the baseline system in terms of EER is 9.83%. This result shows that even in severely degraded speech signal, using the VLRs better speaker modeling is possible.

4) Degraded Train and Test: This experiment is conducted to verify the performance of SV system in a situation where the training and test speech are degraded in a real environment. In this experiment speech is collected through the same sensor. Therefore, the speech used for this experiment does not contain any degradation for sensor variation. The only difference in this experiment compared to noise degraded train and test speech of NIST-2003 is the noise and SNR level varies from the training to testing condition. The DET plots in Fig. 11(d) shows performance of the SV system using VLRs and the baseline system in terms of EER and is 13.4% and 15.3%, respectively. The relative performance improvement in the SV using VLRs over the baseline system in terms of EER is 12.41%. The experimental results show that in the presence of mismatch between training and test speech, VLRs always provide better performance compared to baseline system.

VI. SUMMARY AND CONCLUSION

In this paper, we proposed a new VLROP detection method for clean and degraded speech by utilizing the advantages of HE of LP residual and zero-frequency filtered signal. The performance of proposed method is evaluated using a 60-speaker

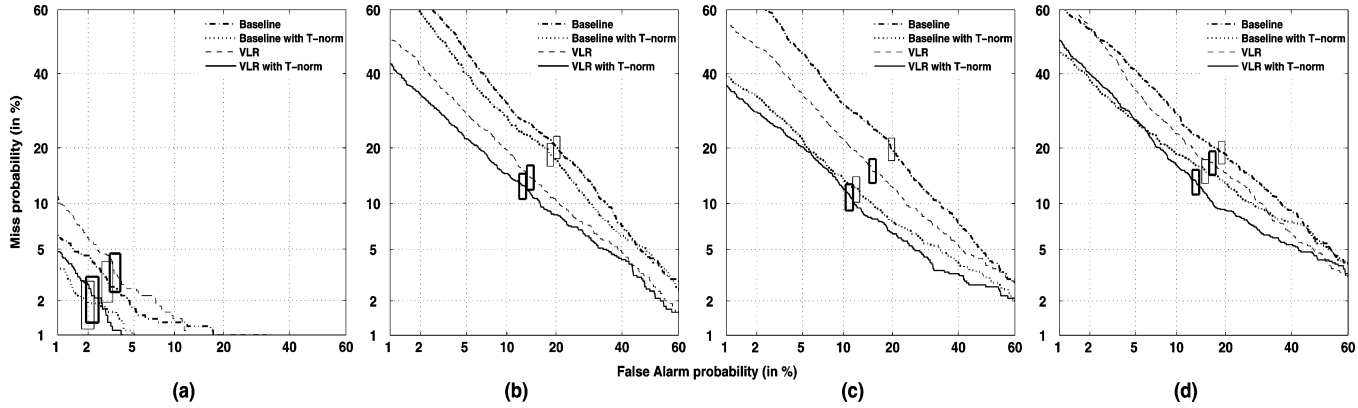

Fig. 11. DET curves showing performance for various experimental setup of IITG MV speaker recognition database. (a) Clean and sensor matched. (b) Clean train degraded test. (c) Degraded train and clean test. (d) Degraded train and degraded test. The boxes indicate the 95% confidence intervals at EER operating points.

TABLE VIII
SUMMARY OF SV PERFORMANCE FOR IITG MV DATABASE WITHOUT (W/O) AND WITH T-NORM

Score normalization	Equal error rate (%)							
	clean sensor matched		clean train degraded test		degraded train clean test		degraded train degraded test	
	Baseline	VLR	Baseline	VLR	Baseline	VLR	Baseline	VLR
w/o T-norm	2.9	3.4	20.2	14.1	19.8	15.3	19.1	16.9
T-norm	1.95	2.25	18.6	12.7	12.2	11	15.3	13.4

subset of the 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 VLRs. A conventional SV system is developed using MFCC as the speaker feature and GMM-UBM as modeling technique and termed as baseline system. The environmental effect is compensated in the cepstral domain using CMS followed by CVN and in the score level using T-norm. The performance of the proposed SV 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 comparable performance 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 therefore be developed by selecting the VLRs.

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

[3] P. Krishnamoorthy and S. R. M. Prasanna, "Enhancement of noisy speech by temporal and spectral processing," *Speech Commun.*, vol. 53, pp. 154–174, Feb. 2011.

[4] B. Yegnanarayana, C. Avendano, H. Hermansky, and P. S. Murthy, "Speech enhancement using linear prediction residual," *Speech Commun.*, 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 Process.*, vol. 8, no. 3, pp. 267–281, May 2000.

[6] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-27, no. 2, pp. 113–120, Apr. 1979.

[7] P. Krishnamoorthy and S. R. M. Prasanna, "Reverberant speech enhancement by temporal and spectral processing," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 2, pp. 253–266, Feb. 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. Speech Audio Process.*, vol. 11, no. 5, pp. 435–446, Sep. 2003.

[9] H. Hermansky, N. Morgan, A. Bayya, and P. Kohn, "RASTA-PLP speech analysis technique," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, San Francisco, CA, Mar. 1992, vol. 1, pp. 121–124.

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

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

[12] R. Teunen, B. Shahshahani, and L. P. Heck, "A model-based transformation approach to robust speaker recognition," in *Proc. Int. Conf. Spoken Lang. Process.*, Beijing, China, Oct. 2000, vol. 2, pp. 495–498.

[13] L. P. Wong and M. Russell, "Text-dependent speaker verification under noisy conditions using parallel model combination," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, Salt Lake City, UT, May 2001, vol. 1, pp. 457–460.

[14] J. Rodriguez, J. Garcica, C. Martin, and L. Hernandez, "Increasing robustness in GMM speaker recognition system for noisy and reverberant speech with low complexity microphone array," in *Proc. Int. Conf. Spoken Lang. Process.*, Philadelphia, PA, Oct. 1996, vol. 3, pp. 1333–1336.

[15] J. L. Gauvain and C. H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 2, pp. 291–298, Apr. 1994.

[16] M. Gales, D. Pye, and P. C. Woodland, "Variance compensation within the MLLR framework for robust speech recognition and speaker adaptation," in *Proc. Int. Conf. Spoken Language Process.*, Philadelphia, PA, Oct. 1996, vol. 3, pp. 1832–1835.

[17] X. Zhang, H. Wang, X. Xiao, J. Zhang, and Y. Yan, "Maximum a posteriori linear regression for speaker recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Dallas, TX, Mar. 2010, pp. 4542–4545.

[18] A. N. Khan and B. Yegnanarayana, "Vowel onset point based variable frame rate analysis for speech recognition," in *Proc. Int. Conf. Intell. Sensing Inf. Process.*, Jan. 2005, pp. 392–394.
[19] 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, Lang. Process.*, vol. 17, no. 4, pp. 556–565, May 2009.
[20] K. N. Stevens, *Acoustic Phonetics*. Cambridge, MA: MIT Pres, 2000.
[21] D. J. Hermes, "Vowel onset detection," *J. Acoust. Soc. Amer.*, vol. 87, pp. 866–873, 1990.
[22] 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 Process.*, vol. 39, no. 9, pp. 2141–2146, Sep. 1991.
[23] J. Y. S. R. K. Rao, C. C. Sekhar, and B. Yegnanarayana, "Neural network based appraoch for detection of vowel onset points," in *Proc. Int. Conf. Adv. Pattern Recognition Digital Tech.*, Dec. 1999, vol. 1, pp. 316–320.
[24] S. R. M. Prasanna and B. Yegnanarayana, "Detection of vowel onset point events using excitation source information," in *Proc. Interspeech*, Lisbon, Portugal, Sep. 2005, pp. 1133–1136.
[25] S. R. M. Prasanna and B. Yegnanarayana, "Extraction of pitch in adverse conditions," in *Proc. Int. Conf. Acoust. Speech, Signal Process.*, May 2004, vol. 1, pp. 109–112.
[26] K. S. R. Murthy and B. Yegnanarayana, "Epoch extraction from speech signals," *IEEE Trans. Audio, Speech, Language Process.*, vol. 16, pp. 1602–1613, Nov. 2008.
[27] K. S. R. Murty, B. Yegnanarayana, and M. A. Joseph, "Characterization of glottal activity from speech signals," *IEEE Signal Process. Lett.*, vol. 16, no. 6, pp. 469–472, Jun. 2009.
[28] J. Makhoul, "Linear prediction: A tutorial review," *Proc. IEEE*, vol. 63, no. 04, pp. 561–580, Apr. 1975.
[29] 'TIMIT Acoustic-Phonetic Continuous Speech Corpus," National Inst. of Standards and Technol., Gaithersburg, MD, NTIS Order PB91-505065, Speech Disc 1-1.1, 1990.
[30] 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, Lang. Process.*, vol. 15, no. 1, pp. 34–43, Jan. 2007.
[31] 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.
[32] "NIST," NIST-Speaker Recognition Evaluations. [Online]. Available: http://www.nist.gov/speech/tests/spk
[33] 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 *Proc. National Conf. Commun. (NCC)*, Bangalore, India, Jan. 2011.
[34] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-28, no. 4, pp. 357–366, Aug. 1980.
[35] F. K. Soong and A. E. Rosenberg, "On the use of instantaneous and transitional spectral information in speaker recognition," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 36, no. 6, pp. 871–879, Jun. 1988.
[36] D. A. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models," *Speech Commun.*, vol. 17, pp. 91–108, Mar. 1995.
[37] X. Zhao, Y. Dong, H. Yang, J. Zhao, and H. H. Wang, "SVM-based speaker verification by location in the space of reference speakers," in *Proc. Int. Conf. Acoust. Speech, Signal Process.*, Honolulu, HI, Apr. 2007, vol. 4, pp. 281–284.
[38] "Linguistic Data Consortium, Switchboard Cellular Part 2 Audio," 2004 [Online]. Available: http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId= LDC 2004S07
[39] D. A. van Leeuwen, A. F. Martin, M. A. Przybocki, and J. S. Bouten, "Nist and NFI-TNO evaluations of automatic speaker recognition," *Comput. Speech Lang.*, vol. 20, no. 1, pp. 128–158, Jan. 2006.
[40] M. K. Omar, J. Navratil, and G. Ramsawamy, "Maximum conditional mutual information modeling for speaker verification," in *Proc. Interspeech*, Lisbon, Portugal, Sep. 2005, pp. 2169–2172.
[41] J. Navratil, G. N. Ramaswamy, and R. D. Zilca, "Statistical model migration in speaker recognition," in *Proc. Interspeech*, Jeju Island, Korea, Oct. 2004.

S. R. Mahadeva Prasanna (M'05) was born in India in 1971. He received the B.E. degree in electronics engineering from Sri Siddartha Institute of Technology, Bangalore University, Bangalore, India, in 1994, the M.Tech. degree in industrial electronics from the National Institute of Technology, Surathkal, India, in 1997, and the Ph.D. degree in computer science and engineering from the Indian Institute of Technology Madras, Chennai, India, in 2004.

He is currently an Associate Professor in the Department of Electronics and Electrical Engineering, Indian Institute of Technology, Guwahati. His research interests are in speech and signal processing.

Gayadhar Pradhan was born in India in 1976. He received the B.E. degree in electronics and communication engineering from Orissa Engineering College, Bhubaneswar, Utkal University, Orissa, India, in 2001 and the M.Tech. degree in electronics and communication engineering from the Indian Institute of Technology, Guwahati, India, in 2009. He is currently pursuing the Ph.D. degree in electronics and electrical engineering at the Indian Institute of Technology, Guwahati.

His research interests are in speech processing and speaker recognition.