SPEAKER VERIFICATION UNDER DEGRADED CONDITIONS USING VOWEL-LIKE AND NONVOWEL-LIKE REGIONS

GAYADHAR PRADHAN

SPEAKER VERIFICATION UNDER DEGRADED CONDITIONS USING VOWEL-LIKE AND NONVOWEL-LIKE REGIONS

A

Thesis submitted

for the award of the degree of

DOCTOR OF PHILOSOPHY

By

GAYADHAR PRADHAN

DEPARTMENT OF ELECTRONICS AND ELECTRICAL ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY GUWAHATI

GUWAHATI - 781 039, ASSAM, INDIA

MARCH 2013

Certificate

This is to certify that the thesis entitled "**SPEAKER VERIFICATION UNDER DEGRADED CONDITIONS USING VOWEL-LIKE AND NONVOWEL-LIKE REGIONS**", submitted by **Gayadhar Pradhan** (09610214), a research scholar in the *Department of Electronics and Electrical Engineering, Indian Institute of Technology Guwahati*, for the award of the degree of **Doctor of Philosophy**, is a record of an original research work carried out by him under my supervision and guidance. The thesis has fulfilled all requirements as per the regulations of the institute and in my opinion has reached the standard needed for submission. The results embodied in this thesis have not been submitted to any other University or Institute for the award of any degree or diploma.

Dated:

Guwahati.

Prof. S. R. Mahadeva Prasanna
Professor
Dept. of Electronics and Electrical Engg.
Indian Institute of Technology Guwahati
Guwahati - 781 039, Assam, India.

To

Lord Dhabaleswar

for His blessings

My son **Asutosh** and wife **Sushri**

for their love and sacrifice

My guide **Prof. S. R. M. Prasanna**

for his guidance and inspiration

&

My **parents** and **parents-in-law**

for their blessings

Acknowledgements

I am obliged to GOD for His divine guidance and blessings. I solely dedicate my thesis to lord Dhabaleswar.

This thesis would not have been possible without the immense help and support of several people in various measures. I would like to convey my acknowledgment to all of them.

First and foremost, I express my sincere gratitude to my research supervisor, Prof. S. R. M. Prasanna for providing me an opportunity to work under his guidance. It is very difficult to describe my feelings in words to acknowledge my supervisor for his continuous guidance in all aspects, constant motivation and support throughout the doctoral studies. I am very much thankful to him for transforming me from an unstructured form to a structured form in every aspect of my life and showing me a different path of life. It would be completely impossible for me to bring the research as well as the thesis to this form without the immense facilities provided by him in the EMST Laboratory and the freedom of work he has given to me.

I am thankful to my doctoral committee members Prof. S. Dandapat, Dr. R. Sinha and Prof. S. Nandi for their encouragement and valuable suggestions on my work. I would like to thank faculty members and the office staffs of the Department of Electronics and Electrical Engineering, IIT Guwahati, for their help in carrying out this research work. I am very much thankful to Mr. Sanjib Das for his kind support and help.

I sincerely thank to Prof. B. Yegnanarayana for his valuable suggestions on my research work during the PRSG meetings, which helped me a lot in my research. I would like to acknowledge E-Security Division, Department of Information technology, New Delhi for providing enormous funding to build advance computational facility in the EMST Laboratory.

I am very much thankful to my friend Haris who helped me whenever I need him.

I am thankful to my friends Deepak K. T and K.K. Ramesh for their assistance in writing my thesis. I am thankful to Soyuj Kumar sahoo, R. C. Mishra sir, Govind. D, Sibasankar Padhy, Malaya Kumar Nath, Biswajit Dev Sarma, Syed Shahnawazuddin, Rohan Kumar Das, Avinab Mishara, Nagaraj Adiga, Ramesh K Bhukya, Sunil. Y and all other members in the EMST Laboratory.

I would like to thank my senior members Dr. S.R. Nirmala, Dr. P. Krishnamoorthy, Dr. M. Sabarimalai Manikandan, Dr. H.S. Jayanna and Dr. D. Pati. My special thanks to Dr. L. N. Sharma

for maintaining the EMST laboratory smoothly.

I am thankful to my wife for her sacrifice and support. I am heartily thankful to my one year lovely son who has sacrificed his valuable one year for me. I am also thankful to Badabhai, Guga, Dipu, Somya, Madhusmita and Dei for their love and support.

I attribute this achievement to my parents and parents-in-law for their constant blessings, support, silent prayers for my success and moreover, making me stand in this position.

Gayadhar Pradhan

Abstract

This thesis proposes a speaker verification system by independent processing of vowel-like regions (VLRs) and non-vowel-like regions (non-VLRs) for achieving better SV performance under clean and degraded conditions. VLRs are defined as the speech regions belonging to vowels, diphthongs and semivowels, and rest of the consonants as non-VLRs. Methods are proposed for detecting VLRs and non-VLRs using excitation source information. The VLR onset point (VLROPs) and end points (VLREPs) are hypothesized and used in an iterative algorithm for detecting the VLRs. Next, for detection of non-VLRs, the linear prediction (LP) residual samples in the VLRs are attenuated significantly to indirectly emphasize the residual samples in the non-VLRs. The modified LP residual samples excite the time varying all pole filter to reconstruct non-VLRs enhanced speech and used for detecting non-VLRs.

For any practical application of a *text-independent* speaker verification (SV) system, along with phonetic variability, the speech signal may be affected by background noise, sensor mismatch and channel mismatch. To reduce the effect of these variabilities, three different methods are proposed for processing the VLRs and non-VLRs during training and testing of a SV system. First, a SV system is developed by using only the VLRs to demonstrate the significance of the VLRs for SV under degraded conditions. Then, the VLRs and non-VLRs are used independently during training and testing of a SV system, and the scores are combined with higher weight on VLRs, for a better SV system under clean and degraded conditions. Finally, a SV system is developed by implicit modeling of VLRs and non-VLRs information to reduce the computational complexity involved in the explicit segmentation of these regions. The experimental results presented in this thesis work shows that the VLRs are more speaker specific and relatively less affected under degraded conditions. A better SV system can be developed under clean and degraded conditions by independent processing of VLRs and non-VLRs with emphasis on the VLRs.

The major contributions of this thesis are as follows:

- Method for the detection of VLROPs and end VLREPs using excitation source information.
- An iterative algorithm for the detection of complete VLRs using VLROP and VLREP.
- Method for the detection of non-VLRs by emphasizing excitation information of non-VLRs in the LP residual.
- Demonstrating significance of VLRs for SV under degraded conditions.
- SV system using VLRs and non-VLRs conditioning and emphasizing the scores from VLRs for better SV under clean and degraded conditions.
- SV by implicit modeling of VLRs and non-VLRs information to avoid the computational complexity involved in explicit segmentation of training and testing speech data.

Keywords: Degraded condition, speaker information, speaker verification (SV), vowel-like region (VLR), non-vowel-like region (non-VLR), VLR onset point (VLROP), VLR end point (VLREP).

Contents

List of Figures

List of Tables

List of Acronyms

ANN	Artificial Neural Network
CDF	Cumulative Distribution Function
CMS	Cepstral Mean Normalization
CMVN	Cepstral Mean and Variance Normalization
CV	Consonant Vowel
CVN	Cepstral Variance Normalization
DCF	Detection Cost Function
DET	Detection Error Tradeoff
D-norm	Distance normalization
DTFT	Discrete Time Fourier Transform
DYPSA	Dynamic Programming Projected Phase-Slope Algorithm
EER	Equal Error Rate
FAR	False Acceptance Rate
FOGD	First Order Gaussian Differentiator
FRR	False Rejection Rate
GCI	Glottal Closure Instant
GMM	Gaussian Mixture Models
HE	Hilbert Envelope
HLDA	Heteroscedastic Linear Discriminant Analysis
H-norm	Handset Dependent Score Normalization
HT-norm	Handset Dependent Test score Normalization
HMM	Hidden Markov Models
IA	Identification Accuracy
IDTFT	Inverse Discrete Time Fourier Transform

IR	Identification Rate
JFA	Joint Factor Analysis
LDA	Linear Discriminant Analysis
LP	Linear Prediction
LPC	Linear Predictive Coefficients
LPCC	Linear Predictive Cepstral Coefficients
LSR	Log-Likelihood Score
LLSR	Log-Likelihood Score Ratio
MAP	Maximum *a Posteriori* Adaptation
MFCC	Mel-Frequency Cepstral Coefficients
MMSE	Minimum Mean Square Error
MR	Miss Rate
MV	Multi-Variability
NAP	Nuisance Attribute Projection
NG	Net Gain
non-VLR	Non-vowel-like region
PDF	Probability Density Function
PLP	Perceptual Linear Prediction
PMC	Parallel Model Combinations
PSD	Power Spectral Density
RASTA	RelAtive SpecTrAl Processing
SCMVN	Segmental Mean and Variance Normalization
SMS	Speaker Model Synthesis
SNR	Signal to Noise Ratio
SR	Spurious Rate
SRR	Signal to Reverberant Ratio
STG	Short Time Gaussianization
STMSN	Short-time Cepstral Mean and Scale Normalization
STSA	Short Time Spectral Amplitude
SV	Speaker Verification

SVM	Support Vector Machine
T-norm	Test Score Normalization
UBM	Universal Background Model
VAD	Voice Activity Detector
VLRs	Vowel-Like Regions
VLREP	Vowel-Like Region End Point
VLROP	Vowel-Like Region Onset Point
VOP	Vowel Onset Point
VQ	Vector Quantization
WCCN	Within Class Covariance Normalization
Z-norm	Zero score Normalization
ZFF	Zero Frequency Filtering
ZFFS	Zero Frequency Filter Signal

List of Symbols

$\boldsymbol{v}$	Arbitrary vector
E_{Avg}	Average energy threshold value
$\boldsymbol{B_c}$	Between-class covariance matrix
$\boldsymbol{B}_{c_{nvl}}$	Between-class covariance matrix for non-VLRs
$\boldsymbol{B}_{c_{vl}}$	Between-class covariance matrix for VLRs
$\boldsymbol{N}_c$	0^{th} order statistics
$\boldsymbol{F}_c$	Centralized 1^{st} order statistics
$X_{mvn}(n,k)$	Cepstral mean and variance normalized feature
$\hat{\boldsymbol{y}}_{clm}$	Claimed i-vector
C^nV	Consonant cluster vowel
$\boldsymbol{\Sigma}_c$	Covariance matrix of c^{th} GMM component
$\boldsymbol{\psi}$	Cumulative Distribution Function
Δ	Delta coefficients
$\Delta\Delta$	Delta-Delta coefficients
$x(n)$	Difference speech signal
F	Dimension of feature vectors
$E(\omega)$	DTFT of LP residual
$\mid \hat{S}_i(e^{j\omega})\mid^2$	Estimated power spectrum of clean speech
$\mid \hat{N}_i(e^{j\omega})\mid^2$	Estimated power spectrum of noise
$\boldsymbol{U}$	Eigen channel matrix
$\boldsymbol{\lambda}$	Eigen value matrix
$\boldsymbol{R}$	Eigen vectors matrix
$\boldsymbol{V}$	Eigen voice matrix
$X(n,k)$	n^{th} feature frame in k^{th} feature space

X	feature vectors
$\boldsymbol{M}_{nvl}$	GMM mean supervector for non-VLRs
$\boldsymbol{M}_{s}$	GMM mean supervector for speech utterance
$\boldsymbol{M}_{vl}$	GMM mean supervector for VLRs
$h_e(n)$	Hilbert Envelope of LP residual
$e_h(n)$	Hilbert transform of LP residual
$\boldsymbol{w}_{nvl}$	i-vector representation of non-VLRs
$\boldsymbol{w}$	i-vector representation of speech utterance
$\boldsymbol{w}_{vl}$	i-vector representation of VLRs
$\boldsymbol{A}$	LDA projection matrix
$\boldsymbol{S}_{\lambda}$	Likelihood for the claimed speaker
$\boldsymbol{S}_{\bar{\lambda}}$	Likelihood for background speakers
$\boldsymbol{\Lambda}$	Linear transform matrix
$\phi(n)$	LP residual phase
$e(n)$	LP residual signal
$\boldsymbol{\mu}_k$	Mean of k^{th} component of the feature vectors
$\boldsymbol{\mu}_I$	Mean of impostor score distribution
μ_c	Mean vector of c^{th} GMM component
$\boldsymbol{\Gamma}$	Nonlinear transformation
$\boldsymbol{S}_n$	Normalized score
SR_1	Non-speech spurious rate
C	Number of components in GMM-UBM
$\boldsymbol{S}$	Original score
$f(y)$	PDF of standard normal distribution
$\boldsymbol{P}$	Projection matrix of NAP
$\left\|S_i(e^{j\omega})\right\|^2$	Power spectrum of noisy speech
r	Rank of a feature
EER_R	Relative improvement in EER
$\boldsymbol{c}$	Session/channel supervector in JFA
$X_{stmsn}(n,k)$	Short-time cepstral mean and scale normalized feature

$\boldsymbol{d}_{st}(n,k)$	Short-time difference of a window of L feature frames
$\boldsymbol{\mu}_{st}(n,k)$	Short-time mean of a window of L feature framess
$\boldsymbol{s}$	Speaker supervector in JFA
$s(n)$	Speech signal
SR_1	Speech spurious rate
$\boldsymbol{\sigma}_k$	Standard deviation of features
$\boldsymbol{\sigma}_I$	Standard deviation of impostor score distribution
$\hat{\boldsymbol{y}}_{tst}$	Test i-vector
$\boldsymbol{T}$	Total variability matrix
$\boldsymbol{T}_m$	Total variability matrix for microphone speech
$\boldsymbol{T}_{nvl}$	Total variability matrix for non-VLRs
$\boldsymbol{T}_p$	Total variability matrix for telephone speech
$\boldsymbol{T}_{vl}$	Total variability matrix for VLRs
$\boldsymbol{m}$	UBM supervector
$\boldsymbol{U}$	Universal background model
S_{nvl}	Verification scores from non-VLRs
S_{vl}	Verification scores from VLRs
$\boldsymbol{w}_c$	VLRs and non-VLRs conditioned i-vector
$\boldsymbol{T_c}$	VLRs and non-VLRs condition $\boldsymbol{T}$ matrix
S_c	VLRs and non-VLRs combined score
η_c	Weight associated with c^{th} component of GMM
$\boldsymbol{W_c}$	Within-class covariance matrix
$\boldsymbol{W}_{c_{nvl}}$	Within-class covariance matrix for non-VLRs
$\boldsymbol{W}_{c_{vl}}$	Within-class covariance matrix for VLRs
$\mathbf{w}_f$	Weight function for enhancement of non-VLRs
w	Weight for score combination
$\hat{y}(n)$	Zero Frequency Filtered Signal

1

Introduction

Contents

Objective

This chapter first describes a speaker verification (SV) system and highlights the issues related to the development of a SV system under degraded conditions, and motivates a solution for achieving better SV performance by independent processing of vowel-like regions (VLRs) and non-vowel-like regions (non-VLRs). VLRs are defined as the speech regions belonging to vowels, diphthongs and semivowels, and rest of the consonants as non-VLRs. Primarily, the source of motivation originates from the fact that the nature of the signal and speaker specific information present in the VLRs is different compared to non-VLRs. By independent processing of VLRs and non-VLRs the gross level mismatch with respect to the sound units may be reduced. Under degraded conditions, a better compensation of degradation effect may be achieved by applying different normalizations to these two different segment types. Also, higher weight can be applied to the scores obtained from the VLRs, which are less degradation affected and more speaker specific compared to non-VLRs. The chapter then defines the research issues to be addressed for the development of a SV system using VLRs and non-VLRs. The chapter concludes with a brief description of the organization of the thesis.

1.1 Speaker verification (SV) system

SV system validates the identity claim of a person [1–3]. A SV system can be classified into *text-dependent* and *text-independent* depending on the text used for training and testing. In a *text-dependent* SV system the verification text is fixed, or known to the system. In a *text-independent* SV system there is no such constraint. Generally, a *text-independent* SV is more challenging due to additional phonetic mismatch between the training and testing utterances [1–3]. Depending on the application each of them have their own merits and demerits. In this thesis work, all studies are presented for a *text-independent* SV system, specifically under degraded conditions like environmental noise, reverberation, sensor mismatch and channel mismatch.

1.1.1 Modular representation of a SV system

Fig. 1.1 shows modular representation of a SV system. The SV process is divided into two phases of operation: training and testing phase. In the training phase, for each speaker a reference model (target model) is built using the training speech. In the testing phase, the similarity between the test speech and claimed model is compared against a verification threshold to make the verification

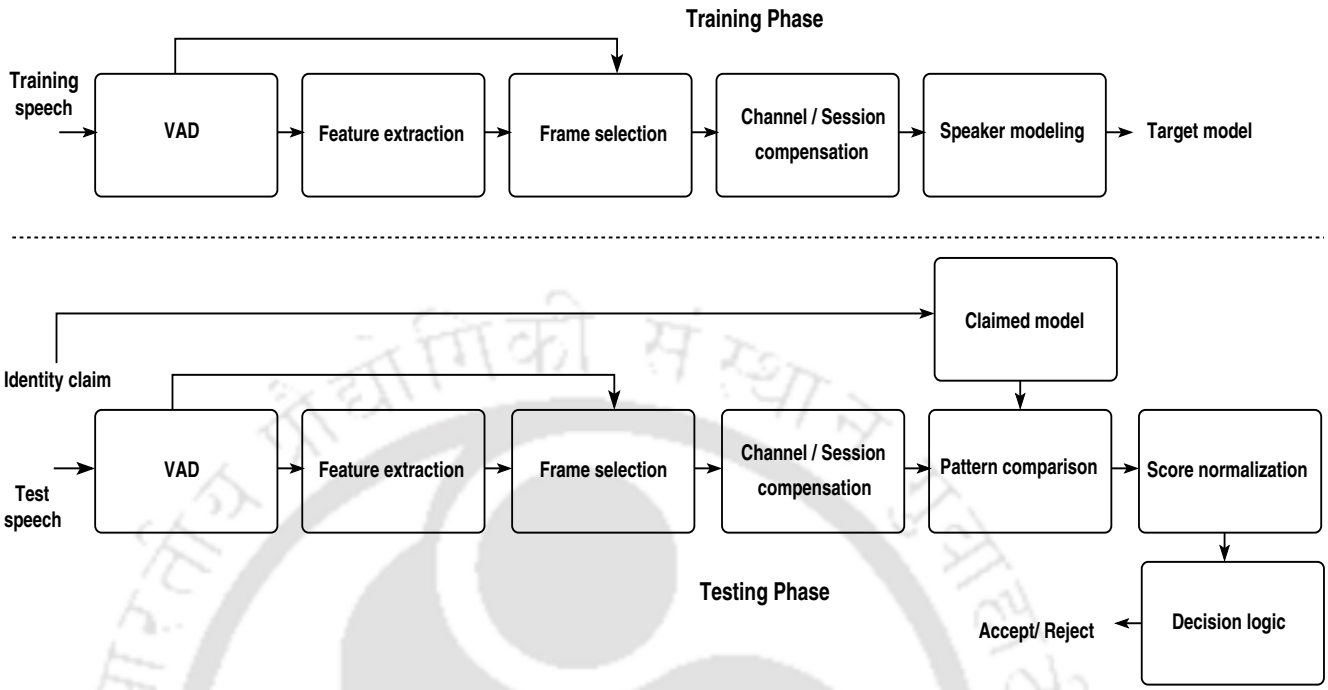

Figure 1.1: Modular representation of SV system.

decision. The speech data used for the development of speaker models and testing contains many other redundant and unwanted information. It is therefore processed through several intermediate stages to suppress the redundant information present in the acoustic speech signal, as follows:

1.1.1.1 Voice activity detection

During training and testing phase of a SV system, speech data is first processed through a voice activity detector (VAD) to separate the speech regions from non-speech regions [1, 4]. Even though VAD is a simple binary classification task, it is very difficult to implement a VAD which works consistently for different types of speech data. A number of VAD algorithms have been proposed using different acoustic features [5–8] and discrimination models [9–13]. Each of these VAD methods have their own merits and demerits depending on accuracy, complexity and robustness. Due to the simplicity in implementation and less computational complexity, for most of the SV studies signal energy is used for VAD. The short-term energy is computed for each analysis frame using the normalized speech signal [-1,1] and the frames having energy above certain threshold are considered as the speech frames. Only these frames are processed further for speaker verification.

1.1.1.2 Feature extraction

With the speaker specific information many other redundant factors like acoustic environment, sensor, channel, language, style etc. are present in the speech data. The feature extraction module transforms speech to a set of feature vectors of reduced dimension in which the speaker specific information are emphasized and other redundant factors are suppressed [1, 2]. A good feature is expected to be robust to varied environmental and recording conditions [14–16], and have minimum intra speaker variability and maximum inter speaker variability [3, 17]. The speaker specific information in speech data is mostly attributed to the unique physiological and behavioral aspects of the speaker [1, 3]. The physiological aspect is due to shape, size and dynamics of the vocal tract and the excitation source [18–22]. The behavioral aspect is due to unique conversational characteristics of the speaker which includes speaking rate, speaking style and use of specific words [23–27]. Several feature extraction methods have been proposed to characterize the vocal tract [4, 18, 19, 28–30], excitation source [20–22] and behavioral characteristics [15, 31, 32] of a person. The vocal tract based features are best performing features [4, 33, 34] and are relatively more affected under degraded conditions [16, 33, 34]. Alternatively, the excitation source and behavioral features are relatively less affected under degraded conditions and less speaker discriminative [3, 14, 16]. Most of the state-of-the-art SV systems use the Mel-frequency cepstral coefficients (MFCCs) appended with first and second order derivatives known as delta (Δ) and delta-delta($\Delta\Delta$) due to their good verification accuracy [35–37]. The MFCCs appended with Δ and $\Delta\Delta$ mostly represent the static and dynamic characteristics of the vocal tract.

1.1.1.3 Channel/ Session variability compensation

As the features extracted from speech data varies depending the category of sound unit, speaker dependent variability and recording condition [3, 38, 39], it is processed though the channel/ session compensation module to compensate the redundant information [40–42]. Channel/ session variability compensation can be performed on feature vectors [40, 41, 43–45] or on supervector [42, 46, 47]. The compensation is applied either in a unsupervised manner [19, 43, 44] or in a supervised manner using large amount of similar development data [42, 48–52]. Therefore performance of these methods depend either on the estimation of the degradation effect or *a priori* knowledge about the testing condition.

1.1.1.4 Speaker Modeling

The speaker models are created using the channel/session compensated feature vectors extracted from the training speech. In this process a set of feature vectors are grouped into their representative vectors. For a *text-independent* SV system commonly used modeling methods include, vector quantization (VQ) [53, 54], Gaussian mixture model (GMM) [4, 55], GMM- universal back ground model (GMM-UBM) [35], artificial neural network (ANN) [56, 57] and support vector machine (SVM) [58]. The modeling methods like VQ, GMM and GMM-UBM estimate the feature distribution for a speaker. Alternatively, ANN and SVM model the boundary between the speakers. With recent advancements in the SV technology, the speech data with variable number of feature vectors are represented using a fixed low dimensional supervectors to preserve maximum speaker dependent variabilities [37, 42]. In [37], the GMM mean supervector is projected to a low rank matrix called as the total variability matrix to get a reduced dimension representation which is called as identity vector or *i*-vector for short [37].

1.1.1.5 Pattern comparison

During the testing phase, an unknown test speech is represented by channel/session compensated feature vectors and compared against the claimed model to obtain similarity score. The similarity measure is done based on the employed modeling method. For instance, Euclidean distance [54], log likelihood score (LLS) [4, 55] and log likelihood score ratio (LLSR) [35] are used as the similarity scores for VQ and GMM and GMM-UBM modeling technique, respectively. In a *i*-vector based SV system, the test speech is represented as the channel/session compensated *i*-vector and the cosine kernel between the claimed *i*-vector and test *i*-vector is used as the similarity measure [37].

1.1.1.6 Score normalization and Decision logic

In the final stage of SV system, the verification score obtained from the claimed model is compared against a decision threshold to accept/reject the claim. The speech data used for the development of model and testing varies between the speakers. Secondly, for the same speaker quality and quantity of test data varies between the trials. As a result, the verification score varies between the trials. Compensation of speaker and text dependent variabilities at the score level is commonly known as score normalization [2, 3]. The score normalization helps to reduce degradation and mismatch effect that are not compensated at the feature and model levels. It is also transforms scores from different

trials into a similar range so that a common verification threshold can be used [2, 3, 59, 60]. There are various score normalization techniques like Hnorm [35, 61], Tnorm [60] and HTnorm [62]. These methods are suitable for lower level degradations like channel and sensor mismatch and requires *a priori* information about the sensor and channel for better performance.

1.1.1.7 Performance measure

A perfect SV system should accept all the true claims and reject all the false claims [19]. Depending on the variability between the training and testing speech some true claims may be rejected and some false claims may be accepted. Therefore the SV performance is measured in terms of false rejection rate (FRR) and false acceptance rate (FAR), more meaningfully in terms of equal error rate (EER) [2, 3]. EER is defined as the error rate at which FAR is equal to FRR. In order to improve the visualization of the SV performance, the detection error tradeoff (DET) curve [63,64] is used for performance measure, where miss and false alarm error probabilities are plotted based on the verification scores assigned to the true and false trials. The DET plot help to find the system threshold to maintain the balance between user convenience and security. Along with the EER the detection cost function (DCF) is also used as a metric for to measure a SV performance [63, 64]. The DCF is a weighted sum of miss and false alarm error probabilities [65]:

$C_{Det} = C_{Miss} \times P_{Miss|Traget} \times P_{Target} + C_{FalseAlarm} \times P_{FalseAlarm|NonTraget} \times (1 - P_{Target})$

Where $P_{Miss|Traget}$ and $P_{FalseAlarm|NonTraget}$ are the system probability of an error for a condition. For instance, $P_{Miss|Traget}$ is the system probability to miss a true trial. Similarly, $P_{FalseAlarm|NonTraget}$ is the system probability to accept a false trail. The parameters C_{Miss}, $C_{FalseAlarm}$ are the cost of the detection errors and P_{Target} is the *a priori* probability of the specified target speaker. For the NIST 2003 speaker recognition evaluation, the parameter values for C_{Miss}, $C_{FalseAlarm}$ and P_{Target} are used as 10, 1 and 0.01, respectively.

1.1.2 Issues in the conventional SV system

A SV system provides good performance when the speech signal is of high quality and free from any mismatch [3]. However, for most of the real world applications of a SV system, the speech signal is affected by different degradations like background noise, reverberation, sensor mismatch and channel mismatch, resulting in degraded speech. The phonetic variability between training and testing speech is another major source of mismatch for a *text-independent* SV system. The speech signal may

also vary depending on the health, emotional condition and aging of the speaker (within speaker variation) [2,3]. The characteristics of the acoustic speech signal varies depending on the sound unit, degradation effect and within speaker variabilities [38,39,66,67]. In the presence of these variabilities, the feature distribution of test speech deviates from that of the train speech. As a result the accuracy of the SV system falls significantly [3, 39]. This difficulty may be removed up to certain extent by imposing the system to operate under certain fixed set of conditions. This will not solve the actual purpose of a SV system, which is meant for remote access. This may be the reason for which main focus of SV research has been in tackling the degradation and mismatch effect.

1.2 Motivation for the present work

As discussed above, most of the existing methods focused on normalization and speaker adaptation methods to tackle the degradation and mismatch effect [2,3,35,36,40,41,43,44,68,69]. Performance of these methods depend either on the estimation of the degradation effect or *a priori* knowledge about the testing condition. On top of existing approaches, the performance of the SV system can be further improved by processing speech regions based on the similarity in the speaker specific information and nature of the signal.

- The features derived from the speech signal vary depending on the category of sound unit [38] and recording environment [39]. The mismatch due to sound units may be reduced by the segmentation of speech into some broad categories and performing matching among them [66,67,70]. For a two class segmentation, the vowel and non-vowel like segmentation is preferable from SV point of view, due to their similarity involved in the production process [71, 72] and may exhibit similar speaker information.

  Fig. 1.2 shows a comparative plot of log likelihood scores for VLR/non-VLR, vowel/non-vowel and voiced/unvoiced segmentations, both in clean and 10 dB white noise added testing conditions. The segmentation is done using the phoneme transcription files available in the TIMIT database [73]. For each category, 16-component Gaussian mixture model (GMM) is trained using the first six sentences from one speaker, while the remaining four sentences are used for testing [35]. In each case, the log likelihood score is the average value of the scores obtained by conditioning on segment class. For instance, in case of VLR/non-VLR segmentation, features from VLRs and non-VLRs are scored using the VLR and non-VLR models, respectively, and

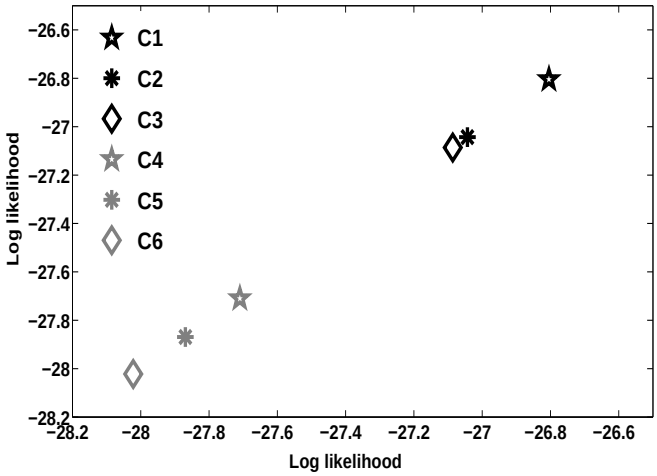

Figure 1.2: Log likelihood scores for different two class segmentations (using reference marking) for one speaker both in clean and noise added testing conditions. The abbreviations C1, C2 and C3 refer to the log likelihood scores for VLRs/non-VLRs, vowel/ non-vowel and voiced/ unvoiced segmentation for clean speech, respectively; C4, C5 and C6 refer to the corresponding log likelihood scores for 10 dB white noise added test speech.

the final score is computed as the average of the two scores. Both for clean and noise added speech, the log likelihood score is maximum for the VLR/non-VLR segmentation. This simple test shows that there is less intra-variabilities between the feature vectors derived from the VLRs and non-VLRs. The better log likelihood score for VLRs/ non-VLRs segmentation compared to vowel/ non-vowel segmentation shows that the signal characteristics and speaker discriminating information present in semivowels and diphthongs are more similar to the vowels compared to other speech regions. we may benefit by segmenting speech into VLRs and non-VLRs and then processing them separately for speaker verification. We conjecture that the broad mismatch with respect to the sound units may be handled.

- VLRs are produced by keeping the vocal tract in an open configuration which offers relatively less obstruction for the air flow and hence high energy regions. Thus, the VLRs have relatively more energy and are robust against degradation due to their impulse-like excitation of the vocal folds [71, 72].
- The major excitation that provides speaker characteristics to the speech signal is the vibration of vocal folds [72]. VLRs are produced using the vocal folds vibration and hence may have relatively more speaker specific information compared to non-VLRs 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. Thus VLRs are more speaker specific and can tolerate more degradation compared to non-VLRs.

- Even though non-VLRs are affected relatively more by degradation due to their low energy nature and non-impulse type excitation, they form sizable number of frames and also contain good amount of speaker information.
- With VLRs and non-VLRs segmentation, efforts may be focused mainly for non-VLRs against degradation. All the relevant regions from speech can be detected and used for SV in an independent and parallel fashion. Under degraded conditions, better compensation of degradation effects may be achieved by applying different normalizations to these two different segment types [74]. Also, higher weight can be applied to the scores obtained from the VLRs, which are more speaker specific than non-VLRs.

1.3 Issues in the development of a SV system using VLRs and non-VLRs

To develop a SV system using VLRs and non-VLRs, the main research issues are as follows:

- The performance of a SV system using VLRs and non-VLRs depends on the accuracy and robustness of VLRs and non-VLRs detection methods. The VLRs may be detected by finding the VLR onset points (VLROPs) and VLR end points (VLREPs). Several methods have been proposed for the detection of vowel onset point (VOP) [75–79]. 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. Therefore robust features are required to deal with degradation.

  The signal characteristics at the end of a VLR are significantly different than at the beginning. At the onset, there is a sudden increase in signal strength, while the signal strength decreases slowly at the end. Due to this, detecting VLREPs is more challenging than detecting VLROPs. An independent method is required for the detection of VLREPs in the speech.

  The non-VLRs are relatively low energy regions compared to VLRs. In a degraded environment, as the level of noise increases, these regions partially or totally merge with the noise. Therefore, special care is required for the development of an automatic non-VLRs detection method.

- As discussed in the previous section, VLRs are more speaker specific and can tolerate more

degradation compared to non-VLRs. Therefore a better SV system may be possible either by using only VLRs or by giving priority on the verification scores obtained from the VLRs. The later case may be achieved either by only segmenting the testing speech into VLRs and non-VLRs or by conditioning a SV system to VLRs and non-VLRs. To demonstrate all these issues, the speaker specific information in the VLRs and non-VLRs, and their significance under degraded conditions need to be investigated in detail. A investigation is also required to find the impact of acoustic mismatch between VLRs and non-VLRs on the SV performance. This requires modification in different stages of a SV system.

- SV by an explicit segmentation of VLRs and non-VLRs always increases the computational complexity. Therefore a method is required for implicit modeling of VLRs and non-VLRs information. The implicit modeling may be done by considering the subspace of VLRs and non-VLRs as independent or by finding the similarity and difference between them, which requires a detailed investigation.

1.4 Organization of the Thesis

To address the issues mentioned in the previous section, this thesis work is organized into six chapters. The content of each chapter is summarized as follows:

- Chapter 2 reviews several existing methods for SV under degraded and mismatch conditions. The review is broadly divided into five sections: speech detection and enhancement, feature extraction and feature combination, feature and supervector normalization, speaker model compensation and score normalization, depending on the similarity of different methods. Summary of the review and the scope for this thesis work is discussed.
- In Chapter 3, a VLRs onset point (VLROP) detection algorithm is proposed using the excitation source information. A VLRs detection method is proposed using the VLROPs. The GMM-UBM based SV systems are developed using only VLRs and speech regions detected by a energy based VAD. For clean and different degraded conditions, performance of both the systems are compared to demonstrate the merits of processing more speaker specific and less degradation affected VLRs for SV task.
- In Chapter 4, the vowel-like region end point (VLREP) event is defined and an iterative algorithm is proposed for detection of the complete VLRs using VLROPs and VLREPs. A method

is proposed for detecting non-VLRs by emphasizing excitation information of non-VLRs in the linear prediction (LP) residual. The SV systems are developed by explicitly modeling VLRs and non-VLRs segments. The merit of proposed approach is evaluated for clean and different degraded conditions using the GMM-UBM and i-vector based SV systems.

- In Chapter 5, a SV system is proposed by implicitly modeling of VLRs and non-VLRs information in the i-vector and the performance is evaluated for clean and different degraded conditions.
- Chapter 6 summarizes the work presented in this thesis, highlights the main contributions of the work and gives some directions for future research.

2

Speaker Verification Under Degraded Conditions - A Review

Contents

Objective

This chapter reviews the literature pertaining to the speaker verification (SV) under degraded conditions. There are many methods presented in the literature to tackle the degradation effect staring from speech signal to verification score. These methods are presented in five sections depending on the similarity in the approach taken. Literature in the area of front-end signal analysis, feature normalization, session/channel variability compensation, speaker model compensation and score normalization are discussed in Section 2.1, 2.2, 2.3, 2.4 and 2.5, respectively. Finally, the chapter concludes with a brief summary of the literature review and the scope for the present work.

2.1 Front-end signal analysis

In a SV system, first the acoustic speech signal is processed to detect the speech regions [35–37, 80] or different constraint regions [66, 70, 81]. The speech signal can be subjected to speech enhancement for reducing the impact of additive noise present in degraded speech [16, 82–84]. This section presents a literature review addressing different approaches for front-end signal analysis.

2.1.1 Detection of speech regions

During training and testing phase of a SV system, first the acoustic speech signal is segmented into speech and non-speech regions, which is popularly known as voice activity detection (VAD) [5, 6, 9, 10]. Then, the features extracted only from the speech segments are processed through different modules of a SV system for speaker modeling and pattern matching. Since, the non-speech regions in the acoustic speech signal do not contain any speaker specific information, inclusion of these regions reduces performance of a SV system [85–87]. Even though VAD is a binary classification task, it is very difficult to implement a VAD which works consistently for clean and degraded speech.

2.1.1.1 Speech detection methods

The only requirement at the VAD stage of a SV system is to identify the speech regions. The VAD methods used for many other speech based applications can be used for a SV system. Several VAD methods have been proposed either by using the features those are specific to the speech segments [5–8] or by constructing discrimination models between speech and non-speech regions [9–13]. The features used for the VAD include, energy based features [6], long term speech information [8], zero crossing rate [7], pitch information [88], periodicity [5] and higher order statistics in linear prediction

(LP) residual domain [89]. The main issue is to find an optimal threshold for speech/non-speech discrimination. Since one threshold is not suitable for both clean and degraded speech, VAD fails in most practical applications. The other group of methods uses statistical models like the Hidden Markov model (HMM) [9], GMM [10], Laplacian model [11] and gamma model [90]. Statistical methods aim to construct different classifiers for speech and noise/silence. These methods do not depend critically on the threshold setting, but their performance depends on noise estimation. Most of these methods are initialized based on the assumption that the initial few frames are non-speech frames, which may not be true for all cases. The performance also depends on the choice of probability distribution and speech specific features.

2.1.2 Impact of speech detection on the SV performance

As reported in [85–87, 91–93], the performance of a SV system varies significantly depending on the employed VAD. Specifically, special care is required at the VAD stage for noise degraded speech [86,93] and short-duration speech for the development of a better SV system [91]. The missing-feature approaches presented in [83,94], suggest that for a severely degraded speech, a better SV system can be developed by using only the less degradation affected speech frames.

2.1.3 Selection of similar speech regions

In a *text-independent* SV system distribution of the features derived from training and testing utterances vary depending on the content of sound units [38, 66, 70]. The mismatch due to different sound units can be reduced by using similar speech regions during training and testing of a SV system [66, 70, 81, 95–99].

2.1.3.1 SV by conditioning

In most of the constraint SV systems, different models are developed by identifying similar speech regions in the training utterance. During testing, the test utterance is segmented into similar speech regions and compared against the corresponding models. Finally the scores obtained from each constraint models are combined to take the verification decision [67, 70, 81, 100, 101].

The constraint SV systems are developed either by identifying frequently used words [81, 100, 102, 103], syllable units [67, 70, 101, 104] or different phoneme units [66, 105, 106]. For example, in [104], the speaker-specific GMM is combined with speaker adapted syllable-based HMM. In [70], different constraint GMM-UBM systems are developed using syllabic sub-units, and finally the verification

scores obtained from each subsystems are combined. As reported in this work, the constraint system without any feature normalization performed significantly better compared to standard GMM-UBM system with feature normalization. In [67], during the training phase, a separate phonetic dependent GMM is developed for each phonetic class and in the verification time the GMM corresponding to the frame label was used in scoring. As reported in this work, the performance is improved significantly compared to conventional (non-phonetic) GMM. In [107], the training and testing utterances are conditioned with a multi-grained model structure, where each speaker model is represented at various levels from all phonemes, to broad classes to individual phoneme, allowing the fine detail. Similar multilevel model building is presented in [108]. In the framework of phonetic speaker verification, the contextual information around keywords is used for phone feature vector generation [109]. In the frame work of joint factor analysis (JFA), different strategies for combining the phone constraint SV systems is presented in [110]. This work shows that a full factor analysis technique like JFA also provides performance improvement by using similar phoneme units during training and testing of a SV system.

Although, no literature is available for comparison of different methods used for the selection of constraint regions, it seems that performance of a constraint SV systems depends on the type of constraint and ability for the detection of constraint regions during training and testing. As most of the constraint SV systems requires a speech recognizer at the front-end of the SV system, it increases the computational complexity at the front-end.

2.1.4 Speech enhancement

In SV systems, the speech enhancement is basically applied for removing the effect of high level additive environmental noise at the signal level before feeding the speech signal to the feature extraction module [16, 82–84].

2.1.5 Speech enhancement methods

Since, the speech enhancement is used as a preprocessing method to remove the additive noise from the speech signal, any suitable speech enhancement method which is efficient to remove the effect of additive noise can be used for the SV systems. Several methods have been proposed for enhancement of the speech signals [82, 84, 111–113]. In the temporal speech enhancement methods, enhancement is performed by identifying the high signal to noise ratio (SNR) regions and enhancing them relative

to the low SNR regions [112, 113]. Alternatively, in the spectral domain methods, an estimate of the clean signal is obtained by estimating the noise and removing it in the spectral or cepstral domain from the noisy speech [111, 114–119], by wavelet denoising methods [120–122] or by using statistical model based denoising methods [123–126]. The temporal and spectral processing methods can also be combined for a better speech enhancement [84, 113, 127].

2.1.6 Impact of speech enhancement on SV systems

Each of the speech enhancement methods addressed above have their own merits and demerits depending on performance and complexity. Due to the simplicity and minimal complexity the spectral subtraction method [111] is often used in the SV systems for speech enhancement [14,16,82,83]. In the spectral subtraction method, for i^{th} analysis frame the estimated power spectrum of noise ($|\hat{N}_i(e^{j\omega})|^2$) is subtracted from the power spectrum of noisy speech ($\left|S_i(e^{j\omega})\right|^2$) to estimate the power spectrum of clean speech ($|\hat{S}_i(e^{j\omega})|^2$) as,

$$|\hat{S}_i(e^{j\omega})|^2 = \begin{cases} \left|S_i(e^{j\omega})\right|^2 - |\hat{N}_i(e^{j\omega})|^2 & \text{if} \quad \left|S_i(e^{j\omega})\right| > |\hat{N}_i(e^{j\omega})| \\ 0 \quad \text{otherwise} \end{cases} \tag{2.1}$$

The noise estimation is done based on the the assumption that the background noise is locally stationary, so that the noise characteristics computed from the non-speech regions is good enough for approximation to the noise characteristics. The estimated speech spectrum obtained by spectral subtraction may contain negative values due to the errors in estimating the noise spectrum, which produces musical noise in the enhanced speech. Several modifications for the standard spectral subtraction method have been proposed to alleviate the speech distortion introduced by the spectral subtraction process. [114, 116, 117, 119, 128]. In the context of speaker identification, in [82] the standard spectral subtraction method [111] is compared with spectral subtraction with compensated musical noise [114] and non-linear spectral subtraction [128]. As reported in this work the standard spectral subtraction method though introducing distortion on the recovered speech, it provides better recognition accuracy.

An alternative spectral enhancement method often used for the enhancement of noise speech is the minimum mean square error (MMSE) estimation of the short time spectral amplitude (STSA) [129, 130]. This method assumes that each of the Fourier expansion coefficients of the speech and of the noise process can be modeled as independent, zero-mean, Gaussian random variables. This method aimed to enhance degraded speech by minimizing the mean squared error between the STSA

of the clean speech and the enhanced speech, which was reported to significantly reduce the problem of musical noise by recursively smoothing *a priori* SNR [129]. As reported in [131], in the context of speaker recognition, MMSE log-STSA [130] outperformed compared to the spectral subtraction [111].

For a SV system, although the speech enhancement provides performance improvement for noise degraded speech signal, it increases complexity of a SV system and reduces the performance for clean and less corrupted speech signal by removing relevant speaker specific information from the speech signal [14, 132].

2.2 Feature normalization

The feature vectors extracted from the speech signal along with speaker specific information contains many other redundant information present in the speech signal [38–41, 43–45]. As a result extracted features varies depending on the acoustic environment, sensor, channel, sound unit, language, style and many other factors [3,29,38,39,44,66,81,133]. The impact of these redundant factors on the feature vectors depends on the employed feature extraction method [3, 14, 16]. To the best of our knowledge no feature extraction method is robust enough to extract only the speaker specific information from the speech signal. Therefore, it is required to normalize the feature vectors before feeding them to the speaker modeling and the testing modules of a SV system. Compensation of different variabilities in the feature domain popularly known as the feature normalization. Feature normalization can be performed in unsupervised manner [19,41,44] or in a supervised manner [40,134].

2.2.1 Feature normalization methods

The feature normalization methods basically aim to scale or warp the feature vectors for suppressing the effect of unwanted variabilities. The feature normalization is performed either by modifying certain statistical properties of the speech signal such as mean and variance [19, 135] or by transforming the feature distribution to a reference distribution [41, 45]. The feature normalization is generally performed on the set of feature vectors selected by the VAD [19, 41, 45, 135].

The convolutive channel effect on the speech signal becomes additive in log spectral or cepstral domain. With an assumption that the channel is remained stationary for the entire utterance, the mean of cepstral feature vectors like MFCC and LPCC over the entire utterance represents the channel effect. In cepstral mean normalization or cepstral mean subtraction (CMS), the mean vector (computed over the entire utterance) is subtracted from each of the feature vectors to remove stationary channel effect

[19]. The additive noise present in the speech signal scales the variance of the feature distribution. The variance of feature vectors can be equalized by dividing each feature vector by the standard deviation computed over the entire utterance. When the cepstral variance normalization (CVN) is applied with the CMS, the feature distributions become a zero mean and unit variance distribution [135]. In the cepstral mean and variance normalization (CMVN), n^{th} feature frame and k^{th} feature space $X(n,k)$ is normalized as,

$$X_{mvn}(n,k) = \frac{X(n,k) - \boldsymbol{\mu}_k}{\boldsymbol{\sigma}_k}. \tag{2.2}$$

Where n and k are the frame index and feature index, respectively. $\boldsymbol{\mu}_k$ and $\boldsymbol{\sigma}_k$ are the mean and standard deviation of the k^{th} component of the feature vectors, respectively.

The global CMS and CVN are not effective to remove the non-stationary channel effect. To capture the time varying properties of the channel, in segmental mean and variance normalization (SCMVN), the mean and variance of the feature vectors are updated over a sliding window [41, 135, 136]. The feature vector to be normalized is located at the center of the sliding window. In this approach, the window should be short enough to capture the time varying properties of the channel and at the same time it should contain sufficient number of feature frames for a good estimates of mean and standard deviation. Normally 3-5 s window is used for SCMVN [3, 41, 136]. The experiment presented in [135] shows that CMVN provides improved performance compared to the CMS and the performance can be further improved by applying SCMVN. Similar to SCMVN, a feature normalization method called as short-time cepstral mean and scale normalization (STMSN) is prosed in [136]. In this approach, mean is normalized similar to SCMVN normalization, where as the variance is normalized by finding the upper and lower bound of the feature components. In STMSN, n^{th} feature frame and k^{th} feature space $X(n,k)$ is normalized as,

$$X_{stmsn}(n,k) = \frac{X(n,k) - \boldsymbol{\mu}_{st}(n,k)}{\boldsymbol{d}_{st}(n,k)}. \tag{2.3}$$

where $\boldsymbol{\mu}_{st}(n,k)$ and $\boldsymbol{d}_{st}(n,k)$ are the short-time mean and short-time difference between the upper and lower bound for a window of L feature frames. Short-time mean and short-time difference are computed as,

$$\boldsymbol{\mu}_{st}(n,k) = \frac{1}{L} \sum_{i=n-\frac{L}{2}}^{n+\frac{L}{2}} X(i,k) \tag{2.4}$$

$$\boldsymbol{d}_{st}(n,k) = \underset{n-\frac{L}{2}\leq i\leq n+\frac{L}{2}}{\mathbf{max}}(X(i,k)) - \underset{n-\frac{L}{2}\leq i\leq n+\frac{L}{2}}{\mathbf{min}}(X(i,k)) \tag{2.5}$$

This work shows that in context of an i-vector based SV system the STMSN provides similar performance compared to the SCMVN. A nonlinear sensor or channel responds differently to different energy regions of the speech signal. Thus, the verification performance can be improved by using SNR dependent CMS [61,137]. In this approach separate CMS is performed for different energy ranges. As reported in [61], SNR dependent CMS does not provide better performance compared to the CMS for handset mismatched experiments.

The other filtering approach used for the removal of channel effect is the RelAtive SpecTrAl (RASTA) processing method [44, 138]. In RASTA filtering method a band pass filter is applied in the log spectral or cepstral domain to suppress the spectral components that changes very slowly or quickly than the typical range of the speech signal. The RASTA filter attenuates modulation frequency components below 1 Hz and above 10 Hz. Therefore it suppress the slow varying convolutive channel effect and also the fast varying modulation frequency components. The RASTA filtering method is signal independent, where in the CMVN the normalization parameters are signal dependent. But in standard RASTA processing method the specified lower cut-off frequency removed significant portions of speaker specific information [41]. According to [138], in the context of PLP feature, RASTA filtering outperformed compared to CMS. As reported in [134], both CMS and RASTA filtering can be combined for better SV performance.

The convolutive channel effect shift mean of the feature distribution and the additive noise due to acoustic environmental degradations scale variance of the feature distribution [41,45]. Both the channel and the environmental effect can be suppressed by modifying the short-time feature distribution to follow a reference distribution, for instance standard normal distribution ($N\ (0,1)$) [41, 45]. In [41], assuming the component of feature vector are independent, each component is processed as a separate stream. A given feature is warped so that its cumulative distribution function (CDF) matches a desirable distribution. The feature warping can be viewed as nonlinear transformation $\boldsymbol{\Gamma}$ from the original feature (X) to warped feature (X') as,

$$X' = \boldsymbol{\Gamma}(X) \tag{2.6}$$

The CDF matching is performed over a sliding window of size L feature frames. Only the central frame of the window is warped based on CDF. This process is repeated for each frame shift. To

determine the warped feature element, the features in the sliding window are shotted in descending order. For window of L feature frames, if the central frame has a rank r (1 to L), then its corresponding CDF value is approximated as,

$$\psi = \frac{L + \frac{1}{2} - r}{L} \tag{2.7}$$

The warped feature is determined by finding the warp value x'

$$\psi = \int_{-\infty}^{x'} f(y)dy \tag{2.8}$$

Where $f(y)$ is the probability density function (PDF) of standard normal distribution.

$$f(y) = \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-y^2}{2}\right) \tag{2.9}$$

According to [41, 135], feature warping slightly outperformed compared to SCMVN but it is computationally more expensive.

A variant of the feature warping, called as short-time Gaussianization (STG) is proposed in [45]. In this method, first a global linear transformation is applied to the features before mapping them to a normal distribution. Linear transformation in the feature space makes resulting features independent. For the original feature set (X) and linear transformed matrix $\mathbf{\Lambda}$, the STG can be implemented using two steps. First the original features are linearly transformed as,

$$Y = \mathbf{\Lambda} X \tag{2.10}$$

Then, short-time windowed feature warping $\boldsymbol{\Gamma}$ is applied on the Y as

$$X' = \boldsymbol{\Gamma}(Y) \tag{2.11}$$

In STG each feature vector is warped independently. The STG performed better compared to feature warping [45].

A supervised feature normalization method called as feature mapping is proposed in [40]. This method basically aims on transforming the feature obtained from different channel conditions to a common channel independent feature space. A channel independent root GMM is trained by using large amount of speech data from different channel conditions. The channel dependent models are developed by adapting the root GMM using the channel dependent speech data. For a given input

utterance, first the most likely channel dependent model is detected. Then, each feature vector is mapped to the channel independent space by finding the top scoring Gaussian in the channel dependent GMM. The mapping of a feature X to a channel independent feature Y as,

$$Y = (X - \mu_j^{CD})\frac{\sigma_j^{CI}}{\sigma_j^{CD}} + \mu_j^{CI} \tag{2.12}$$

Where μ_j^{CD} and σ_j^{CD} are the mean and diagonal covariance of the j^{th} mixture component of the channel dependent GMM (top scoring Gaussian component for the given feature vector X). μ_j^{CI} and σ_j^{CI} are the corresponding mean and diagonal covariance of the root GMM. This method showed good performances in channel compensation but it required large amount of different channel data to train the channel dependent models. The extended version of feature mapping method are presented in [139, 140]. The method proposed in [139] uses an iterative clustering approach for effective feature mapping models in the absence of handset and channel dependent data for building the root model. In [140] a channel dependent piecewise linear transformations used for feature mapping, which does not require detection of the top-1 Gaussian component.

It is also possible to combine different feature normalization methods for better normalization of feature vectors [49, 134, 141, 142]. For instance, in [142], both MFCC and LPCC feature vectors are normalized using combination of RASTA filtering and CMVN. In [141] combination of RASTA filtering, feature mapping and CMVN is used. Combination of RASTA filtering, STG and heteroscedastic linear discriminant analysis (HLDA) is used in [49]. HLDA mostly used for speech recognition systems. It applies a linear transformation on the feature vectors that reduces dimensionality while preserving the discriminative ability of features [143]. In [134], combination of CMS, RASTA filtering, feature warping, HLDA and feature mapping is used. In these works, performance of individual normalization methods are not presented. However, as reported in [134], combination of different feature normalization methods provides significant performance improvement. Out of the different feature normalization methods discussed, as a single feature normalization method, most of the SV system used SCMVN due to less complexity and nearly similar performance compared to different complex feature normalization methods [41, 135, 136].

2.3 Session/Channel variability compensation

In most of the state-of-the-art SV systems, the speech utterances having different number of feature vectors are represented by a fixed large dimensional vector, called as supervector [46, 47, 144]. A supervector is created by mapping all feature vectors of an utterance to a large fixed dimensional vector using a kernel function. Depending on the kernel function there are different supervector representations [36, 46, 47, 144, 145]. For instance, Gaussian supervector for a speech utterance is created by concatenating the mean vectors of the adapted GMM [36, 145, 146]. For a C component adapted GMM and F dimensional feature vector, the speech utterances of variable length are represented by CF dimensional supervectors. In a supervector representation, speech utterances of variable length are represented as a single point in the supervector space. Therefore for a given speaker, any variation between the speech utterances recorded in different sessions can be compensated by finding variation between them in supervector space [36, 46, 47, 145, 146]. This type of normalization is popularly known as inter-session compensation. It is difficult to avail multiple sessions speech data for each enrolled speaker. The intersession compensation is performed using a similar development data set [49, 50, 80]. The speakers in the development data set are independent to the enrolled speakers. In a *text-independent* SV system, the variation between different sessions of speech utterances arises from speaker dependent factors (health, emotion, etc), phonetic content and due to different degradation effects. Therefore, the explicit inter-session compensation helps to remove different unwanted variabilities not compensated at feature normalization.

2.3.1 Session/Channel variability compensation methods

As explained above, for inter-session variability compensation, it is required to model explicitly the session/channel variabilities. Different methods have been proposed for this purposed [36,37,50,80]. In JFA framework a Gaussian supervector is considerered as a linear sum of components from a speaker and a session/cahannel subspaces [36, 42, 146]. The supervector M_s representing a speech utterance is decomposed as,

$$\boldsymbol{M}_s = \boldsymbol{s} + \boldsymbol{c} \tag{2.13}$$

where $\boldsymbol{s}$ and $\boldsymbol{c}$ represent the speaker and session/channel supervector, respectively. For a C component adapted GMM and F dimensional feature vectors, $\boldsymbol{s}$ and $\boldsymbol{c}$ are statistically independent and normally distributed CF dimensional supervectors.

The speaker supervector of a randomly selected speaker is represented as,

$$\boldsymbol{s} = \boldsymbol{m} + \boldsymbol{V}y + \boldsymbol{D}z \tag{2.14}$$

Where $\boldsymbol{m}$ is the UBM mean supervector of dimension $(CF \times 1)$, $\boldsymbol{V}$ is the rectangular matrix (eigen voice matrix) of dimension $CF \times R_s$, which span a subspace of low rank R_s, $\boldsymbol{D}$ is the diagonal matrix of the factor analysis model of dimension $CF \times CF$. y and z are normally distributed random vectors. The component of y and z describe the speaker factor.

The session/channel variability is model as,

$$\boldsymbol{c} = \boldsymbol{U}x \tag{2.15}$$

Where $\boldsymbol{U}$ is a rectangular matrix (eigen channel matrix) of dimension $CF \times R_c$, which lies in the low dimensional subspace R_c. The component of x represents the session/ channel factor, where x is normally distributed random vector.

In JFA, first the subspaces corresponding to the speaker and the session/channel $(\boldsymbol{V}, \boldsymbol{D}, \boldsymbol{U})$ are estimated using suitable development data set. Then, the speaker and channel factors $(y, z, x\ )$ are estimated for the given speech utterance. Finally, the channel supervector $\boldsymbol{c}$ is discarded and the speaker supervector $\boldsymbol{s}$ is used for scoring. Comparison of different JFA scoring methods is given in [147]. The JFA provides significant performance improvement through session/channel variability compensation [36, 148].

In nuisance attribute projection (NAP) method, with a assumption that session/channel variability lies in a speaker independent low dimensional subspace, the session/channel variabilities compensation is done by finding the nuisance subspace dimensions [48, 49, 80]. In this method an appropriate projection matrix is estimated to remove the nuisance (session/channel) directions from a supervector. The speech utterances projected to the complimentary space of the nuisance subspace for session/channel variability compensation. The projection matrix $\boldsymbol{P}$ is given as

$$\boldsymbol{P} = \boldsymbol{I} - \boldsymbol{R}\boldsymbol{R}^t \tag{2.16}$$

where $\boldsymbol{R}$ is a rectangular matrix containing the eigen vectors corresponding to few of the top eigen values of the within-class covariance matrix. The projection matrix is learned using a development dataset with sufficiently large number of speakers, each having several sessions recording. After train-

ing the projection matrix, the session/channel compensation is applied on the given speech utterance supervector.

As reported in [49,80], session/channel compensation using NAP provides significant performance improvement. According to [149], the nuisance (session/channel) dimensions removed by NAP contain speaker specific information. To maintain speaker variability, scatter difference analysis (SDA) method is used for optimizing the NAP projection matrix. As reported in [149], a modest improvement is observed by using SDA for NAP training over the standard NAP baseline.

In [37,150,151], the linear discriminant analysis (LDA) is used for the session/channel compensation. In this approach, the feature vectors are projected down to a set of new orthogonal axes where the intra-class variance caused by the channel is minimized and inter-class variance is increased. The projection matrix is composed of the eigen vectors corresponding to the best eigen values of the eigen analysis equation as,

$$(\boldsymbol{W_c^{-1}B_c})\boldsymbol{v} = \boldsymbol{\lambda v} \tag{2.17}$$

where $\boldsymbol{W_c}$ is the within-class covariance matrix, $\boldsymbol{B_c}$ is the between-class covariance matrix, $\boldsymbol{v}$ is an arbitrary vector, and $\boldsymbol{\lambda}$ is the diagonal matrix of eigen values. The LDA projection matrix $\boldsymbol{A}$ composed by the best eigenvectors (those having highest eigenvalues). Similar to NAP projection matrix, the LDA projection matrix is learned using a development dataset with sufficiently large number of speakers, each having several sessions recording. After training the projection matrix, the session/channel compensation is done by projecting the given speech supervector to the low dimensional subspace using the LDA matrix.

Unlike NAP and LDA, where the nuisance dimensions are removed by projecting the supervectors to a lower dimensional subspace, a method called as within class covariance normalization (WCCN) is proposed to weight instead of completely removing the dimensions [50,152]. In this approach, a set of upper bounds are defined on the classification error metric to reduce the error rate. The feature vectors are transformed using a matrix which minimizes the upper bounds on the classification error metric and hence minimizes the classification error. The transformation matrix $\boldsymbol{B}$ is obtained by Cholesky decomposition of the inverse of the within-class covariance matrix $\boldsymbol{W}$ as,

$$\boldsymbol{W}^{-1} = \boldsymbol{B}\boldsymbol{B}^t \tag{2.18}$$

Comparison of different session/channel compensation methods are presented in [37,148,150]. Accord-

ing to [37, 150], the channel factors estimated in JFA also contains speaker information. As reported in [148], JFA without speaker factors gives similar compared to SVM both with linear and nonlinear kernels compensated with NAP. The JFA provides better performance compared to SVM when speaker factors are added. It is also possible to combine different session/channel compensation methods for better compensation of nuisance factors [37, 150]. The experiment presented in [37] shows that for an i-vector based SV system, the best performance is obtained when LDA is followed by WCCN compared to each individual session/channel compensation methods and combination of NAP and WCCN.

2.4 Speaker model compensation

These approaches are aimed to modify the component of the speaker model to the testing condition instead of compensating the degradation effect from the feature vectors [39, 68, 69, 153–155].

2.4.1 Speaker model compensation methods

Assuming the availability of statistical model of noise or environment, a speaker model compensation method called as parallel model compensation is used in [68, 153, 154]. The background noise model is generated using the available noise samples. The clean models and noise model are then combined in the log-spectral domain to obtain the best possible estimate of the degraded speech models. Similar parallel model compensation, the other model compensation method called as Jacobian environmental adaptation is used in [156, 157]. Although these methods provide performance improvement under degraded conditions, *a priori* information about the testing conditions is required and may not be possible under all practical scenarios.

A method based on multi-condition training and missing feature theory is proposed to improve the matching between speaker model and test features [39]. The multi-conditioned training is done by adding simulated white noise with different SNR values to the clean speech. The missing feature theory approach is used to ignore noise variations outside the given training conditions. The experiments presented in [39] shows that for both identification and verification task, the multi-conditioned training provides significant performance improvement compared to clean training condition, and the multi-condition training with missing feature theory provides best performance. In [158], multi-conditioned training is performed using single SNR colored noise. The colored noise samples are generated by filtering a white Gaussian sequence that leads to a power spectral density proportional to $1/f^a$, where

$a \in [0\ 2]$. As reported in [158], for speaker identification task the color noise based multi-conditioned training provides better performance compared white noise multi-conditioned training. For clean speech, both the approaches performed poorer compared to clean training condition.

A better compensation of channel effect can be achieved when the channel of the background model matches to that of the speaker [159]. Based on this assumption, a method called as speaker model synthesis (SMS) is proposed to synthesize the speaker models from a unknown channels [69]. This approach can be viewed as a model domain approach of feature mapping method [40]. Similar to the feature mapping method, a channel independent root GMM is trained by using large amount of speech data from different channel conditions. The channel dependent background models are developed by adapting the root GMM using the channel dependent speech data. The speaker models are created by adapting corresponding channel dependent model. Since all models are developed by adapting a common root model, there is a correspondence between Gaussian components in the root model and channel dependent models. For a channel without corresponding enrollment data, a model is synthesized for that enrolled speaker. The model synthesis is done by finding the mean shift, variance scale and weight scale. The parameters of the $i\ th$ Gaussian component of the speaker model in Channel-1 $(\mu_{i,c_1}, \sigma_{i,c_1}, \eta_{i,c_1})$ is transformed to the corresponding component in channel-2 $(\widetilde{\mu}_{i,c_2}, \widetilde{\sigma}_{i,c_2}, \widetilde{\eta}_{i,c_2})$ as,

$$\begin{aligned} \widetilde{\mu}_{i,c_2} &= \mu_{i,c_1} + (\mu_i^{CD2} - \mu_i^{CD1}) \\ \widetilde{\sigma}_{i,c_2} &= \sigma_{i,c_1}(\tfrac{\sigma_i^{CD2}}{\sigma_i^{CD1}}) \\ \widetilde{\eta}_{i,c_2} &= \eta_{i,c_1}(\tfrac{\eta_i^{CD2}}{\eta_i^{CD1}}) \end{aligned} \tag{2.19}$$

where $(\mu_i^{CD1}, \sigma_i^{CD1}, \eta_i^{CD1})$ and $(\mu_i^{CD2}, \sigma_i^{CD2}, \eta_i^{CD2})$ are the parameters of the i^{th} Gaussian component in channel dependent root model. During testing, the most likely channel dependent model is detected by an automatic method and the likelihood ratio is computed with respect to the corresponding channel dependent speaker model. As reported in [40], feature mapping and SMS provides similar performance in context of channel mismatch. A variant of SMS called as cohort based SMS is proposed in [155]. New speaker model is synthesized by using a set of cohort speaker models. Cohort based SMS utilizes speaker specific cohort subsets as *a priori* knowledge of channels for speaker model synthesis. As reported in this work, the cohort based SMS outperformed compared to background model based SMS.

2.5 Score normalization

In the final stage of a SV system, the verification score obtained from the claimed model is compared against a decision threshold to accept/ reject the claim. The speech data used for the development of model and testing varies between the speakers. Secondly, for the same speaker quality and quantity of test data varies between the trials. As a result, the verification score varies between the trials. Compensation of different variabilities at the score level is commonly known as score normalization [2, 3]. The score normalization helps to reduce degradation and mismatch effect that are not compensated at feature and model levels. It also transforms scores from different trials into a similar range so that a common speaker independent verification threshold can be used [2, 3, 59, 60].

The score normalization methods basically aim to scale the overall score distribution (both target and impostor score distributions), most commonly to a zero mean and unit variance distribution [59, 60, 160] as,

$$\boldsymbol{S}_n = \frac{\boldsymbol{S} - \boldsymbol{\mu}_I}{\boldsymbol{\sigma}_I}. \tag{2.20}$$

where $\boldsymbol{S}_n$ is the normalized score, $\boldsymbol{S}$ is the original score and $\boldsymbol{\mu}_I$ (mean) and $\boldsymbol{\sigma}_I$ (standard deviation) are the normalization parameters. In most of the practical applications of a SV system, it is difficult to avail enough speech utterances for the target speakers to estimate the target score distribution. Therefore, the normalization parameters are generally estimated from the impostor score distribution. The set of speech files used for the estimation of the normalization parameters may be speaker dependent or speaker independent [2,3]. Score normalization by using speaker dependent normalization parameters provides better performance compared to the speaker independent normalization parameters [2,3,60].

2.5.1 Score normalization methods

Since from Li and Porter observation [160], several score normalization methods have been proposed. In [161, 162], score normalization is done in the form of likelihood ratio as,

$$\boldsymbol{S}_n = \frac{\boldsymbol{S}_\lambda}{\boldsymbol{S}_{\bar{\lambda}}} \tag{2.21}$$

where $\boldsymbol{S}_\lambda$ and $\boldsymbol{S}_{\bar{\lambda}}$ are the likelihood for the claimed speaker model and cohort speaker models, respectively. For each target speaker, the close cohort speakers are selected offline in the training process. To avoid cohort variation and reduce the computational load, the cohort speaker models later replaced by a unique background model [163–165]. The background model is generally developed by pooling

similar speech data from a large set of speakers. These normalization methods relay more on the estimation of anti-speaker hypothesis (test speech not spoken by the speaker in the target model) in the Bayesian hypothesis test. A score normalization method based on background model and *a posteriori* probability (WAMP) is proposed in [166]. In this method, the similarity score ratio is further normalized as *a posteriori* probability using the Bayes' rule.

In most of the modern SV systems, both target and impostor scores are normalized using Eq. (2.20). Depending on different possibilities for the estimation of the normalization parameters (μ_i and σ_i), several score normalization methods have been proposed. In zero normalization (Z-norm), a set of impostor utterances are tested against the speaker model to obtained the normalization parameters [60, 167]. The normalization parameters are speaker model dependent and they are computed offline during the training process. Generally, impostor utterances from the cohort speaker set provides better performance compared to the generic impostor set [2, 3]. The verification score varies differently depending on the handset used for recording the test speech data [61, 159]. The performance of a SV system can be improved by using the handset information in score normalization. To deal with the handset mismatch between training and testing speech data, a handset dependent Z-norm method, called as the handset normalization (H-norm) is proposed in [35]. In this method, handset dependent normalization parameters are estimated by testing each speaker model against different possible handset dependent speech utterances from the impostors. In the testing process, the most likely handset is detected automatically and the corresponding normalization parameters are used for score normalization.

A different method for the estimation of the normalization parameters, called as the test normalization (T-norm) is proposed in [60]. The T-norm is quite similar to the likelihood normalization based on the cohort speakers [161, 162]. In this method, the normalization parameters are estimated using impostor models instead of impostor test utterances. During the verification process, the test speech is simultaneously compared against the claimed model and a set of impostor models to estimate the normalization parameters. The score obtained from the claimed model is normalized using the normalization parameters estimated the impostor score distribution. This type of normalization works better when the number of impostor models are large [60]. The T-norm can be applied using a common set of impostor models or speaker dependent cohort models. The T-norm provides better result for the speaker dependent impostor models [59, 168]. In [168], based on the city-block vector distance,

a speaker adaptive cohort selection method for T-norm (known as AT-norm) is proposed. For a given set of impostor utterances and T-norm models, the impostor utterances are scored against all the T-norm models and the speaker model. Then, the city-block vector distance is used to select a subset of the nearest T-norm model as the cohort of the speaker model. Similar to AT-norm, an approximation of Kullback-Leibler (KL) divergence is used for the distance measure in the KL-T-norm [59]. The KL divergence between the speaker model and a large set of T-norm models are computed to find the nearest cohort speaker set. Both AT-norm [168] and KL-T-norm [59] outperformed compared to the T-norm. In [169] an adaptive T-norm method is proposed for normalizing score drift in the progressive model adaption. In this method, whenever a speaker model is adapted, the corresponding T-norm speaker models are also adapted using the utterances from T-norm speakers. As reported in [169], the adaptive T-norm performed better compared to the T-norm. Like H-norm, handset dependent T-norms (HT-norm) provides better verification result [62].

The T-norm provides better performance compared to the Z-norm [60]. According to [170], T-norm improves the SV performance by reducing the false acceptance rate. Alternatively, T-norm increases the verification time, since the normalization parameters are computed during verification process (on-line). Z-norm and T-norm can be combined for better normalization of verification score [167]. As reported in [167], Z-norm followed by T-norm (ZT-norm) provides better verification result. The performance of Z-norm and T-norm based methods depends on the selection of impostor speakers for the estimation of normalization parameters [2, 3]. In [171], Distance normalization (D-norm) is proposed to deal with the problem of availability of pseudo-impostor data. A Monte-Carlo based symmetric KL distance is used to obtain a set of target speaker and impostor data using target model and background model, respectively. The main advantage of D-norm is that it does not require any normalization data in addition to the background model. According to [171], the D-norm performs slightly poorer compared to Z-norm.

Out of the different normalization methods discussed above most of the modern SV systems use ZT-norms for their better SV accuracy [36,37,42]. According to [42,172], even in case of the complete factor analysis model like JFA, Z-norm, T-norm and their combinations is essential to remove the variabilities from the verification score. Although Z-norm and T-norm can be effective in reducing the variabilities in the verification score, their performance depends on the selection of cohort speakers and other side information like recording handset, channel and environmental conditions. For instance,

the H-norm and the HT-norm are only applicable when the number of possible hand is known *a priori* and large number of utterances from each handset is available. Performance of these methods depend on the automatic detection of handset type which may make classification error. Similarly, these normalization methods may seriously fail if the cohort speakers are badly selected [3]. The SV performance also depends on the similarity in the recording condition of the cohort speakers to that of the training and testing condition of the SV system. For a SV under degraded condition, it is difficult to avail cohort speakers speech utterances from the similar recording conditions. Therefore the score normalization methods may not help much to remove the degradation effect, which needs to be studied to understand the effect of score normalization under degraded conditions.

2.6 Summary and scope for present work

In this chapter, a brief review is presented addressing different approaches for SV under degraded conditions. The literature shows that the accuracy of SV system falls significantly in the presence of background noise, sensor and channel mismatches. The phonetic variability between training and testing utterance is another major source for reducing the performance of a SV system. Several efforts have been made starting from the front-end signal analysis to score normalization for the development of a robust SV system. Although many results in the literature are encouraging, most of them focused on a particular type of mismatch. For instance, feature normalization, session/channel variability compensation and score normalization methods are explicitly used for the compensation of sensor and channel variabilities between the training and testing speech signals. Most of these methods fail for higher level of degradations. The speaker model compensation methods provide improved performance for noise degraded speech signal. But performance of these methods depends on the *a priori* information about the testing conditions, and most of these methods provide poorer performance for clean and less corrupted speech signal.

A particular method may not compensate different variabilities present in the speech signal. For example, the full factor analysis method like JFA with feature normalization requires score normalization to further compensate the variabilities [42, 172]. The performance of a compensation method used in particular module depends on the accuracy of the methods used in the previous modules. For example, the performance of an i-vector based system with session/channel compensation depends on the feature normalization [136].

For any practical application of a *text-independent* SV system, along with phonetic variability, the speech signal may be affected by background noise, sensor and channel mismatches. In such a scenario a better SV system for clean and degraded speech signal may be possible by compensating mismatch due to sound units and degradation effect. Without knowledge of degradation effect, this may be achieved by modifying the front-end signal analysis module to tackle the degradation effect and phonetic variability. The feature normalization and channel/session variability compensation methods can be used for further reducing the variabilities.

The existing SV systems mostly uses a VAD for selection of speech frames. The performance of a SV system depends on the robustness of VAD, specifically for the noise degraded speech signal [85–87, 91–93]. Even selection of speech regions by a robust VAD may not reduce the degradation effect for highly corrupted speech signals [39, 83, 94] and the phonetic variability between the training and testing speech signals [66, 67, 70, 81, 100–106]. A few attempts have been made to develop a better SV system for severely degraded speech signal by neglecting more corrupted speech frames [83, 94]. Although performance is improved under degraded conditions, the performance always reduces for clean and less corrupted speech signal by neglecting speaker specific speech frames, and these methods may not reduce the variabilities due to sound units. The methods explicitly used for reducing the text variability between the training and testing speech signal by conditioning [66, 67, 70, 81, 100–106] may not help to reduce degradation effect. These methods mostly uses complex speech recognizer at the front end which increases complexity of SV system. The SV performance also depends on the type of conditioning and ability for the detection of condition regions during training and testing process. The speech enhancement methods used for the removal of degradation effect, provide performance improvement for noise degraded speech signal, but most of these methods reduces the performance for clean and less corrupted speech signal by removing relevant speaker specific information from the speech signal [14, 132].

Without neglecting the speaker specific speech frames, the degradation effect and variability due to sound units may be minimized up to an extent by classifying the speech signal depending on SNR and speaker specific information. The degradation effect may be minimized by emphasizing the less degraded speech regions and variability due to sound units by processing similar speech regions during training and testing of the SV system. This may be achieved by segmenting the speech regions into two classes, which avoids complexity involved in a front-end speech recognizer.

In the proposed SV system using VLRs and non-VLRs, the VLRs in the speech signal are high SNR regions [71, 72] and less affected under different degraded conditions. By emphasizing these regions the impact of degradation effect on the feature vectors may be reduced. The independent processing of VLRs and non-VLRs will maintain conditioning between the training and testing utterances. This will help to reduce the phonetic variability in a *text-independent* SV system. Under degraded conditions, better compensation of degradation effects may be achieved by applying different normalizations to these two different segment types [74]. To address all these issues, the investigations in this thesis are planned as,

(i) Chapter 3 explores the significance of processing less degradation affected VLRs for speaker verification under clean, sensor mismatch and noise degraded speech signal. The performance of a SV system using VLRs depends on the robustness of the detection of VLRs. For this, a VLR detection algorithm is proposed using the VLR-onset point (VLOPs). The VLROPs are detected using two robust excitation source features namely HE of the LP residual [78] and zero frequency filtered signal [173]. The GMM-UBM based SV systems are developed using only VLRs and speech regions detected by an energy based VAD. The feature vectors are normalized using CMVN to reduce the channel effect. The verification scores are normalized in the score domain using T-norm. The performance of both the systems are compared for clean, sensor mismatch and noise degraded speech signal to demonstrate the significance of VLRs for speaker verification task.

(ii) Chapter 4 explores the merit of conditioning and emphasis of less degraded speech regions by independent processing of VLRs and non-VLRs. The VLROPs and VLR end points (VLREPs) are hypothesized and used in an iterative algorithm for detecting the VLRs. Next, for detection of non-VLRs, the LP residual samples in the VLRs are attenuated significantly to indirectly emphasize the residual samples in the non-VLRs. The modified LP residual samples excite the time varying all pole filter to reconstruct non-VLRs enhanced speech and used for detecting non-VLRs. To maintain conditioning, GMM-UBM based and *i*-vector based SV systems are developed by using VLRs and non-VLRs independently during training and testing of the SV systems. To emphasize the less degraded speech regions, the verification scores obtained from VLRs and non-VLRs are combined with higher weight on VLRs. For both the systems, feature vectors are normalized in the feature domain using CMVN. For the *i*-vector based SV systems,

the session/channel compensation is performed using LDA and WCCN. The performance of both the systems are evaluated for clean, sensor mismatch and noise degraded speech signal with and without VLRs and non-VLRs conditioning to demonstrate the significance of VLRs and non-VLRs conditioning for SV task.

(iii) Chapter 5 explores a VLRs and non-VLRs conditioned SV system by implicitly modeling of these regions in the *i*-vectors to avoid the complexity involved in the explicit segmentation of speech regions during training and testing of the SV system. The performance of implicitly conditioned SV system is evaluated for clean and noise degraded speech signal to demonstrate the significance of implicit conditioning.

(iv) Chapter 6 summarizes the contributions of this thesis towards development of a speaker verification system under degraded conditions and outline future research directions made possible by the present work.

3

Speaker verification using vowel-like regions

Contents

Objective

In conventional approach, the speech regions are separated from the non-speech regions and features extracted from the speech regions are processed for speaker verification. The objective of this chapter is to demonstrate that a better speaker verification system can be developed under degraded conditions by using features only from the more speaker specific and less degradation affected speech regions. This can be achieved by using knowledge of vowel-like regions (VLRs). VLRs have impulse-like excitation and therefore information about the vocal tract system may be better manifested in them. Also the VLRs are relatively high signal to noise ratio (SNR) regions. Speaker information extracted from VLRs may therefore be more speaker specific and relatively less affected under degraded conditions. Due to this, better speaker modeling and reliable testing may be possible. A method is developed for the detection of VLRs using the VLR onset points VLROPs. VLROP is defined as the instant at which onset of the VLR takes place.

3.1 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 [3]. 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 [39, 68, 174]. 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 [35, 36, 40, 41, 43–45, 50, 60, 61, 144]. In the second group, the parameters of the speaker model are biased towards the testing environment to match the testing conditions [39, 68, 69, 174].

In the first group of techniques, the compensation is done at the signal level, feature level, model 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) [112, 113] or signal to reverberation ratio (SRR) regions and enhance them in the time domain [175] or estimate the noise and subtract in frequency domain [111] or estimate

reverberation and eliminate the same in the cepstral domain [127, 175]. The methods used to remove the degradation effect in feature level include filtering techniques like, cepstral mean subtraction (CMS) [43], relative spectral (RASTA) filtering [44], and various feature transformed techniques like cepstral variance normalization (CVN), feature mapping [40], feature warping [41] and short-term Gaussianization [45]. The methods used for compensation of degradations effect at the model level include joint factor analysis (JFA) [36], nuisance attribute projection (NAP) [144] and within class covariance normalization (WCCN) [50]. The score domain methods include, various score normalization techniques like Hnorm [35, 61], Tnorm [60] and HTnorm [62]. Performance of these methods depend on estimation of degradation effect and availability of suitable development data. These methods are suitable for the lower level of degradations like channel and sensor mismatches. In the second group of techniques, methods like speaker model synthesis (SMS) [69], parallel model combinations (PMC) [68], multi-condition model training [39] and microphone array [174] are used for adapting model parameters to the testing environment. Such type of techniques require *a priori* information about the testing conditions 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 point (VLROP). 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 the work presented in this chapter. The proposed approach is motivated from the earlier studies on using the high SNR or SRR regions from the production perspective for speech enhancement [112, 113, 127, 175].

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 [77, 176], 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 [79]. Majority of the errors are observed to be due to the cases of semivowels and diphthongs [79]. However, for the SV 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 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 [72]. 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 system under degraded condition.

In conventional 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 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 VLRs from the production perspective to different degradations.

The rest of the Chapter is organized as follows: Methods for the detection of VLROPs and VLRs are described in Section 3.2. Proposed speaker verification system using VLRs is described in Section 3.3. The experimental studies are described in Section 3.4. The experimental results are discussed in Section 3.5. The summary of the work are mentioned in Section 3.6.

3.2 Detection of VLROP and VLRs in Degraded Speech using excitation source information

The VLRs are prominent regions in the speech signal due to high amplitude, periodicity, long duration and less 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 [75], zero-crossing rate, energy and pitch information [76], training neural network with the trends in energy, zero crossing rate and spectral flatness at the VOP [77] and using excitation source information [78]. A combined method using the excitation source, spectral peaks and modulation spectrum information is proposed for the detection of VOP [79]. 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. 3.1 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 the linear prediction (LP) residual is relatively less affected by various degradations and hence a pitch extraction method under adverse conditions is proposed in [177]. Also a VOP detection method is exclusively developed using this information in [78]. 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

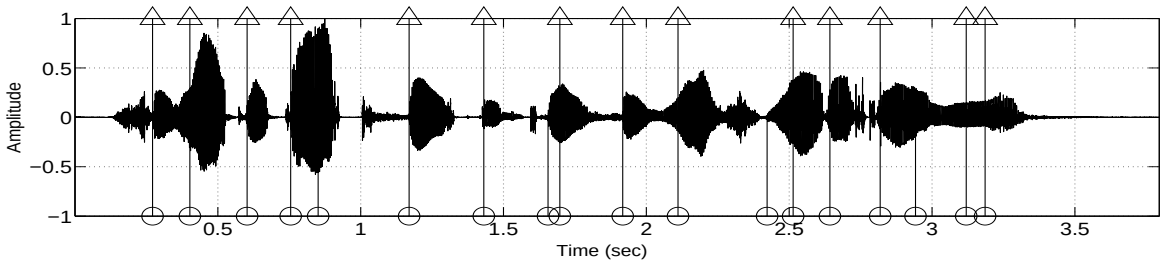

Figure 3.1: Speech signal of the text *she had your dark suit in greasy wash water all year* taken from TIMIT database with reference VOPs (arrows) and reference VLROPs (circles).

enhanced periodicity information. A zero frequency filtering (ZFF) approach is proposed for detecting epochs in speech [173]. Since, the zero frequency resonator exploits only signal energy around zero frequency region and attenuates all other information [173], the resonator output signal may provide robustness to various degradations. The strength of excitation derived from ZFF signal (ZFFS) has been demonstrated earlier to have better discriminating ability at the unvoiced-voiced transitions [178]. It is therefore proposed that by combining the features derived from the HE of the LP residual with features from the 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*.

3.2.1 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, 10^{th} order LP analysis is performed to estimate the linear prediction coefficients (LPCs) [179]. 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 [78].

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

$$e_a(n) = e(n) + je_h(n) \tag{3.1}$$

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

$$e_h(n) = IDTFT(E_H(\omega)), \tag{3.2}$$

where

$$E_H(\omega) = \begin{cases} +jE(\omega), -\pi \leq \omega < 0 \\ -jE(\omega), 0 \leq \omega \leq \pi \end{cases} \tag{3.3}$$

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.,

$$h_e(n) = |e_a(n)|\,. \tag{3.4}$$

$$h_e(n) = \sqrt{e^2(n) + e_h^2(n)}. \tag{3.5}$$

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

$$\phi(n) = \tan^{-1}\left(\frac{e_h(n)}{e(n)}\right). \tag{3.6}$$

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) [78, 79]. This convolved output is termed as *VLROP evidence from HE of LP residual*. Fig. 3.2 shows a portion of speech signal shown in Fig. 3.1, 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. 3.2(d) and sufficient for finding the VLROP evidence.

3.2.2 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 [173]. 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

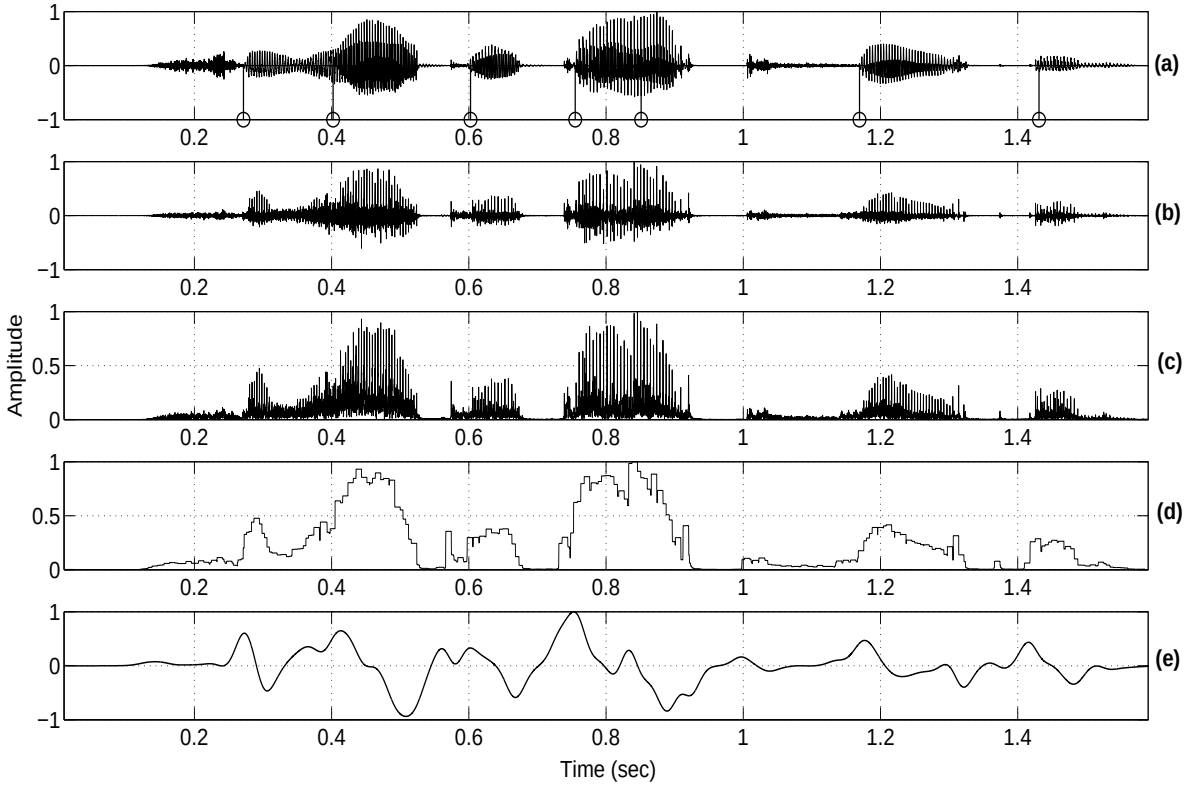

Figure 3.2: Steps involved in VLROP evidence using HE of LP residual (a) A portion of speech signal of the text *she had your dark suit in greasy wash water all year* taken from TIMIT database with reference VLROPs (circles), (b) LP residual, (c) HE of LP residual (d) smoothed excitation contour (e) VLROP evidence using HE of LP residual

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) [173]. 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 [173]:

- Difference input speech signal $s(n)$

$$x(n) = s(n) - s(n-1) \tag{3.7}$$

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

$$y(n) = -\sum_{k=1}^{4} a_k y(n-k) + x(n) \tag{3.8}$$

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

- Remove the trend i.e.,

$$\hat{y}(n) = y(n) - \bar{y}(n) \tag{3.9}$$

  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.

Fig. 3.3 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. 3.3(b). The zero crossings give the locations of epochs and are shown in Fig. 3.3(c). The slope around the zero crossings are proposed earlier as strength information [178]. The epochs with their strength are given in Fig. 3.3(d). Thus first order difference of ZFFS (Fig. 3.3(b)) given in Fig. 3.3(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. 3.3(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. 3.3(f) and 3.3(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 sec).

Finally, the *VLROP evidence using the excitation source information* is obtained by adding the two evidences and normalizing by 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 (0.05) to eliminate the spurious ones. Peak locations in the combined evidence are hypothesized VLROPs. Since the excitation features used for the VLROP evidence are robust to different degradations [177,178], the performance of the VLROP detection does not critically depend on the threshold. It was observed experimentally that a threshold value from 0.03 to 0.1 provides nearly the same performance.

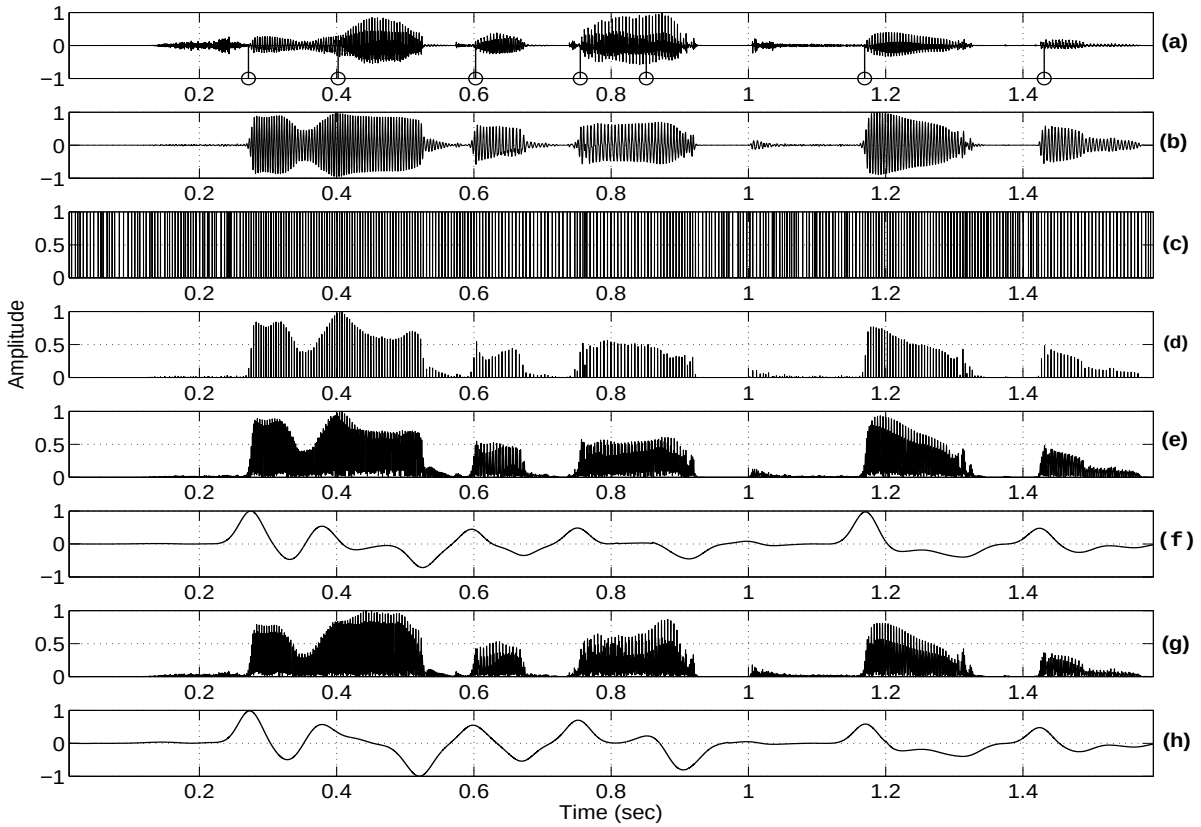

Figure 3.3: Steps involved in zero frequency filtering (a) A portion of speech signal of the text *she had your dark suit in greasy wash water all year* taken from TIMIT database with reference VLROPs (circles), (b) zero frequency filtered signal (ZFFS), (c) epoch locations (d) strength of excitation (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

3.2.3 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* [73]. 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 [180]:

- *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.

The performance of proposed VLROP detection algorithm is given in Table 3.1. For comparison, individual performances of HE of LP residual, ZFFS and a method based on VOP detection [79] are also given in the same table. The performance of ZFFS is better in terms of identification accuracy. The performance of proposed method is better in terms resolution. The 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 poorer performance compared to ZFFS, it combines well to improve the resolution of VLROP. The possible reason for high spurious VLROPs in case of the VOP detection method is due to the emphasis provided for low energy regions by peak enhancement [79].

Table 3.1: Performance of VLROP detection methods using excitation source information and based on the VOP detection method for speech signals from TIMIT database. The abbreviations VOP, HE, ZFFS and ESI refer to performance due to VOP detection method, HE of LP residual, zero frequency filtered signal and proposed excitation source method.

Method	IR (%)	MR (%)	SR (%)	IA (ms)
VOP	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 mixed with two different noises from the NOISEX-92 database [181]: *factory-1* and *white noise*. The energy level of the noise is scaled such that the overall SNR of the noise degraded speech is maintained at 3 dB. The performance of the proposed VLROP detection and based on VOP detection methods for noise degraded speech are given in the Table 3.2. By comparing the Tables 3.1 and 3.2, it can be observed that the spurious rate in case of 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 terms of resolution compared to each individual feature and the performance is almost same for both clean and degraded speech.

Table 3.2: Performance of VLROP detection methods using excitation source information and based on the VOP detection method for noise degraded speech. The abbreviations VOP, HE, ZFFS and ESI refer to performance due to 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	VOP	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	VOP	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

3.2.4 Detection of VLRs from degraded speech

The Fig. 3.4(b), shows the VLROP evidence for a segment of speech taken from NIST-2003 speaker recognition database [182]. The corresponding evidences for *white noise* degraded speech are shown in the Fig. 3.4(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. The Figs. 3.4(a) and (b) show that the hypothesized VLROPs nearly correspond to the starting of VLRs. The Figs. 3.4(c)-(f) show 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.

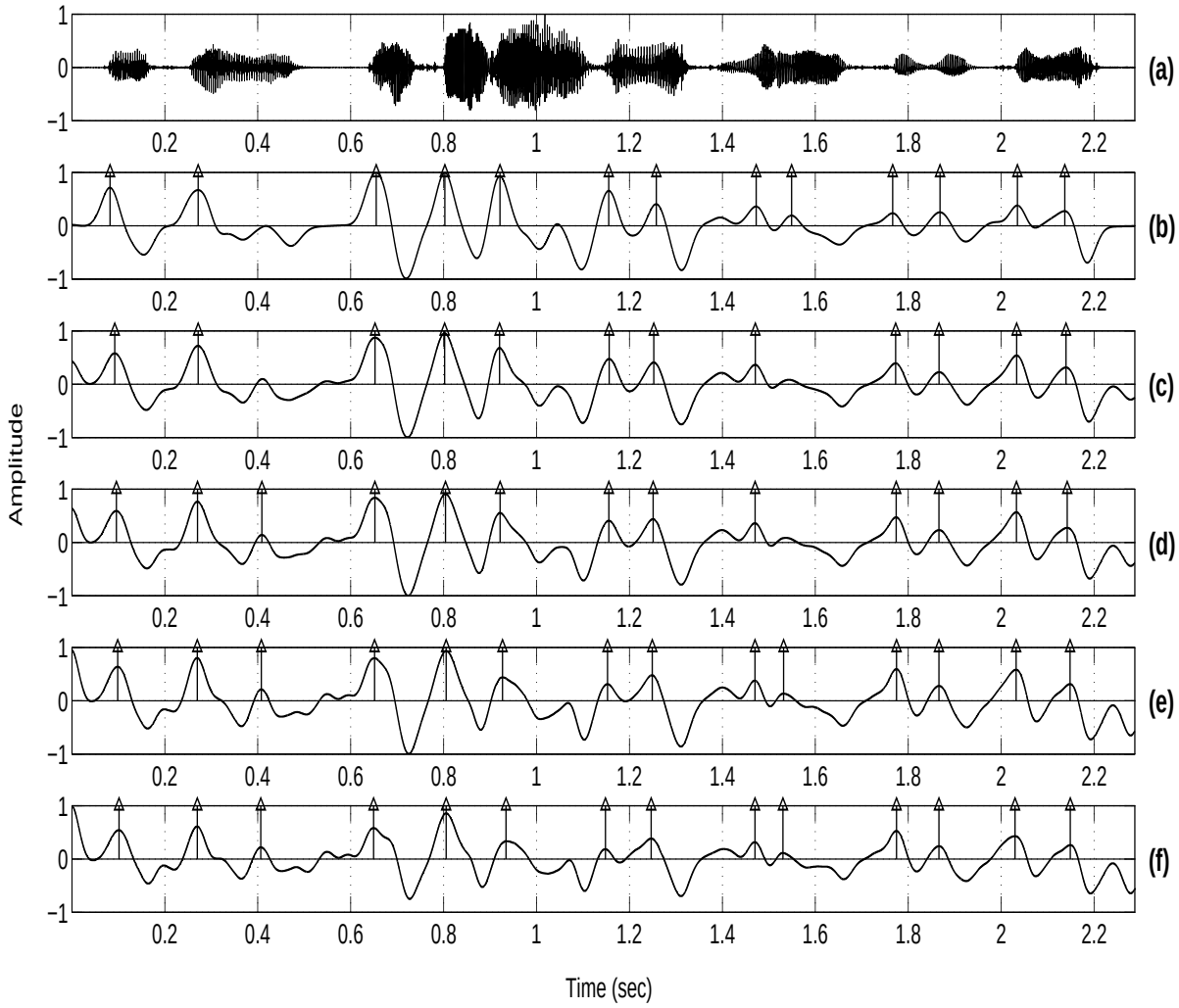

Figure 3.4: VLROP evidences for degraded speech. (a) Segment of speech taken from NIST-2003 speaker recognition database, (b) VLROP evidence for clean speech, (c)-(f) VLROP evidences for *white noise* degraded speech with over all SNR level 9 dB, 6 dB, 3 dB and 0 dB, respectively.

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. 3.5. 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 ($0.06 \times \textit{average energy}$), which are represented in dotted lines. In case of clean

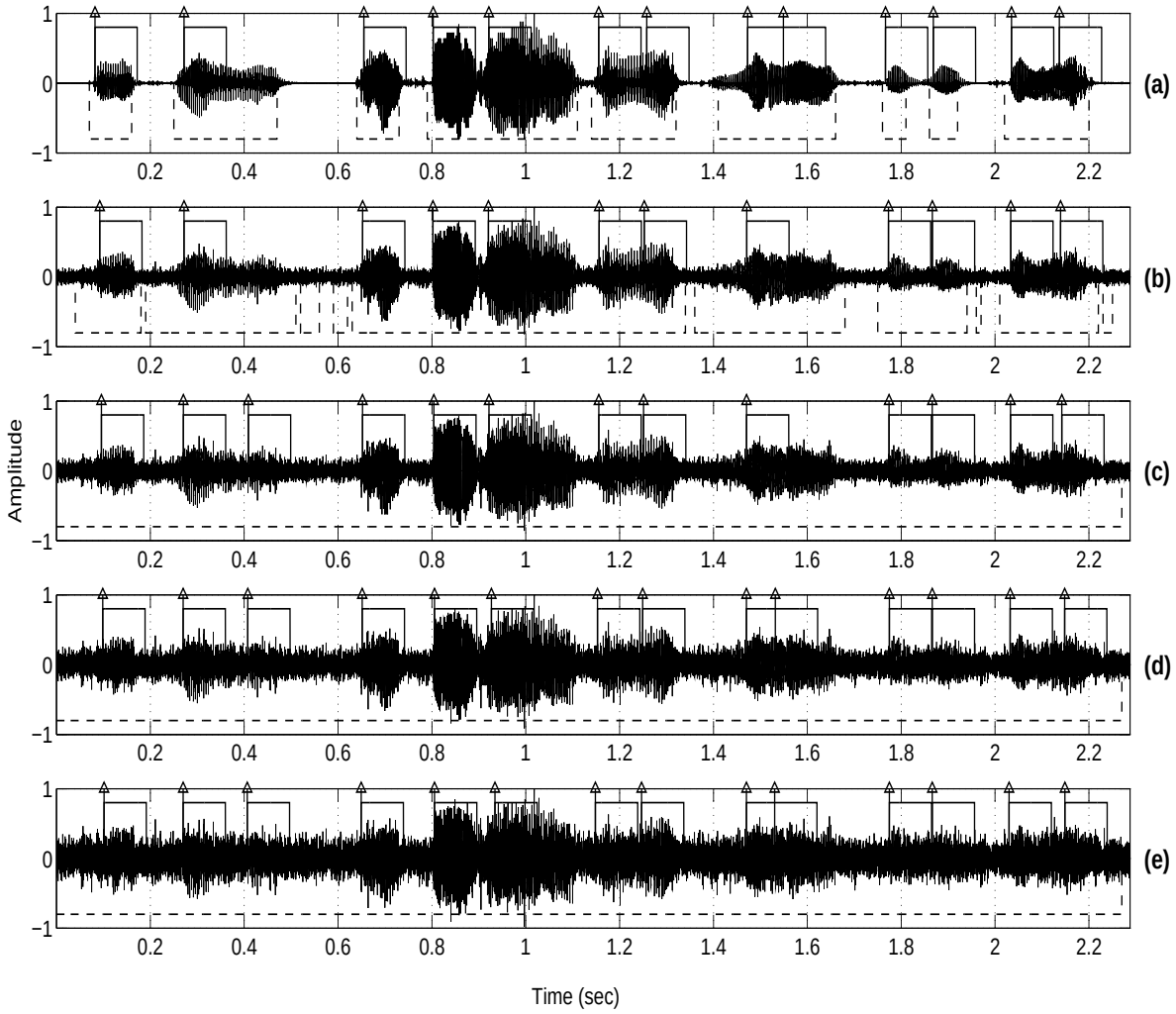

Figure 3.5: VLRs (solid lines) and speech regions (dotted lines) detection in degraded condition. (a) Segment of speech taken from NIST-2003 speaker recognition database, (b)-(e) noise degraded speech with over all SNR level 9 dB, 6 dB, 3 dB and 0 dB, respectively.

speech shown in Fig. 3.5(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. Figs. 3.5(b)-(e) show 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.

The effectiveness of proposed algorithm can be better investigated using clean speech (head-mounted microphone recordings) and degraded speech (digital voice recorder recordings) of IITG Multi-Variability (MV) speaker recognition database [133]. The speech file shown in Fig. 3.6(a) is a segment of speech recorded with head-mounted microphone (clean speech) and the same segment speech recorded in parallel with digital voice recorder (degraded speech) is shown in Fig. 3.6(c). The VLROP evidences corresponding to clean and degraded speech are shown in Fig. 3.6(b) and (d), respectively. The Fig. 3.6 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

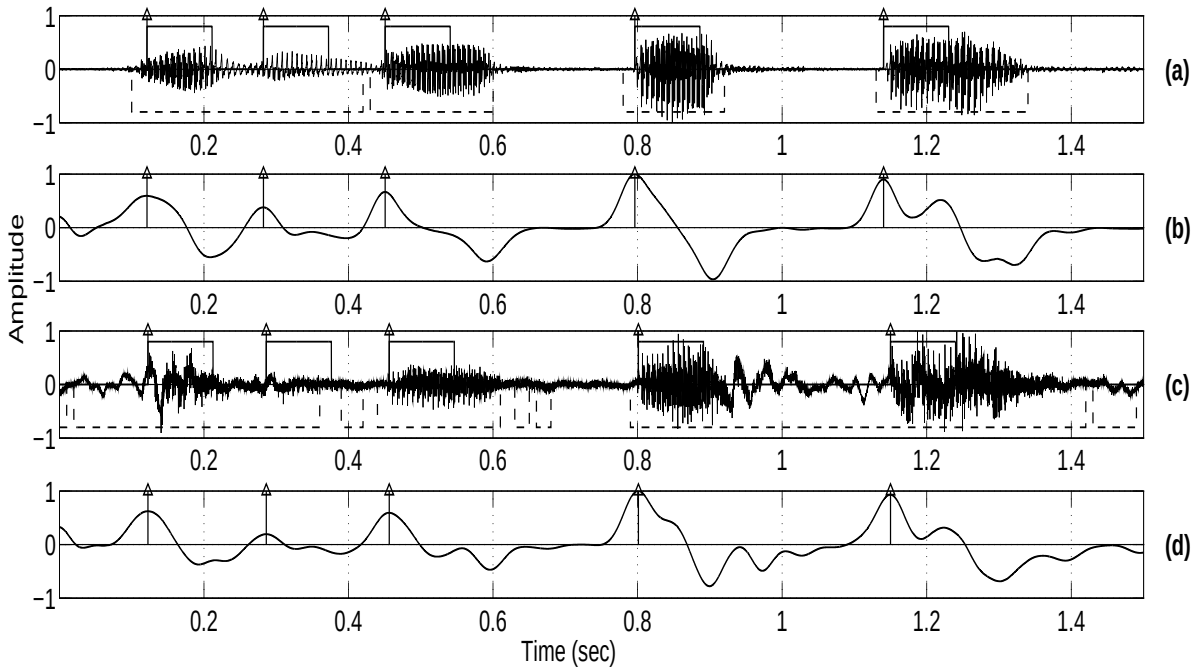

Figure 3.6: VLRs (solid lines) and speech regions (dotted lines) detection for clean and degraded speech of IITG MV speaker recognition database. (a) Clean speech (speech recorded with head-mounted microphone), (b) VLROP evidence for clean speech, (c) degraded speech (speech recorded in parallel with digital voice recorder), (d) VLROP evidence for degraded speech.

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 condition.

3.3 Speaker Verification using VLRs

3.3.1 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 two different noises from the NOISEX-92 database [181]: *factory-1* and *white noise* are taken to generate the noise mixed 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 files are 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 mounted close to the speaker, inbuilt tablet PC microphone, two mobile phones and one digital voice recorder. 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 with headphone microphone and inbuilt tablet PC microphone are at 16 kHz and stored with 16 bits/sample resolution. Speech recorded with 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 with 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.

3.3.2 Detection of VLRs

As described in Section 3.2, VLROPs are determined using the excitation source information derived from the speech signal. Using each hypothesized VLROP as the anchor point, 100-ms segments following VLROPs are labeled 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.

3.3.3 Feature extraction

During the training and testing process, the speech signal is processed in frames of 20 ms duration at a 10 ms frame shift. For each 20 ms Hamming windowed frame, MFCCs are calculated using 22 logarithmically spaced filters. The first 13 coefficients, excluding the zeroth coefficient, are used as a feature vector [18]. Delta (Δ) and delta-delta ($\Delta\Delta$) features are computed using the two preceding and two succeeding feature vectors from the current feature vector. Thus the feature vector will be of 39 dimension with 13 MFCC, 13 ΔMFCC and 13 $\Delta\Delta$MFCC.

3.3.4 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 [55]. 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 with 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.

3.3.5 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 conditions. 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 [35]. 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 $U = \{\mu_c, \Sigma_c, \eta_c\}$, $c = 1, ...C$, 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 [35]. 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 [60] [183]. Hence in the present work the T-norm is employed as the score normalization technique [60].

3.3.6 Baseline SV system

In order to compare the SV performance obtained using VLRs, we have developed a 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.

3.3.7 Performance Comparison

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

$$EER_R = \frac{(EER_B - EER_V)}{EER_B} \times 100\ \% \tag{3.10}$$

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.

3.4 Experimental Studies

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

(i) *Original NIST-2003*: NIST-2003 speaker recognition database is used for the performance evaluation.

(ii) *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.

(iii) *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.

(iv) *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, thirty hours of UBM speech was selected from randomly selected 250 male and 250 female speech of switchboard cellular part 2 Audio database [184]. 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 [35]. 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 VLROPs 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, 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 with digital voice recorder, 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 with head-mounted microphone is more clean compared to other sensors. Accordingly, the speech recorded with digital voice recorder is considered as degraded speech and speech recorded with head-mounted microphone is considered as clean speech.

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

(i) *Clean and sensor matched*: Speech recorded with head-mounted microphone is used for training and testing.

(ii) *Clean train and degraded test*: Speech recorded with head-mounted microphone is used for training and speech recorded with digital voice recorder is used for testing.

(iii) *Degraded train and clean test*: Speech recorded with digital voice recorder is used for training and speech recorded with head-mounted microphone is used for testing.

(iv) *Degraded train and test*: Speech recorded with digital voice recorder 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: head-mounted microphone and digital voice recorder. 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 non-availability 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.

3.5 Results and Discussions

This section tabulates the various experimental results of the SV studies and also discusses the possible reasons for the trends in each case.

3.5.1 NIST-2003 speaker recognition database

3.5.1.1 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_{Avg})) to a comparable higher threshold ($0.3 \times \ E_{Avg}$). The detection estimation trade-off (DET) plots given in Fig. 3.7(a) shows the performance of the SV system in terms of equal error rate (EER) for threshold values $0.006 \times \ E_{Avg}$, $0.06 \times \ E_{Avg}$, $0.1 \times \ E_{Avg}$, $0.15 \times \ E_{Avg}$, $0.2 \times \ E_{Avg}$ and $0.3 \times \ E_{Avg}$ is 7.12%, 6.91%, 7.04%, 7.6%, 8.67% and 10.07%, respectively. The effect of energy threshold on the SV performance is summarized in Fig. 3.8. This shows 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 ($0.06 \times \ E_{Avg}$) is fixed as the energy threshold for the baseline system for further experiments.

The DET plots in Fig. 3.7(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 [64]. The 95% confidence intervals at EER operating points do not overlap. The relative improvement in EER for the SV system using VLRs

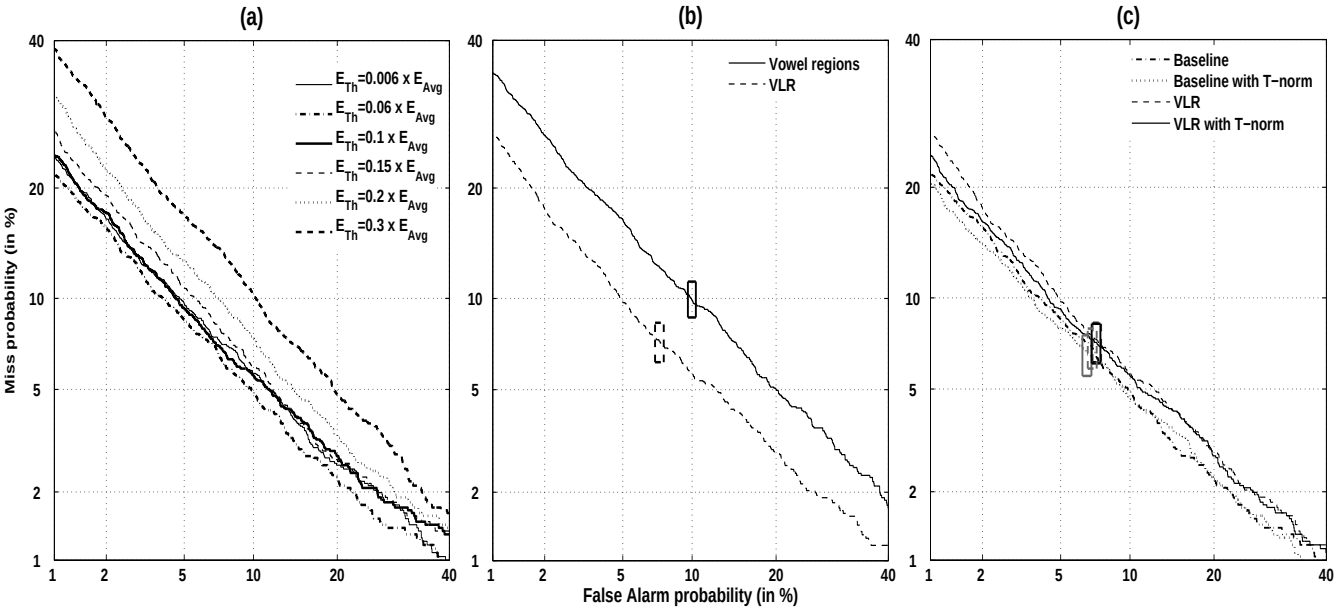

Figure 3.7: DET curves showing performance for various experimental setup of NIST-20003 speaker recognition database. (a) Effect of energy threshold on speaker verification performance, (b) performance using vowel regions and VLRs, (c) performance of baseline system and SV system using VLR. The boxes indicate the 95% confidence intervals at EER operating points.

Table 3.3: 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 × 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

is 35.85% compared to the SV system using only vowel regions. Thus VLROPs are preferable over VOPs. The DET plots in Fig. 3.7(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 3.3. 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 [185, 186].

It is also observed that the T-norm provides more improvement to the baseline compared to the

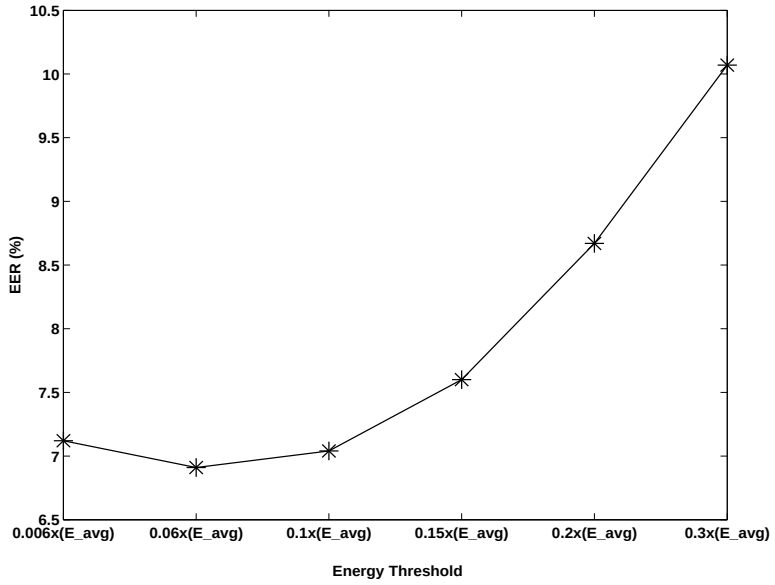

Figure 3.8: Performance (in EER) of the baseline SV system for different energy thresholds (E_{Avg}) on NIST-20003 speaker recognition database

SV system using VLRs. The additional score normalization like T-norm is generally 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 3.4. 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 the Fig. 3.7(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 more speaker information and improved performance in baseline system may be due to the significantly more number of features used for training and testing.

Table 3.4: 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

3.5.1.2 Noise degraded NIST-2003 test speech

For most of the practical implementation of a 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 mixed with two different noises from the NOISEX-92 database: *factory-1* and *white noise*. For each noise, the noise level is scaled such that the overall SNR of noise degraded speech is maintained as 9 dB, 6 dB, 3dB and 0 dB, respectively. The DET plots in Figs. 3.9(a)-(d) show performance of the SV system using VLRs and the baseline system for *factory-1 noise* degraded test speech for 9 dB, 6 dB, 3 dB and 0 dB SNR, respectively. The DET plots in the Fig. 3.9 (e)-(h) shows performance of the SV using VLRs and the baseline system for *white noise* degraded 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 3.5 and are summarized in Fig. 3.10.

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

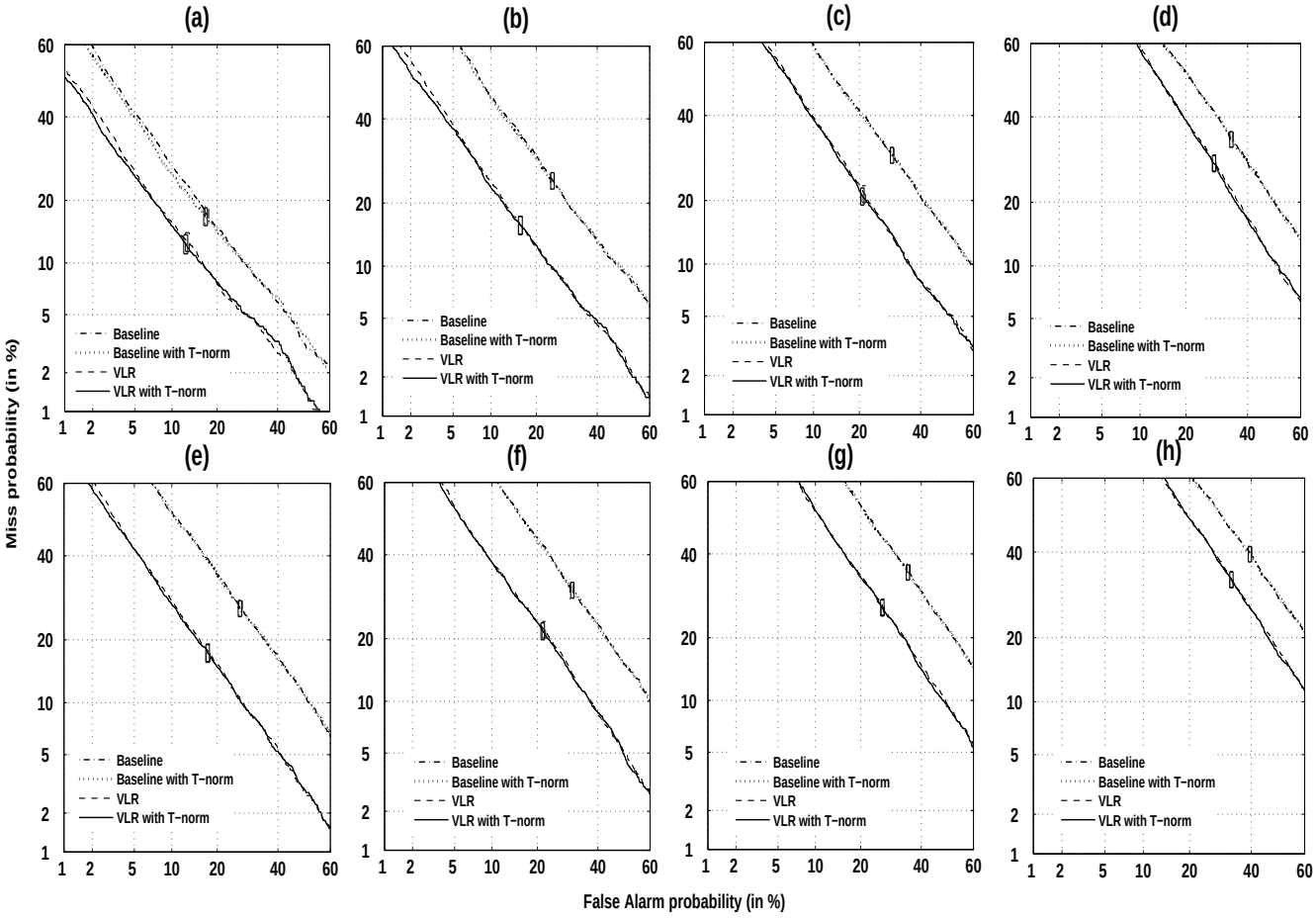

Figure 3.9: DET curves showing performance for noise degraded NIST-2003 test speech. (a)-(d) *factory-1 noise* degraded test speech for SNR level 9 dB, 6 dB, 3 dB and 0 dB, (e)-(h) *white noise* degraded test speech for SNR level 9 dB, 6 dB, 3 dB and 0 dB. The boxes indicate the 95% confidence intervals at EER operating points.

Table 3.5: 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

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

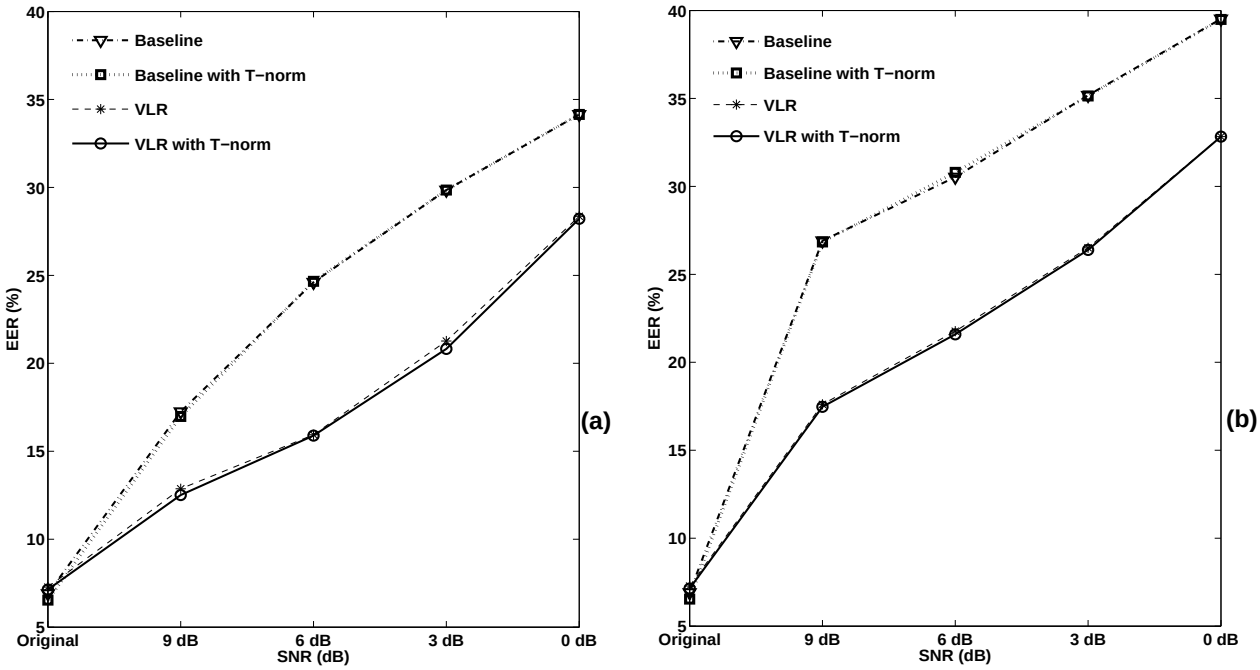

Figure 3.10: Summary of SV performance for noise degraded NIST-2003 test speech without (w/o) and with T-norm. (a) *factory-1 noise* degraded test speech for SNR level 9 dB, 6 dB, 3 dB and 0 dB, (b) *white noise* degraded test speech for SNR level 9 dB, 6 dB, 3 dB and 0 dB.

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.

3.5.1.3 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 are mixed with two different noises from the NOISEX-92 database: *factory-1* and *white noise*. 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. 3.11(a)-(d) shows the performance of SV system using VLRs and

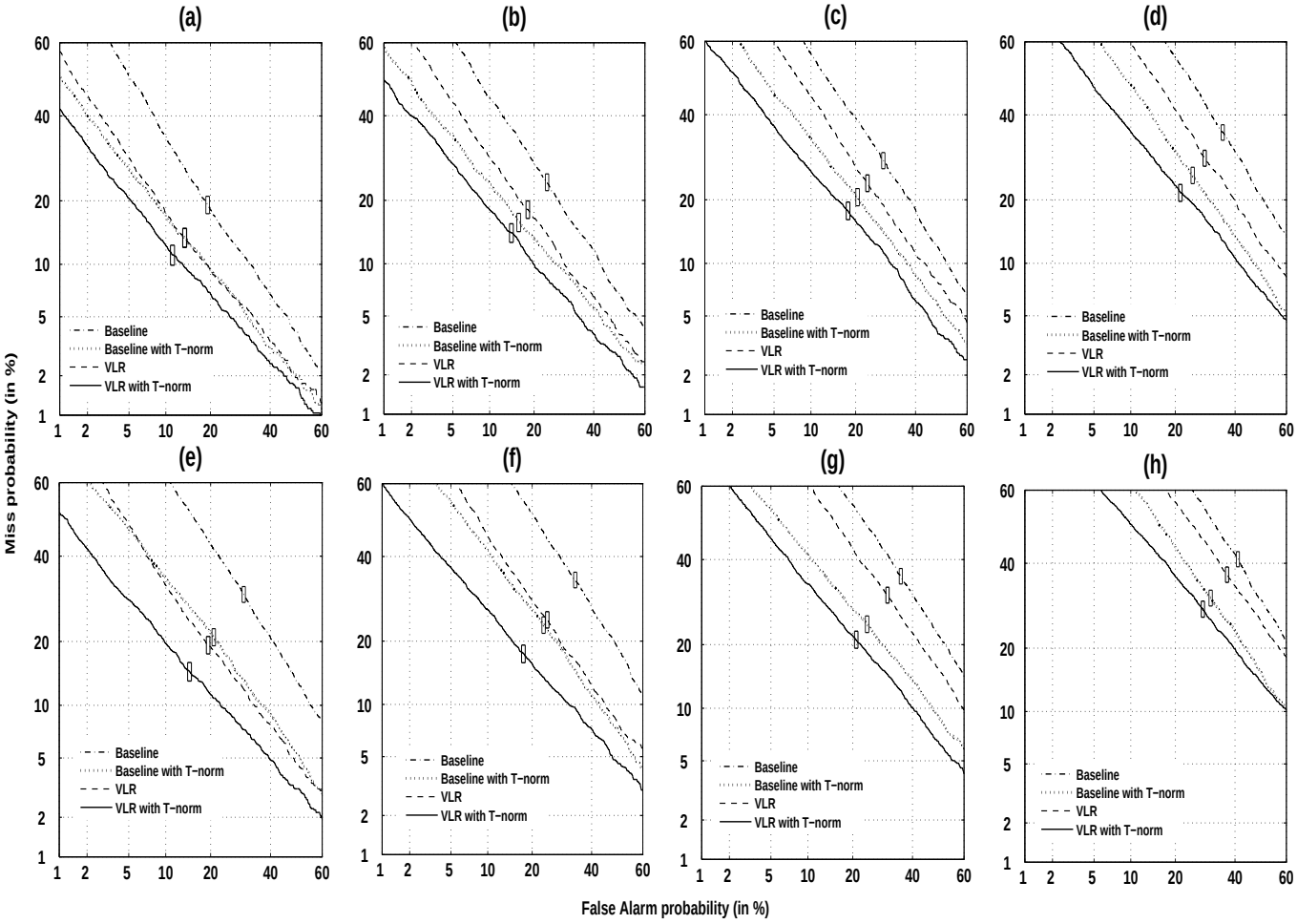

Figure 3.11: DET curves showing performance for noise degraded NIST-2003 train speech. (a)-(d) *factory-1 noise* degraded train speech for SNR level 9 dB, 6 dB, 3dB and 0 dB, (e)-(h) *white noise* degraded train speech for SNR level 9 dB, 6 dB, 3 dB and 0 dB. The boxes indicate the 95% confidence intervals at EER operating points.

Table 3.6: 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

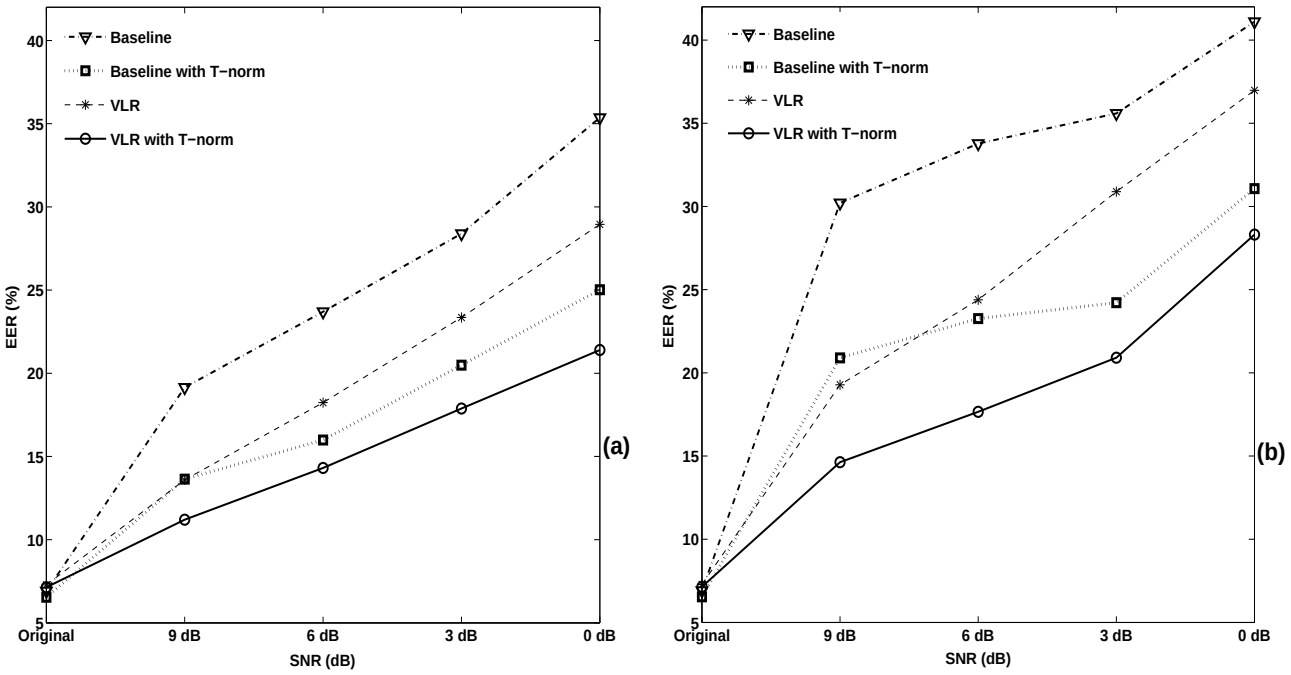

Figure 3.12: Summary of SV performance for noise degraded NIST-2003 train speech without (w/o) and with T-norm. (a) *factory-1 noise* degraded train speech for SNR level 9 dB, 6 dB, 3 dB and 0 dB, (b) *white noise* degraded train speech for SNR level 9 dB, 6 dB, 3 dB and 0 dB.

baseline for *factory-1 noise* degraded train speech for 9 dB, 6 dB, 3 dB and 0 dB SNR, respectively. The DET plots in the Fig. 3.11(e)-(h) shows performance of the SV using VLRs and the baseline system for *white noise* degraded 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 3.6 and are summarized in Fig. 3.12. 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.

3.5.1.4 Noise degraded NIST-2003 train and test speech

This set of experiment 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 training and testing speech of NIST-2003 are mixed with two different noises from the NOISEX-92 database: *factory-1* and *white noise* 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 mixed using same noise and SNR level. The DET plots given in Fig. 3.13(a)-(d) shows the performance of SV system using VLRs and baseline for *factory-1 noise* degraded train and test speech under 9 dB, 6 dB, 3 dB and 0 dB SNR, respectively. The DET plots in the Fig. 3.13(e)-(h) shows performance of the SV using VLRs and the baseline system for *white noise* degraded 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 3.7 and are summarized in Fig. 3.14. 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.

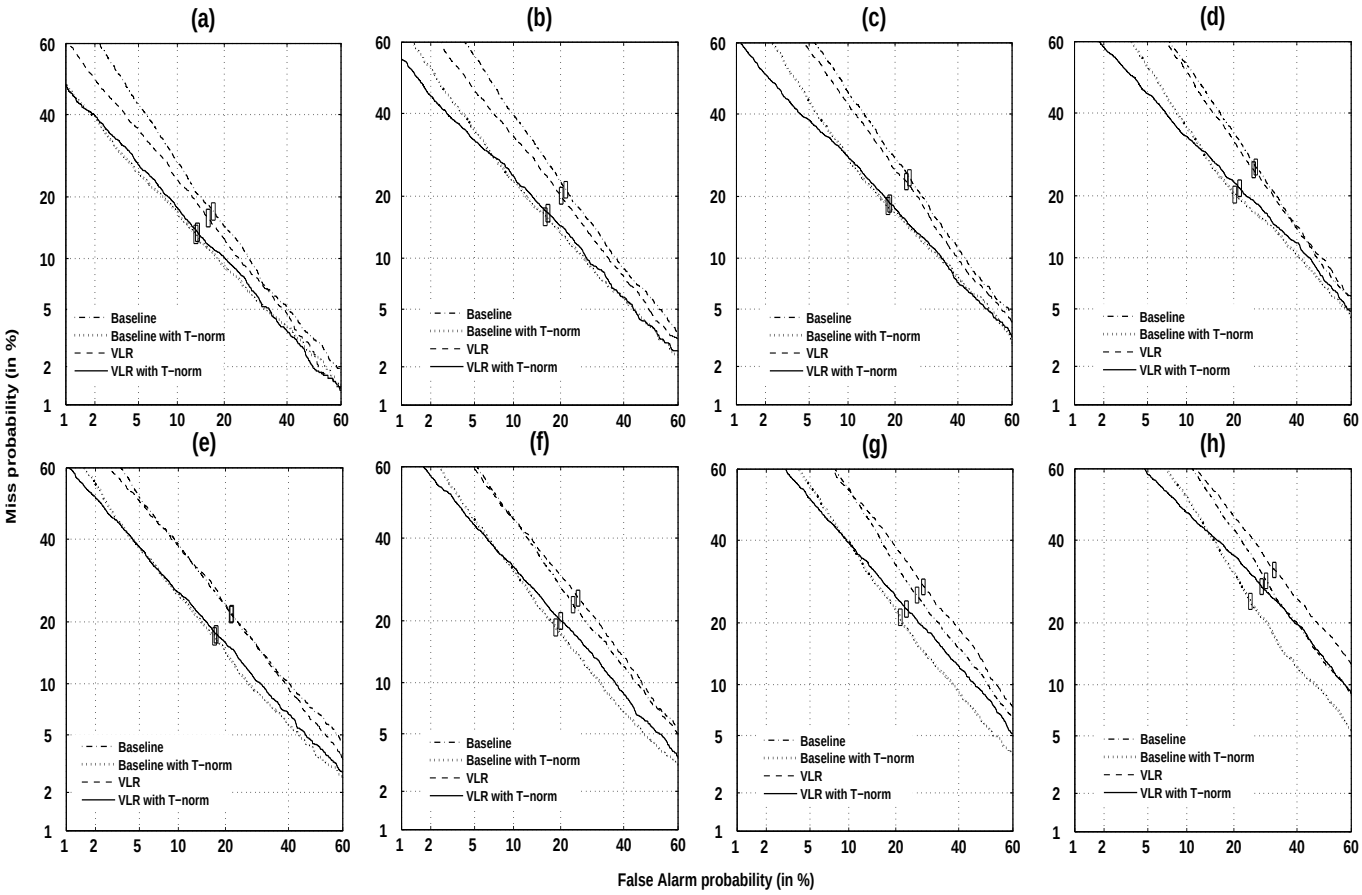

Figure 3.13: DET curves showing performance for noise degraded NIST-2003 train and test speech. (a)-(d) *factory-1 noise* degraded train and test speech for SNR level 9 dB, 6 dB, 3 dB and 0 dB, (e)-(h) *white noise* degraded train and test speech for SNR level 9 dB, 6 dB, 3 dB and 0 dB. The boxes indicate the 95% confidence intervals at EER operating points.

Table 3.7: 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

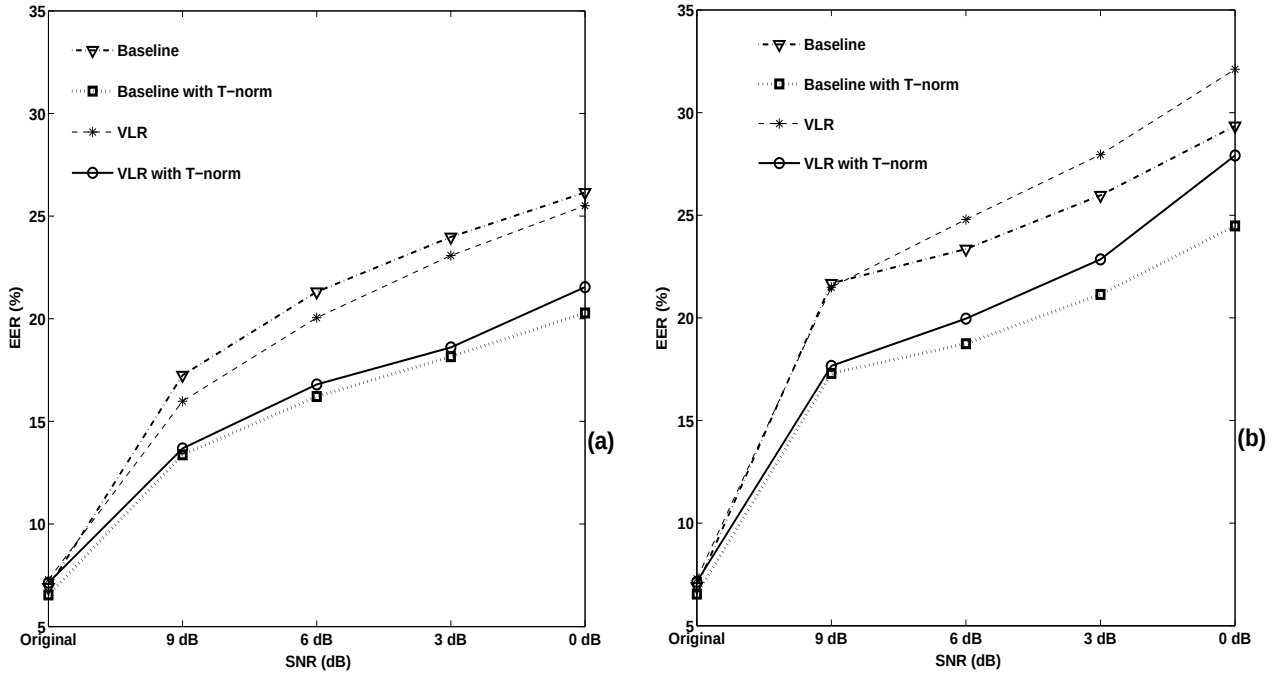

Figure 3.14: Summary of SV performance for noise degraded NIST-2003 train and test speech without (w/o) and with T-norm. (a) *factory-1 noise* degraded train and test speech for SNR level 9 dB, 6 dB, 3 dB and 0 dB, (b) *white noise* degraded train and test speech for SNR level 9 dB, 6 dB, 3 dB and 0 dB.

3.5.2 IITG MV speaker recognition database

3.5.2.1 Clean and sensor matched

The performance of baseline system and SV system using VLRs on IITG-MV speaker recognition database are given in the Table 3.8 and are summarized in Fig. 3.16. 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. 3.15(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.

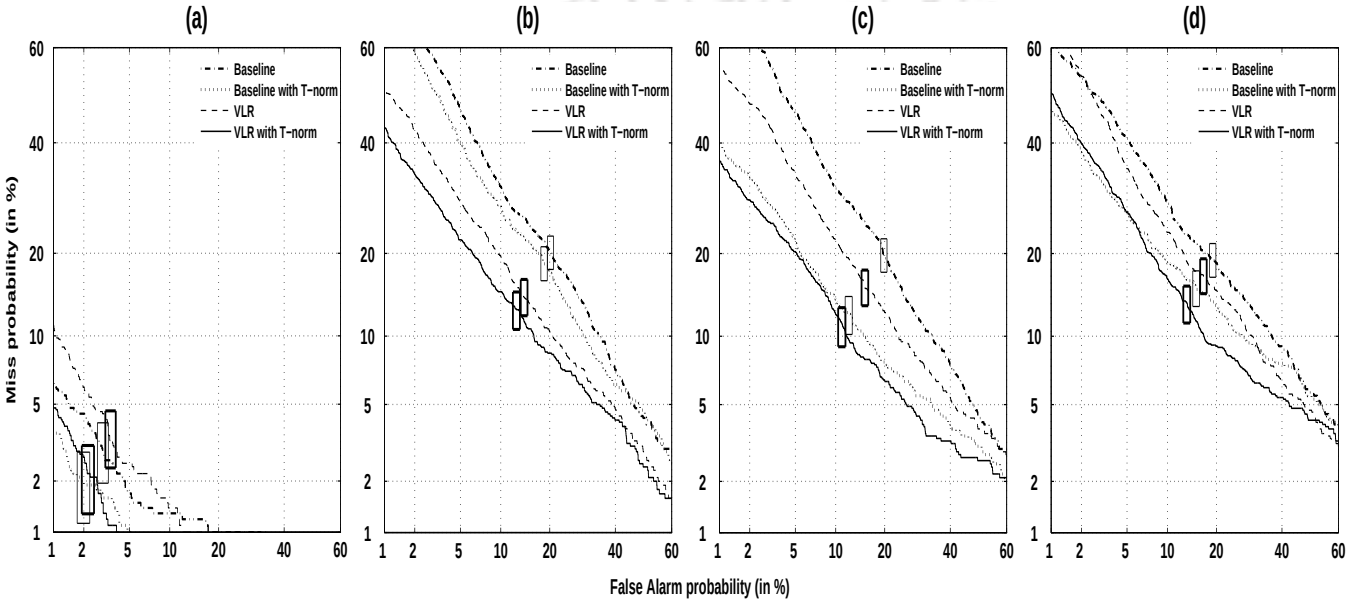

Figure 3.15: DET curves showing performance for various experimental setup of IITG MV speaker recognition database. (a) Clean and sensor matched, (b) clean train degraded test, (c) degraded train and clean test, (d) degraded train and degraded test. The boxes indicate the 95% confidence intervals at EER operating points.

Table 3.8: 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

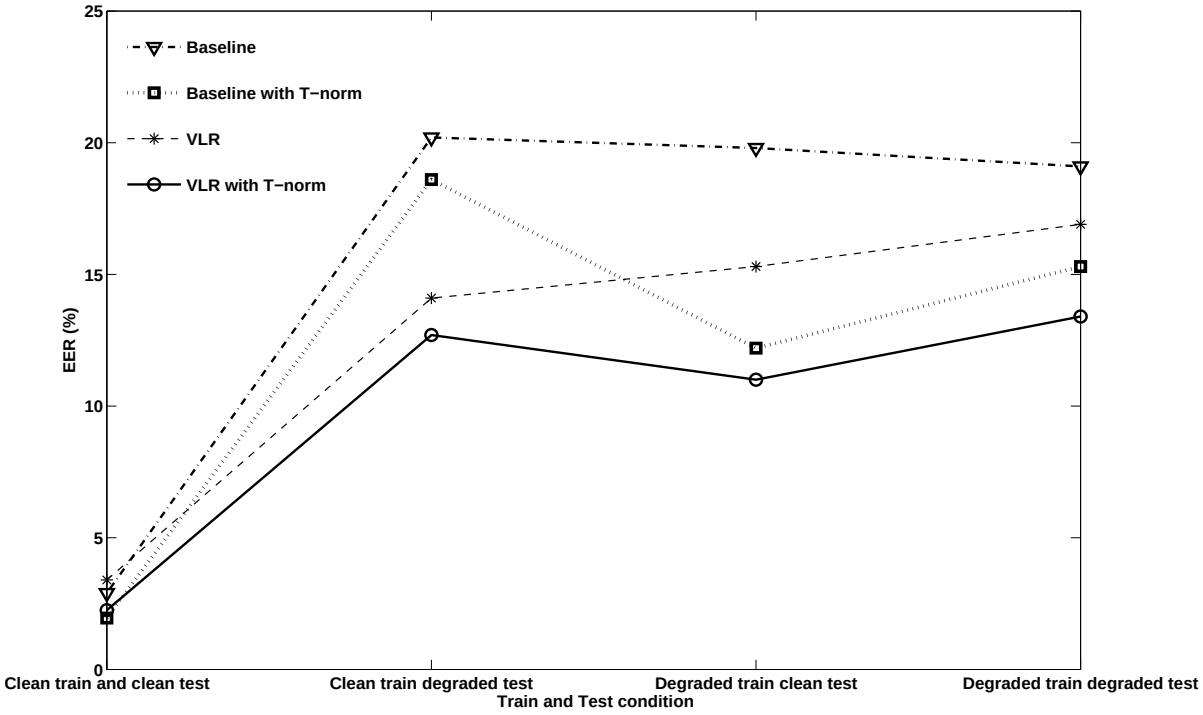

Figure 3.16: Summary of SV performance for IITG MV database without (w/o) and with T-norm.

3.5.2.2 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 with digital voice recorder 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 are collected over different sensors. The DET plots in Fig. 3.15(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.

3.5.2.3 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. 3.15(c) shows performance of the 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.

3.5.2.4 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. 3.15(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.

3.6 Summary

In this work we proposed a new VLROP detection method for clean and degraded speech by utilizing the advantages of HE of the LP residual and zero frequency filtered signal. 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 segments following VLROPs are labeled as VLRs. SV systems are developed using the speech regions detected by energy based VAD (baseline) and the VLRs. For both the systems MFCC is used as the speaker feature and GMM-UBM is used as the modeling technique. Degradation effect is compensated in the cepstral domain using CMS followed by CVN and in the score level using T-norm. The performance of the SV systems are evaluated on NIST-2003 speaker recognition database. In the second level, two different noises are taken from NOISEX-92 database to create noise mixed NIST-2003 speech. Performance of the SV systems are evaluated for different degraded conditions on noise mixed NIST-2003 speaker recognition database. Finally, performance of the SV system is evaluated on the IITG MV speaker recognition database for clean and real environmental degraded speech.

This work shows that for clean speech, the proposed SV system gives poorer performance compared

to the baseline. Alternatively, under degraded conditions, the proposed system provides significantly improved performance. The poor performance of the proposed SV system for clean speech may be due to two possible reasons, (1) significantly less number of feature vectors (fixed 100-ms segments from VLROPs) or (2) the non-VLRs contain different speaker information. To address these issues, next chapter presents a method for detection of complete VLRs and non-VLRs. The effect of VLRs and non-VLRs detection on SV performance is described in detail. A SV system is developed by conditioning VLRs and non-VLRs for better SV performance under clean and degraded speech.

4

Speaker verification by explicit segmentation of vowel-like and non-vowel-like regions

Contents

Objective

The initial exploration of a SV system using VLRs described in the previous chapter shows that VLRs are more speaker specific and less affected by degradations. In this study, fixed 100-ms segment following a VLROP is labeled as a VLR and other speech regions due to the non-VLRs are neglected. As a result, for clean and matched speech the SV performance is poorer compared to the SV system using all speech regions. This chapter proposes a SV system using VLRs and non-VLRs to improve the SV performance both for the clean and degraded conditions. To achieve this, methods are proposed for detecting complete VLRs and non-VLRs. The VLRs and non-VLRs are used independently during training and testing of a SV system to reduce gross level mismatch due to sound units and achieve better compensation of degradation effects by applying different normalization to these two different energy regions. Finally, the scores are combined with higher weight on VLRs, which are more speaker specific.

4.1 Introduction

Conventional speaker verification, where the VLRs and non-VLRs are processed together may not be the best choice to maintain statistical matching between the speaker model and test features [95,97]. Previous studies using different constraints for selecting features also suggest that the performance of GMM-based SV systems can be improved by applying conditioning during training and testing [66, 70, 81]. The performance of such constrained SV systems depends on the type of constraint and ability for the detection of constraint regions during training and testing.

The initial exploration of a SV system using VLRs described in the previous chapter shows that VLRs are more speaker specific and less affected by degradations. By using VLRs during training and testing of a SV system, performance can be improved under degraded conditions. A limitation of this study is that it selected fixed 100-ms segments from VLRs. As a result, when VLRs were longer than 100 ms, some portions were excluded. Also, the other speech frames due to non-VLRs are neglected. Even though non-VLRs are affected relatively more by degradation due to their low energy and non-impulsive excitation, they account for a large number of frames and also contain good speaker information. The work presented in Chapter 3 can therefore be extended in two directions: (1) developing a method that performs automatic detection of complete VLRs and also a method that detects non-VLRs, and (2) with the help of these two methods, detecting and using all the

relevant regions from the speech signal for speaker verification in an independent and parallel fashion. With this segmentation, efforts may be focused mainly for non-VLRs against degradation. Under degraded conditions, better compensation of degradation effects may be achieved by applying different normalizations to these two different segment types [74]. Also, higher weight can be applied to the scores obtained from the VLRs, which are more speaker specific than non-VLRs [187].

An algorithm is proposed for robust segmentation of speech into VLRs and non-VLRs. Both a GMM-UBM and an *i*-vector based SV system are developed using MFCCs. The SV performance is evaluated using VLRs and non-VLRs independently for clean as well as different degraded test conditions, assuming that the knowledge about the degradation is not known. The verification scores are then combined with more weight to VLRs. For the NIST-2003 speaker recognition database [182] and IITG-MV speaker recognition database [188], the proposed approach outperforms the conventional approach of using the same regions without conditioning.

The novel contributions of the work presented in this chapter are the following:

- Defining the vowel-like region end point (VLREP) event.
- An iterative algorithm for the detection of complete VLRs using the vowel-like region onset point (VLROP) (Chapter 3) and proposed VLREP.
- A method for detecting non-VLRs by emphasizing excitation information of non-VLRs in the linear prediction (LP) residual.
- Speaker verification studies using VLRs and non-VLRs on GMM-UBM and *i*-vector based speaker verification systems.

The rest of the Chapter is organized as follows: methods for the detection of VLROPs and VLREPs and selection of VLRs are described in Section 4.2. Selection of non-VLRs and segmentation of the speech signal into VLRs and non-VLRs are described in Section 4.3. Proposed SV systems using VLRs and non-VLRs are described in Section 4.4. The experimental studies are presented in Section 4.5. The experimental results are discussed in Section 4.6. The summary of the work are mentioned in Section 4.7.

4.2 Detection of vowel-like regions from speech

The objective of work presented in the previous Chapter was to demonstrate robustness of VLRs for speaker verification. However, we may benefit further by selecting whole VLRs and using them

for speaker verification. The detection of an entire VLR involves identifying its begin and end points, namely, the VLROP and VLREP. For a particular vowel-like cluster, one VLROP and one VLREP are sufficient to detect the VLRs. Knowledge of VLREPs enhances the true detection and reduces the spurious detection of VLROPs. If the detection of VLROPs and VLREPs is robust, then by anchoring these two points, the exact VLRs can be selected. As described in Chapter 3, the excitation source information used for the detection of VLROPs is robust to different degradations. Therefore, present work uses the same for the detection of VLREPs and develops an iterative algorithm for the identification of VLRs.

4.2.1 Detection of the VLR end point (VLREP)

Fig. 4.1(b) shows a segment of speech for the words **the sea** taken from a conversational sentence. Selecting VLRs by assuming a fixed length VLR after the VLROP (dotted line in Fig. 4.1(b)) (Chapter 3) or by finding the valley [32] is not accurate in most cases (Fig. 4.1(c)). The signal characteristics at the end of a VLR are significantly different than at the beginning. At the onset, there is a sudden increase in signal strength, while the signal strength decreases slowly at the end. Due to this, detecting VLREPs is more challenging than detecting VLROPs. Therefore, to detect VLREPs processing is done from right to left using a FOGD which is double in length and standard deviation compared to the VLROP detector. The increase in FOGD size accumulates evidence over a longer window, providing good evidence for the detection of VLREP events, even in the case of weak transitions.

Both the smoothed HE of the LP residual (Eqn. (3.5)) and the second order difference of the ZFFS (Eqn. (3.9)) are convolved with the FOGD from right to left. The *VLREP evidence* is obtained by adding the two evidences and normalizing with respect to 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 a threshold (0.04) to eliminate the spurious ones. The peak locations in the combined evidence are hypothesized VLREPs. The excitation features used for the VLREP evidence are robust to different degradations [177,178] and evidences are obtained by convolving with a FOGD. The VLREP evidence does not directly depend on the signal energy. The VLREP evidence is nearly same for clean and noise added speech (Figs.4.4 (d) and (i)). Therefore performance of the VLREP detection does not critically depend on the threshold. It was observed experimentally that a threshold value from 0.03 to 0.08 provides nearly the same performance.

The speech segment in Fig. 4.1(b) is processed through the VLROP detection algorithm (as de-

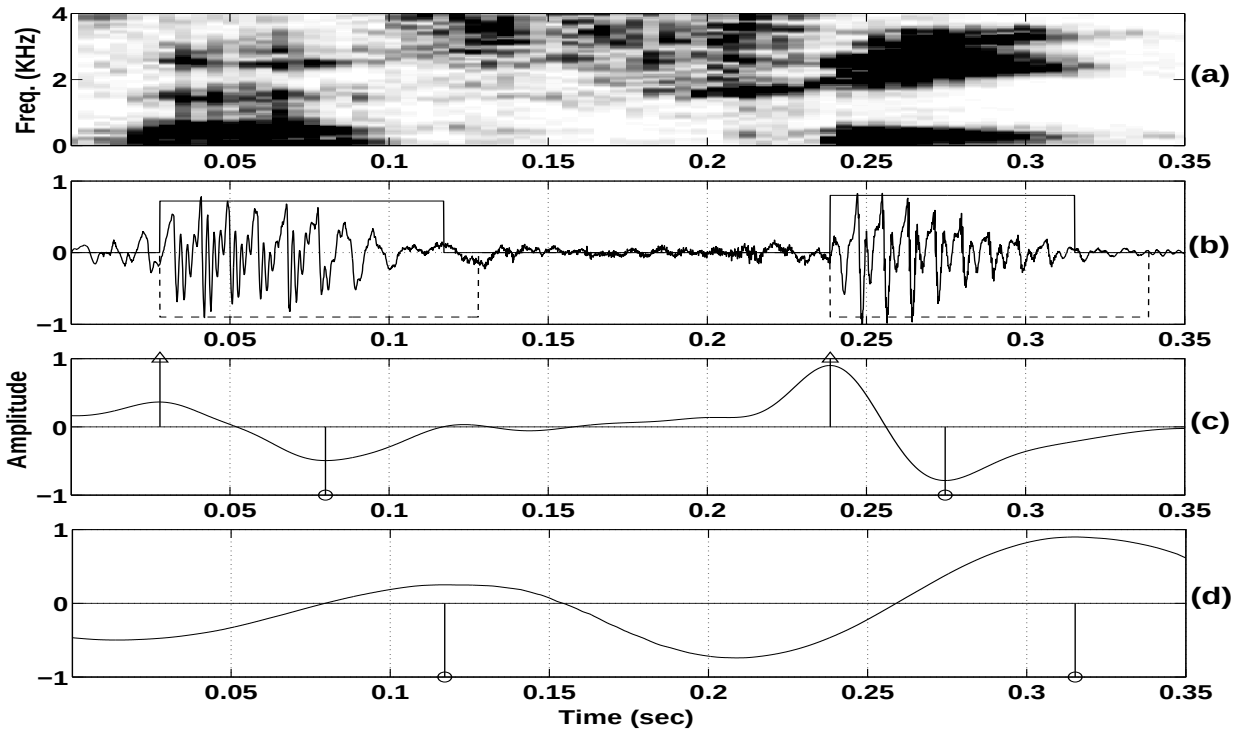

Figure 4.1: Importance of VLR end point (VLREP) for detection of VLRs. (a) Wideband spectrogram of the speech signal given in (b), (b) a segment of speech signal for the words "the sea" with detected VLRs. Solid lines using proposed method and dotted lines by considering 100 ms right to the VLROP as VLR, (c) VLROP evidence using excitation source information with hypothesized VLROPs (arrows) and VLREPs by finding valley to the hypothesized VLROPs (circles) (d) VLREP evidence using excitation source information with hypothesized VLREPs (circles).

scribed in Chapter 3) to find the *VLROP evidence* and is given in Fig. 4.1(c). Arrows in the VLROP evidence correspond to the hypothesized VLROPs. By comparing the hypothesized VLROPs and the wideband spectrogram given in Fig. 4.1(a), for each VLR, detected VLROP is accurate. By comparing Figs. 4.1(a), (b) and (c), the valley point in the VLROP evidence corresponds to the maximum negative transition point within the VLRs. Hence, if the valley is used as the VLREP, a significant amount of VLR will be lost. The VLREP evidence using the proposed VLREP detection algorithm is given in Fig. 4.1(d) and the peaks in the VLREP evidence nearly correspond to the end point of each VLR. The selection of VLRs by independent detection of VLROPs and VLREPs is shown in Fig. 4.1(b) (solid lines).

4.2.2 Detection of VLRs using VLROPs and VLREPs

Due to independent detection, the number of hypothesized VLROPs and VLREPs will differ, with misses and spurious detections occurring for both events. Therefore, neither hypothesized VLROPs

nor VLREPs can be considered as references for one-to-one matching. Detection of VLROPs and VLREPs depends on the preceding and succeeding sound units. For a vowel-semivowel cluster, the rate of change of signal strength is maximum at the beginning and end of the cluster. Hence, VLROP and VLREP evidences are maximal at the starting and ending points, respectively. This knowledge can be exploited to avoid missing VLRs within the cluster. Considering all these issues, an algorithm is proposed to, (1) force the detection of missing cases, if other evidence is sufficiently strong, and (2) reduce the spurious detection of one event using the knowledge of the other event. The algorithm works in two stages as explained in *Algorithm* 1.

Fig. 4.2 illustrates the selection of VLRs using the proposed algorithm. Fig. 4.2(c) and (d) show the VLROP and the VLREP evidences using excitation source information for the speech signal given in Fig. 4.2(b). The arrows and the circles in the evidences correspond to the hypothesized VLROPs and VLREPs, respectively. By comparing the evidences with the wideband spectrogram given in Fig. 4.2(a), both the evidences accurately discriminate the VLRs from the nasal sound units (speech region around 0.2 s and at the end of VLR at 0.8 s). From Fig. 4.2(c), the VLROP evidence is always maximal at the start of the cluster. Similarly, for VLREPs the evidence is always maximal at the end of the cluster. If the valley point is considered as a VLREP, in most cases some portion of the VLRs will be missed. But by using VLREPs, the VLRs can be selected with maximum accuracy for VLR clusters. Fig. 4.2(c) shows that the proposed algorithm forces removal of the weak evidences around 0.2 s and 1.2 s. But this will not affect the detection of VLRs. The algorithm adds the VLREP missed in the hypothesized VLREP (VLREP around 0.4 s). By comparing the Figs. 4.2(a) and (b) in terms of the detected VLRs (solid line), they are nearly accurate.

4.2.3 Performance of VLROP and VLREP detection

The performance of the proposed VLROP and VLREP detection methods is evaluated for clean speech using the same set of speech files, described in the previous Chapter. The only difference lies in the locations of the reference markings. In the present evaluation, for a vowel-semivowel cluster only the beginning and end point of the cluster are considered as the reference VLROP and VLREP, respectively. Using these manually marked references, the performance of the proposed method is measured using three parameters (Chapter 3): Identification rate (IR), Spurious rate (SR) and Identification accuracy (IA). To evaluate robustness in degraded environments, the same set of TIMIT speech files are mixed with four different noises: white, factory, babble and vehicle noise of

Algorithm 1 Correction of hypothesized VLROPs and VLREPs.

$Th_1 = 40\%$ of the maximum VLROP evidence
$Th_2 = 40\%$ of the maximum VLREP evidence

Stage 1: Forced detection

$i = 1, j =$Number of VLREPs
while $i <$ Number of VLROPs **do**
  if VLROP evidence at i^{th} and $(i+1)^{th}$ VLROPs $> Th_1$ **then**
    Search VLREPs between i^{th} and $(i+1)^{th}$VLROP
    if Number of VLREPs found $= 0$ **then**
      Declare immediate valley point after i^{th} VLROP as a VLREP
    end if
  end if
  $i = i + 1$
end while
while $j > 1$ **do**
  if VLREP evidence at j^{th} and $(j-1)^{th}$ VLREPs $> Th_2$ **then**
    Search VLROPs between j^{th} and $(j-1)^{th}$VLREP
    if Number of VLROPs found $= 0$ **then**
      Declare immediate valley point before j^{th} VLREP as a VLROP
    end if
  end if
  $j = j - 1$
end while

Stage 2: Spurious elimination

$i = 1, k =$Number of VLROPs
while $i < k$ **do**
  Find number of VLREPs ($\boldsymbol{P}$) between i^{th} and $(i+1)^{th}$ VLROPs
  if $\boldsymbol{P} = 0$ **then**
    Remove weaker evidence VLROP from the hypothesized list
    $i = i, k = k - 1$
  else
    if $\boldsymbol{P} > 1$ **then**
      Preserve the highest evidence VLREP and remove other VLREPs
    end if
    $i = i + 1$
  end if
end while
if If one or more VLREPs exist after the last VLROP **then**
  Preserve highest evidence VLREP
else
  Remove last VLROP from the hypothesized list
end if
If there is any VLREP before the first VLROP, remove it.

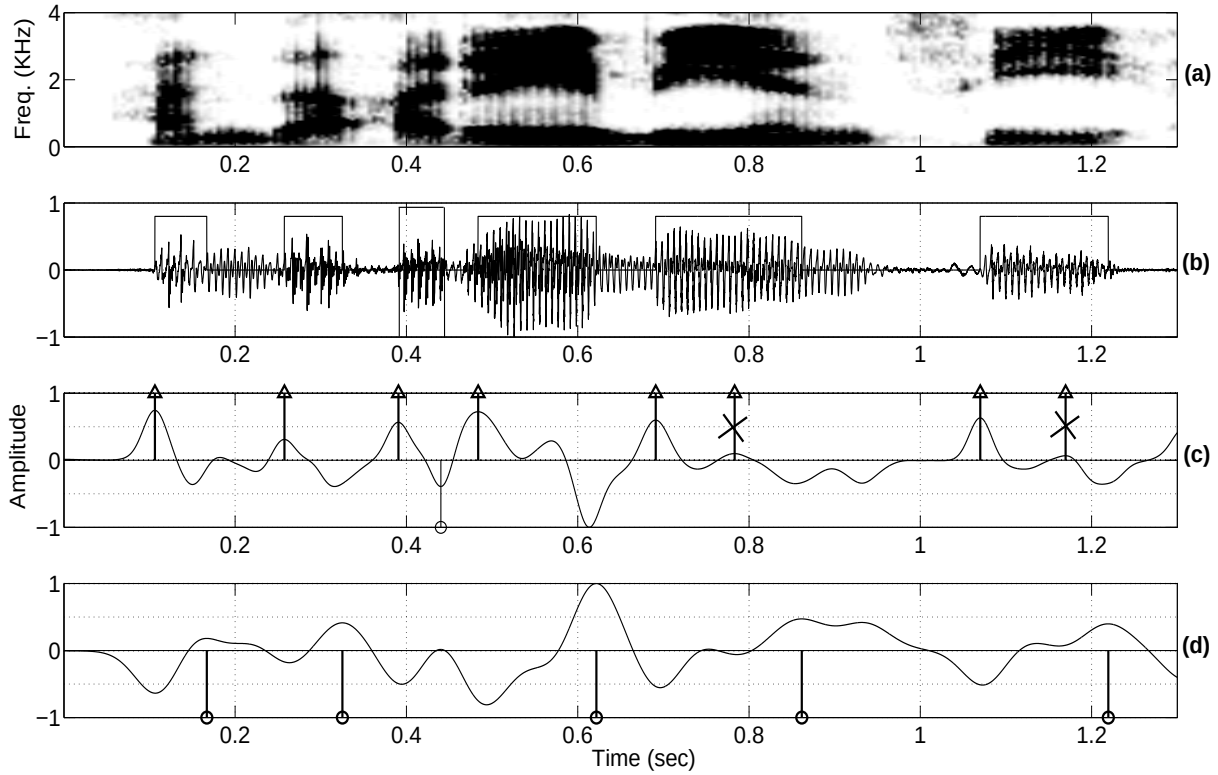

Figure 4.2: Detection of VLRs (a) Wideband spectrogram of the speech signal given in (b), (b) a segment of speech signal (And the Arabian sea) with detected VLRs (solid lines), (c) VLROP evidence using excitation source information with hypothesized VLROPs (arrows), (d) VLREP evidence using excitation source information with hypothesized VLREPs (circles)

the NOISEX-92 database [181]. The energy level of the noise is scaled such that the overall signal to noise ratio (SNR) of the noise added speech is maintained at 15, 10, 5 and 0 dB, respectively. The performance of the proposed VLROP and VLREP detection methods for clean and noise added speech is given in Table 4.1 and Table 4.2, respectively. For comparison, the performance before and after correction is also given in the tables. The performance improvement is measured in terms of net gain (NG) defined as follows:

$$NG = IR_p - IR_e \tag{4.1}$$

where IR_p and IR_e are the identification rate of the proposed and existing methods for the same spurious rate. The existing and proposed methods correspond to VLROP (VLREP) detection without and with correction using VLREPs (VLROPs), respectively.

In case of VLROP detection, the evidence of true VLROPs is much stronger than for the spurious

ones and should be associated with true VLREPs. Based on this logic, the spurious ones are reduced by referring to VLREP evidence. As a result the IR is improved for the same SR. In the case of VLREP detection, the evidence of true VLREPs is relatively weaker compared to that of VLROP evidence and hence the IR is less. However, by using the logic that every VLREP should have a VLROP, the IR is improved significantly. This marginally increases the SR. Both IR and SR are expressed in percentage. The gain in one parameter may result in the loss of performance in the other parameter. Since the relation between IR and SR is nonlinear, the net gain is computed using the relation given in Eqn.(4.1). In this work, the trade-off between SR and IR is computed for 50 different thresholds starting from a very low threshold (0.01) to relatively higher threshold (0.5) for each incremental threshold value of 0.01. It is observed through pilot experiments that the SR of the baseline VLROP detection algorithm roughly varies between 5% to 10% for the clean speech. Therefore the IR and net gain for clean and noise added speech files are computed considering a minimum SR (5%) as the

Table 4.1: Performance of proposed VLROP detection method using excitation source information for speech signals from TIMIT database. The abbreviations IR, SR, IA and NG refer to identification rate, spurious rate, identification accuracy and net gain, respectively.

		Without Correction			**Correction using VLREP**			
Noise	**SNR**	**IR**	**SR**	**IA**	**IR**	**SR**	**IA**	**NG**
Clean speech		94.32	5	8.53	95.15	5	9.25	0.83
White	**15dB**	94.27	5	9.01	95.80	5	9.32	1.53
	10dB	94.14	5	9.93	95.65	5	10.15	1.51
	5dB	93.43	5	10.46	94.05	5	10.49	0.61
	0dB	88.11	5	11.99	91.80	5	12.33	3.69
Factory-1	**15dB**	94.34	5	9.64	95.71	5	10.02	1.37
	10dB	93.75	5	10.36	95.51	5	10.59	1.76
	5dB	89.66	5	11.76	94.68	5	12.17	5.02
	0dB	81.47	5	12.23	87.59	5	13.43	6.12
Babble	**15dB**	94.50	5	10.53	95.47	5	10.76	0.97
	10dB	88.80	5	10.44	92.45	5	11.03	3.65
	5dB	73.32	5	11.54	84.49	5	12.16	11.17
	0dB	64.43	5	13.01	75.06	5	13.38	10.63
Vehicle	**15dB**	94.25	5	9.13	95.58	5	9.54	1.33
	10dB	94.05	5	9.34	95.51	5	9.77	1.46
	5dB	93.77	5	9.61	95.30	5	10.02	1.53
	0dB	92.13	5	10.22	94.90	5	10.56	2.77

Table 4.2: Performance of proposed VLREP detection method using excitation source information for speech signals from TIMIT database. The abbreviations IR, SR, IA and NG refer to identification rate, spurious rate, identification accuracy and net gain, respectively.

		Without Correction			Correction using VLROP			
Noise	**SNR**	**IR**	**SR**	**IA**	**IR**	**SR**	**IA**	**NG**
Clean speech		93.75	5	9.86	96.18	5	10.29	2.43
White	**15dB**	94.17	5	8.91	96.64	5	9.30	2.47
	10dB	93.06	5	8.87	96.32	5	9.54	3.26
	5dB	92.63	5	10.18	96.15	5	10.57	3.52
	0dB	88.07	5	10.37	90.83	5	10.89	2.76
Factory-1	**15dB**	93.14	5	8.55	95.89	5	9.62	2.75
	10dB	92.73	5	9.65	95.82	5	9.87	3.09
	5dB	89.62	5	10.40	95.34	5	10.88	5.72
	0dB	85.73	5	11.23	87.91	5	11.68	2.18
Babble	**15dB**	92.48	5	9.01	95.78	5	9.49	3.30
	10dB	89.05	5	10.19	93.63	5	10.65	4.58
	5dB	86.15	5	11.63	84.27	5	12.25	-1.88
	0dB	81.53	5	12.49	73.93	5	12.72	-7.60
Vehicle	**15dB**	92.95	5	7.82	95.73	5	8.16	2.78
	10dB	92.82	5	7.99	95.69	5	8.47	2.87
	5dB	92.49	5	8.35	95.47	5	8.92	2.98
	0dB	92.23	5	9.64	93.59	5	10.18	1.36

operating point. The net gain is positive in all the cases after correction except for 5 dB and 0 dB babble noise added speech. This may be due to the speech-like nature of babble noise. This shows that VLROPs and VLREPs complement each other for improving the detection of VLRs and hence their significance.

4.3 Detection of non-VLRs and segmentation of speech into VLRs and non-VLRs

Unlike VLRs, non-VLRs contain both voiced and unvoiced speech. The non-VLRs are relatively low energy regions compared to VLRs. In a degraded environment, as the level of noise increases, these regions partially or totally merge with the noise. Therefore, detecting non-VLRs is more challenging than detecting VLRs. After detecting VLRs, non-VLRs may be detected by first classifying the speech signal into speech/non-speech regions (popularly known as voice activity detection (VAD)), and then labelling speech regions as VLRs and non-VLRs. The robustness of non-VLR detection depends on

the robustness of both VAD and VLR detection.

A number of VAD algorithms have been proposed using energy based features [6], long term speech information [8], zero crossing rate [7] and periodicity [5]. The main issue is to find an optimal threshold for speech/non-speech discrimination. Since one threshold is not suitable for both clean and degraded speech, VAD fails in most practical applications. Another group of methods uses statistical models like the Hidden Markov model (HMM) [9], GMM [10], and Laplacian model [11]. Statistical methods aim to construct different classifiers for speech and noise/silence. These methods do not depend critically on the threshold setting, but their performance depends on noise estimation. Most of these methods are initialized based on the assumption that the initial few frames are non-speech frames, which may not be true for all cases. The performance also depends on the choice of probability distribution and speech specific features. The performance is not evaluated separately for VLRs and non-VLRs. For a SV task under degraded conditions, missing some portion of non-VLRs is tolerable compared to spurious detection i.e., selecting non-speech frames as speech (discussed in Section 4.6.2.1). The only requirement is detection accuracy should not critically depend on threshold setting and noise estimation.

4.3.1 Detection of non-VLRs

From the production point of view, non-VLRs may not be as discriminative as VLRs in terms of energy. The rising and falling edges of non-VLRs are less prominent than for VLRs. Therefore, it is difficult to detect these edges directly from the speech signal. If the speech signal is processed to emphasize the non-VLRs, then these edges may be more prominent. One way is to deemphasize the excitation strength in the VLRs, which results in indirect enhancement of the non-VLRs. In continuous speech, the non-VLRs are always either succeeded or preceded by VLRs. If the VLRs are directly deemphasized in the signal domain, the VLROPs and VLREPs may be forced to appear as the begin or end points of non-VLRs due to sudden truncation. Alternatively, if VLRs are deemphasized in the LP residual domain, the sudden truncation effect may be minimized due to the smoothing effect of the resynthesis filter. For the detection of non-VLRs, the begin and end points of a non-VLR cluster are sufficient, not necessarily the start and end points of each sound unit.

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 compute the LP residual signal as shown in Fig. 4.3 (c). Using the detected VLRs (solid line in Fig. 4.3 (b)), the LP residual signal is weighted with the weight

function given in Eqn. (4.2), to obtain the weighted LP residual signal as shown in Fig. 4.3 (d). The normalized weighted LP residual signal is used to excite the LP all pole filter to reconstruct an enhanced non-VLRs speech signal as shown in Fig. 4.3 (e). The reconstructed speech signal is then processed through the algorithm as explained in Section 3.2 to find the *non-VLR evidence*, as shown in Fig. 4.3 (f).

$$\mathbf{w}_f = \begin{cases} 0.05, & \text{VLRs} \\ 1, & \text{Other regions} \end{cases} \tag{4.2}$$

When the smoothed excitation source contour derived from the HE of LP residual is convolved by the FOGD, the degradation portion is significantly deemphasized in the non-VLROP evidence. Alternatively, ZFFS helps to detect the voiced consonants having low energy and reduces the effect of degradation. Therefore combining the evidence obtained from the HE of LP residual with that of ZFFS increases the true detection and reduces the spurious detection.

Unlike VLRs, the signal characteristics at the begin and the end of a non-VLR are approximately the same. Hence, the immediate valley point to the detected peak (starting point) can be considered as the ending point. The detection accuracy and the robustness of proposed non-VLRs detection algorithm can be seen by comparing the non-VLROP evidences given in Figs. 4.3 (f) and (i). The speech signal given in Fig. 4.3 (b) is a segment of speech taken from the IITG-MV speaker recognition database [188] recorded with a head-mounted microphone, which is treated as clean speech for the present work. The speech signal given in Fig. 4.3 (h) is the same segment of speech recorded with a digital voice recorder which is kept at 2 to 3 feet distance from the speaker (degraded speech). By comparing the spectrogram with the detected VLRs (solid line) and non-VLRs (dotted line), it can be stated that the detection accuracy is robust to real environmental degradation. As the evidence is robust, the non-VLRs detection may not critically depend on the threshold setting.

4.3.2 Segmentation of speech into VLRs and non-VLRs

Fig. 4.4 (b) shows a segment of speech taken from the NIST-2003 speaker recognition database [182]. The VLROP, VLREP and non-VLRs evidences are shown in Fig. 4.4 (c), (d) and (e), respectively. The white noise added speech signal for 0 dB SNR is given in Fig. 4.4 (g). The VLROP, VLREP and non-VLRs evidences for noisy speech are shown in Fig. 4.4 (h), (i) and (j), respectively. By comparing the evidences for clean and noisy speech, it can be seen that each evidence is robust

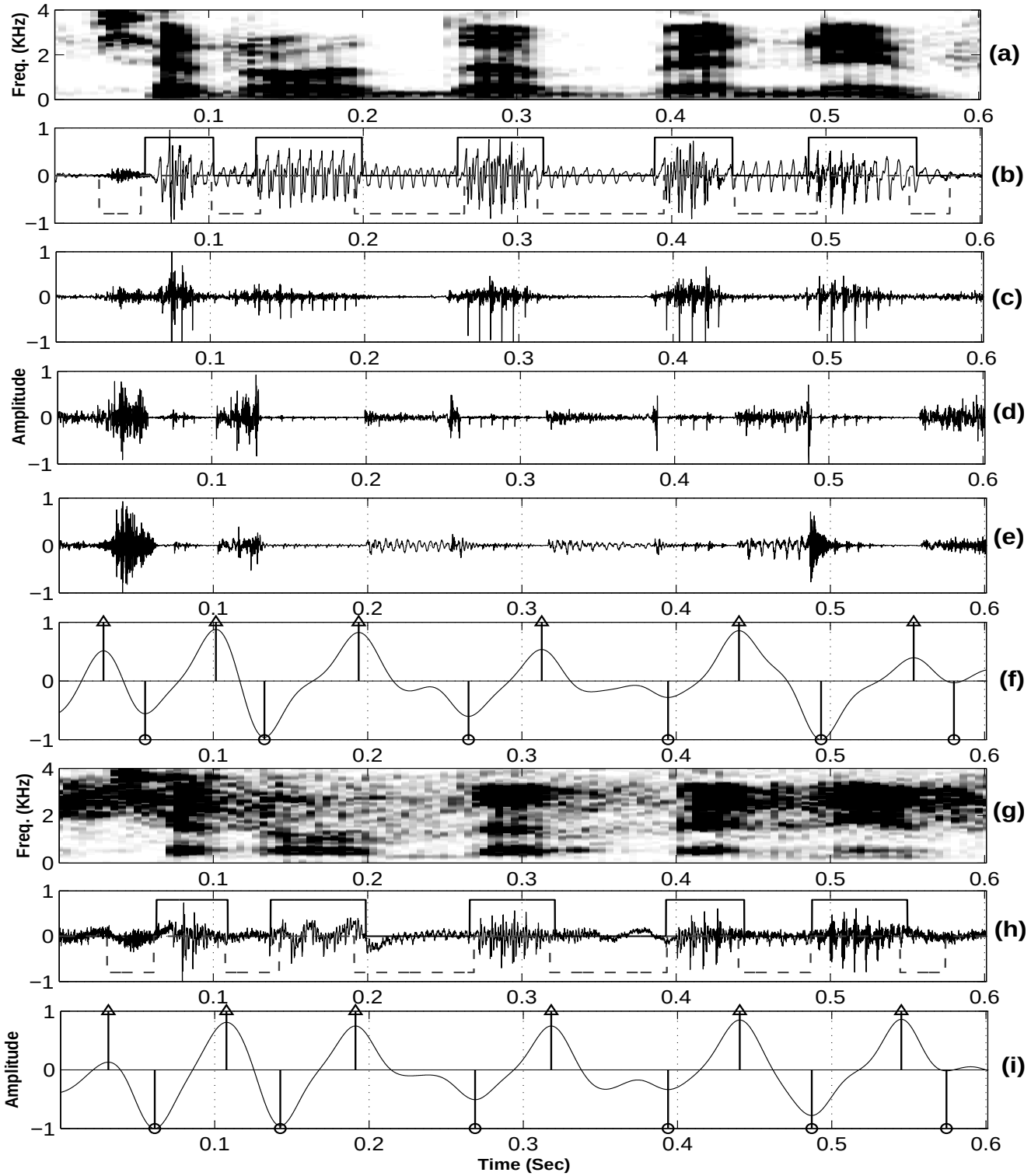

Figure 4.3: Detection of VLRs (solid lines) and non-VLRs (dotted line) for a segment of speech taken from IITG-MV speaker recognition database. (a) Wideband spectrogram of the clean speech, (b) clean speech (speech recorded over sensor H01), (c) LP residual, (d) weighted LP residual, (e) reconstructed speech, (f) non-VLRs evidence for clean speech, (g) wideband spectrogram of the degraded speech, (h) Degraded speech (speech recorded in parallel over sensor D01), (i) non-VLRs evidence for degraded speech.

to degradation and the hypothesized begin and end points nearly correspond to the same instants for clean and degraded speech. For noisy speech, the low SNR non-VLRs are completely merged in the background noise. But, the evidence in these regions is sufficiently higher than the noise-only portion. All the evidences are almost zero for silence and degraded portions. Therefore, the proposed VLR/non-VLR segmentation method is robust.

4.3.3 Performance of VLRs and non-VLRs detection

The performance of the proposed VLRs and non-VLRs segmentation method is evaluated for the same set of speech files and noise levels as explained in Section 4.2.3. The performance is measured in terms of frame error rate using the following parameters:

- *Identification rate (IR)*: The percentage of reference VLR/non-VLR frames that are matched to the detected regions;
- *Speech spurious rate* (SR_1): The percentage of detected VLR frames that are matched with the reference non-VLR regions and vice versa;
- *Non-speech spurious rate* (SR_2): The percentage of detected VLR/non-VLR frames that are matched with the reference non-speech regions.

4.3.3.1 Performance of VLRs detection

The performance of the proposed VLR detection method for clean and noise added speech is given in Table 4.3. For comparison, the performance of VLR detection using only VLROPs is also given in the table (Chapter 3). In both the VLRs detection methods, the IR and SR depends on the SR and the IA of the VLROPs detection method. Therefore, it is reasonable to expect that the SR rate of VLRs detection will be higher than the SR of VLROPs detection. The performance of VLRs detection methods are compared in terms of IR and NG by considering a 10% spurious rate ($SR_1 + SR_2$) as the operating point. The performance of proposed VLRs detection is better than the detection of VLRs using only VLROPs. This result shows that using VLROP and VLREP information significantly enhances the detection of VLRs. Labeling each 100-ms segments after a VLROP as a VLR usually either misses a portion of the VLR or incorrectly labels part of a non-VLR or silence as a VLR, resulting in a higher SR than for the proposed method. The net gain is computed using the relation given in Eqn. (4.1), and it is found to be positive in all cases demonstrating the significance of the proposed method.

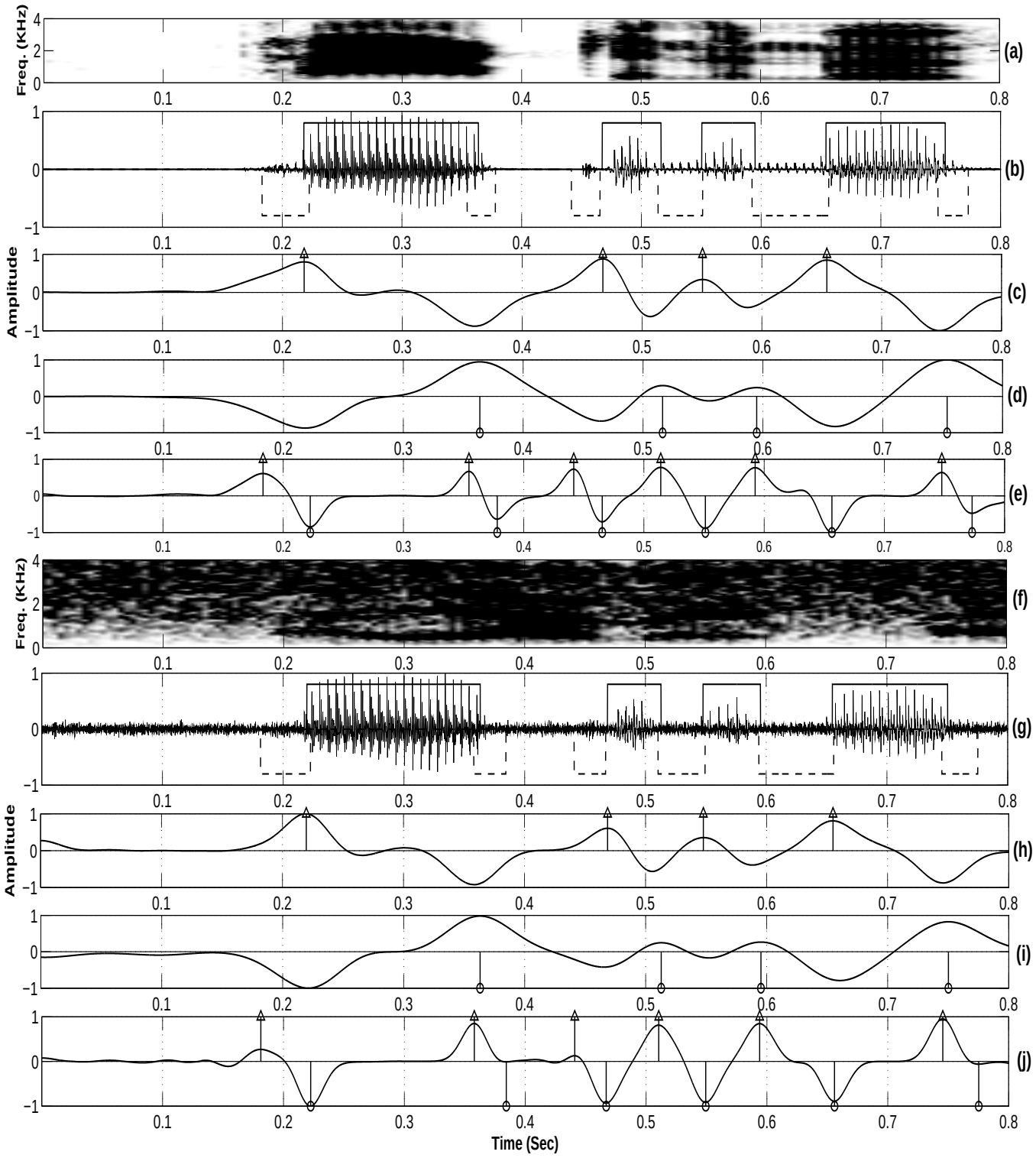

Figure 4.4: Segmentation of speech into VLRs (solid lines) and non-VLRs (dotted line). (a) Wideband spectrogram of the speech signal given in (b), (b) a segment of speech taken from NIST-2003 speaker recognition database, (c) VLROP evidence, (d) VLREP evidence, (e) non-VLRs evidence, (f) Wideband spectrogram for the noise added speech signal, (g) white noise added speech signal with an average SNR of 0 dB, (h) VLROP evidence for the noisy signal, (h) VLREP evidence for the noisy signal, (i) non-VLRs evidence for the noisy signal.

Table 4.3: Performance of VLRs detection method for 10% spurious rate ($SR_1 + SR_2$). The abbreviations IR, MR, SR_1, SR_2 and NG refer to identification rate, miss rate, speech spurious (non-VLRs as VLRs) rate, non-speech spurious rate (silence/noise as VLRs) and net gain respectively.

		VLROP-100 ms			**VLROP & VLREP**			
Noise	**SNR**	IR	SR_1	SR_2	IR	SR_1	SR_2	**NG**
Clean speech		65.44	9.67	0.33	**88.27**	9.64	0.36	24.82
White	**15dB**	65.29	9.81	0.19	**88.43**	9.58	0.42	24.26
	10dB	65.08	9.78	0.22	**87.60**	9.52	0.48	23.47
	5dB	63.34	9.66	0.34	**85.24**	9.41	0.59	23.69
	0dB	58.12	8.78	1.22	**82.51**	8.09	1.91	30.21
Factory-1	**15dB**	64.70	9.74	0.26	**88.96**	9.75	0.25	24.69
	10dB	63.80	9.56	0.44	**87.32**	9.63	0.37	24.80
	5dB	56.82	9.03	0.97	**85.51**	9.22	0.78	27.45
	0dB	48.18	8.68	1.32	**79.83**	8.54	1.46	27.28
Babble	**15dB**	65.09	9.42	0.58	**86.15**	8.47	1.53	25.33
	10dB	61.22	8.51	1.49	**82.66**	8.13	1.87	27.23
	5dB	49.35	7.46	2.54	**73.01**	7.11	2.89	23.62
	0dB	42.42	6.39	3.61	**61.38**	6.61	3.39	19.69
Vehicle	**15dB**	64.53	9.84	0.16	**88.54**	9.77	0.23	25.52
	10dB	63.91	9.81	0.19	**88.16**	8.92	1.08	23.89
	5dB	63.23	9.72	0.28	**86.43**	8.84	1.16	23.96
	0dB	62.74	9.70	0.30	**84.28**	8.73	1.27	22.32

4.3.3.2 Performance of non-VLRs detection

The performance of the proposed non-VLR detection method for clean and noise added speech is given in Table 4.4. For comparison, the performance of non-VLR detection using detected VLRs and energy VAD is also given in the table. For the detection of non-VLRs using VLRs and energy VAD, the speech signal is processed in parallel through an energy based VAD and the VLRs detection algorithm. The speech regions having energy above the threshold of VAD are marked as the speech regions. After identifying the VLRs, the remaining speech regions are treated as non-VLRs. As explained in the Section 4.2.3, the trade-off between the SR ($SR_1 + SR_2$) and IR is computed for 50 threshold values. For both the non-VLRs detection methods, the SR depends on their ability to classify the speech frames from the non-speech frames, and indirectly on the identification accuracy of the VLRs detection method. This is because the VLRs speech frames missed by VLRs detection method may appear as the spurious frames (SR_1) due to their higher signal strength. Therefore the

Table 4.4: Performance of non-VLRs detection method for 20% spurious rate ($SR_1 + SR_2$). The abbreviations IR, MR, SR_1, SR_2 and NG refer to identification rate, miss rate, speech spurious rate (VLRs as non-VLRs), non-speech spurious rate (silence/noise as non-VLRs)and net gain, respectively.

		VLRs & Energy VAD			**Proposed**			
Noise	**SNR**	IR	SR_1	SR_2	IR	SR_1	SR_2	**NG**
Clean speech		72.93	16.92	3.08	**74.97**	16.52	3.48	2.04
White	**15dB**	76.98	14.06	5.94	**80.49**	16.22	3.78	3.51
	10dB	69.56	10.43	9.57	**83.23**	15.28	4.72	13.67
	5dB	56.20	9.82	10.18	**81.36**	14.33	5.68	25.16
	0dB	47.49	9.65	10.35	**74.72**	11.70	8.30	27.23
Factory-1	**15dB**	74.53	12.53	7.47	**79.25**	15.88	4.12	4.72
	10dB	68.35	10.08	9.92	**79.67**	15.22	4.78	11.32
	5dB	53.92	9.88	10.12	**80.26**	14.71	5.29	26.34
	0dB	44.27	9.59	10.41	**76.70**	13.84	6.16	32.43
Babble	**15dB**	69.20	10.60	9.40	**78.91**	16.13	3.87	9.71
	10dB	62.85	9.85	10.15	**71.09**	15.10	4.90	8.24
	5dB	43.37	8.97	11.03	**66.19**	13.77	6.23	22.82
	0dB	36.15	8.25	11.75	**52.05**	12.19	7.81	15.90
Vehicle	**15dB**	69.70	11.77	8.23	**76.46**	18.04	1.96	6.76
	10dB	58.87	10.26	9.74	**76.77**	17.86	2.14	17.90
	5dB	49.51	9.57	10.43	**77.34**	17.47	2.53	27.83
	0dB	43.17	8.84	11.16	**76.57**	17.38	2.62	33.40

IR of both the non-VLRs detection methods are computed by considering 20% SR as the operating point. The net gain is computed similar to the VLRs detection method and found to be positive in all cases. The improvement in the IR for the proposed non-VLRs detection method is due to its robustness to the spurious detection. The detection of silence/noise as non-VLRs is significantly less than the detection of non-VLRs using VLRs and energy VAD. In case of non-VLRs detection using VLRs and VAD, most of the missed VLRs are detected as non-VLRs and for noisy speech due to failure of VAD, most of the non-speech frames are detected as non-VLRs.

4.4 Speaker verification using VLRs and non-VLRs

As described in Section 4.3, VLRs and non-VLRs are identified. In the proposed approach, VLRs and non-VLRs are used independently during training and testing of the SV system. Finally, the scores obtained from VLRs and non-VLRs are weighted and added.

A GMM-UBM based SV system [35] is developed to measure the merit of conditioning and the

effect of speech detection on SV performance. The GMM-UBM based SV systems are developed using the speech frames selected based on the following: energy based VAD, concatenating VLRs and non-VLRs (robust VAD), labeling 100-ms segments following VLROPs as VLRs, and for different mismatched conditions (VLRs train, non-VLRs test and vice versa). The SV system that uses energy based VAD is the *baseline system*. For the baseline SV system, the energy threshold is tuned on the NIST-2002 speaker recognition database. The SV system developed by concatenating the detected VLRs and non-VLRs is called the *SV system without conditioning*.

In order to further investigate the merits of conditioning for the SV task, we have developed a total variability i-vector based SV system [37]. The performance of the i-vector based SV system is compared with and without conditioning on the VLRs and non-VLRs. For an i-vector based SV system, the combination of linear discriminant analysis (LDA) and within class covariance normalization (WCCN) provides better performance than a combination of nuisance attribute projection (NAP) and WCCN [37]. Hence, the performance of i-vector based SV systems are compared using LDA and WCCN as the session/ channel compensation methods.

In this work, the feature extraction (39 dimensional MFCC), feature normalization (CMS followed by CVN) and speaker modeling using GMM-UBM (adapting only the mean parameters of the UBM) are remained same as described in the previous Chapter.

4.4.1 Total variability i-vector based SV system

In the total variability i-vector based SV system [37], the GMM mean supervectors are projected to a low rank matrix to get a reduced dimension representation, which is called as identity vector or i-vector for short. The GMM mean supervector for a speaker utterance ($\boldsymbol{u}$) is created by concatenating the mean vectors of the adapted GMM. The low rank projection matrix represents the dominant speaker and channel variabilities simultaneously and hence is called the total variability matrix. For a given total variability matrix $\boldsymbol{T}$, the i-vector $\boldsymbol{w}$ can be related to the GMM mean supervector $\boldsymbol{M}_s$ as,

$$\boldsymbol{M}_s = \boldsymbol{m} + \boldsymbol{T}\boldsymbol{w} \tag{4.3}$$

where $\boldsymbol{m}$ is the speaker and channel independent supervector (UBM mean supervector).

Given a UBM represented by a weighted sum of C component Gaussian densities as $\boldsymbol{U} = \{\boldsymbol{\mu_c}, \boldsymbol{\Sigma_c}, \boldsymbol{\eta_c}\}$, $c = 1, 2, ..., C$, where $\boldsymbol{\mu_c}$, $\boldsymbol{\Sigma_c}$ and $\boldsymbol{\eta_c}$ are the mean vector, covariance matrix and weight associated

with mixture c, respectively and a sequence of L speech feature vectors $\{\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_L}\}$ of dimension F, the 0^{th} order ($\boldsymbol{N}_c$) and the centralized 1^{st} order ($\boldsymbol{F}_c$) Baum-welch statistics of the speech frames on the c^{th} component of the UBM are given by,

$$\mathbf{N_c} = \sum_{\mathbf{t=1}}^{\mathbf{L}} \mathbf{P(c|x_t, U)} \tag{4.4}$$

$$\boldsymbol{F_c} = \sum_{t=1}^{L} P(c|\boldsymbol{x}_t, \boldsymbol{U})(\boldsymbol{x}_t - \boldsymbol{\mu}_c) \tag{4.5}$$

where, $c = 1, 2, ..., C$ is the component index in the UBM, $P(c|\boldsymbol{x}_t, \boldsymbol{U})$ is the posterior probability of the mixture component c generating the feature vector $\boldsymbol{x}_t$ and $\boldsymbol{\mu}_c$ is the mean of UBM component c.

The learning of the total variability matrix $\boldsymbol{T}$ from the development data and the extraction of i-vectors from the training and test speech utterances are done using the methodology described in [37]. It uses a variant of probabilistic principal component analysis modified to operate on the Baum-Welch statistics of the speech data computed using the UBM. For a given $\boldsymbol{T}$, the estimated i-vector $\hat{\boldsymbol{w}}$ is computed as,

$$\hat{\boldsymbol{w}} = (\boldsymbol{I} + \boldsymbol{T'}\boldsymbol{\Sigma}^{-1}\boldsymbol{N(u)}\boldsymbol{T})^{-1}\boldsymbol{T'}\boldsymbol{\Sigma}^{-1}\boldsymbol{F(u)} \tag{4.6}$$

where $\boldsymbol{N(u)}$ and $\boldsymbol{\Sigma}$ are diagonal matrix of dimension $CF \times CF$ whose diagonal blocks are $N_c\boldsymbol{I}$ and $\boldsymbol{\Sigma_c}$, respectively. $\boldsymbol{F(u)}$ is supervector of dimension $CF \times 1$ generated by concatenating all 1^{st} order Baum-welch statistics ($\boldsymbol{F}_c$) for a given utterance $\boldsymbol{u}$. In the training and testing phases, speech utterances are represented in the form of i-vectors. Let $\hat{\boldsymbol{y}}_{clm}$ and $\hat{\boldsymbol{y}}_{tst}$ represent the i-vectors of the claimed and the test speakers utterances respectively, then the verification of a claim is performed by computing the cosine kernel score between these two i-vectors as,

$$\text{Score} = \frac{< \hat{\boldsymbol{y}}_{clm},\ \hat{\boldsymbol{y}}_{tst} >}{\|\hat{\boldsymbol{y}}_{clm}\|\ \|\hat{\boldsymbol{y}}_{tst}\|} \tag{4.7}$$

4.4.2 Session/channel compensation

The i-vector extracted from the speech utterance contains both channel and speaker variabilities. The performance of an i-vector based SV system can be improved by applying different session/ channel compensation methods. The following are the different session/ channel variability compensation methods used in our experiments.

4.4.2.1 Linear discriminant analysis

In LDA, the feature vectors are projected down to a set of new orthogonal axes where the intra-class variance caused by the channel is minimized and inter-class variance is increased [37, 151]. The projection matrix is composed of the eigen vectors corresponding to the best eigen values of the eigen analysis equation as,

$$(\boldsymbol{W_c^{-1}B_c})\boldsymbol{v} = \boldsymbol{\lambda v} \tag{4.8}$$

where $\boldsymbol{W_c}$ is the within-class covariance matrix, $\boldsymbol{B_c}$ is the between-class covariance matrix, $\boldsymbol{v}$ is an arbitrary vector, and $\boldsymbol{\lambda}$ is the diagonal matrix of eigen values [37].

4.4.2.2 Within class covariance normalization

In WCCN, a set of upper bounds are defined on the classification error metric to reduce the error rate [50]. The feature vectors are transformed using a matrix which minimizes the upper bounds on the classification error metric and hence minimizes the classification error [37]. The transformation matrix $\boldsymbol{B}$ is obtained by Cholesky decomposition of the inverse of the within-class covariance matrix $\boldsymbol{W}$ as, $\boldsymbol{W}^{-1} = \boldsymbol{BB}^t$. As suggested in [37] the best result is obtained when LDA is followed by WCCN. For this purpose $\boldsymbol{W}$ is calculated in the projected space of the LDA.

4.5 Experimental Studies

To verify the merit of conditioning and effect of speech selection on SV performance, the performance of a GMM-UBM based SV system is evaluated on the NIST 2003 speaker recognition database [182] and the IITG-MV speaker recognition database [188]. In order to evaluate the performance of the SV systems under mismatched conditions, the test speech files of NIST-2003 speaker recognition database are mixed with four different noises from the NOISEX-92 database: white, factory, babble and vehicle noise. The energy level of the noise is scaled such that the overall SNR of the noise added speech is maintained at 15, 10, 5 and 0 dB, respectively. For the IITG-MV speaker recognition database, the performance is evaluated by varying the test speech from clean sensor matched to different sensor and environmental mismatched conditions. Due to unavailability of suitable development data for the IITG-MV speaker recognition database, the performance of the *i*-vector based SV system is evaluated only for the NIST-2003 speaker recognition database under original and different noise degraded test conditions.

4.5.1 GMM-UBM based SV system

The main motivation of this work is to find the effect of conditioning on VLRs and non-VLRs on SV performance for clean and different degraded conditions, assuming no *a priori* knowledge about the testing environment. For all the experiments on NIST 2003 speaker recognition database, thirty hours of UBM training speech is selected from randomly selected 250 male and 250 female speech files of switchboard cellular Part 2 Audio database [184]. Using these samples, different 1024-component UBMs are trained [35] for different speech selection methods. For instance, using the speech frames selected by energy based VAD (0.06× average energy), VLRs selected by the proposed approach, concatenating VLRs and non-VLRs, labeling 100 ms segments following VLROPs as VLRs, and finally considering different non-VLRs detection methods. For each system, at the time of model adaptation and testing, the respective UBM is used.

To evaluate the performance of the SV system for real environmentally degraded speech, four sets of experiments are conducted on the IITG-MV speaker recognition database. We use a set of 100 speakers who are speaking conversational English from the IITG-MV database: 75 male speakers and 25 female speakers. To vary the SV condition from clean matched speech to different mismatched degraded speech, speaker models are developed using the initial 2 minutes of speech data recorded with a head-mounted microphone in the first session. The test utterances are 10 utterances from each speaker with durations of 30-45 seconds, recorded in a second session using a head-mounted microphone, tablet PC, mobile phone, and digital voice recorder. Therefore, for the 100 speaker 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. The UBMs are built as described previously using three hours of speech selected from head-mounted microphone recordings.

4.5.2 *i*-vector based SV system

After a detailed analysis of SV performance for different speech segments and the merit of conditioning using a GMM-UBM based SV system, the SV performance is evaluated using VLRs and non-VLRs detected by the proposed method without and with conditioning for an *i*-vector based SV system. The effect of VLRs and non-VLRs detection on the SV performance is evaluated for *i*-vector based SV with LDA and WCCN. For this set of experiments the non-VLRs are also detected using the proposed VLRs and statistical model based VAD [9]. The non-VLR frames are selected by following a

similar procedure as in case of non-VLRs detection using energy VAD and VLRs. The main reason for not including the baseline is due to its lower performance compared to the SV without conditioning (concatenation of VLRs and non-VLRs).

The Switchboard cellular corpus-2 is used as the development data [184]. The UBM remains the same as in the case of the GMM-UBM based SV system. A total variability matrix of rank 400 is created using 1872 speech utterances taken from the development database. For each SV system, the total variability matrix is created using the respective speech frames. The LDA and WCCN matrices are created using the same development data which is used for learning the total variability matrix. For the SV without conditioning, the LDA dimension is fixed at 200. For the SV systems using the VLRs and non-VLRs conditioning, the LDA dimension is fixed at 150. The choice of this dimension is tuned for the best performance. In preliminary experiments, it is observed that for the conditioned systems, the verification accuracy is reduced by combining WCCN with LDA compared to only LDA. This may be due to poor estimation of the WCCN matrix using a particular speech region. To compensate for this, the WCCN matrix is created by weighting and using the i-vectors from the VLRs and non-VLRs. The weight for the class under trial is given as 0.8 and for other class as 0.2. For example, when VLRs are used for training and testing of the SV system, the i-vectors corresponding to VLRs are weighted by 0.8 and i-vectors corresponding to non-VLRs are weighted by 0.2. These weights are tuned on the NIST-2002 speaker recognition database. It is experimentally verified that any weight between 0.7 and 0.9 provides approximately the same performance.

4.5.3 Combination of VLRs and non-VLRs SV system

The scores obtained from VLRs and non-VLRs are combined linearly with more weight to VLRs. Let, for a particular speaker, the match scores obtained using VLRs and non-VLRs be S_{vl} and S_{nvl}, respectively. For each speaker, the match scores (S_c) are combined using a linear weighted sum,

$$S_c = w \times S_{vl} + (1 - w) \times S_{nvl} \tag{4.9}$$

The weight (w) is fixed to 0.6 for all the experiments on both the NIST-2003 speaker recognition database and the IITG-MV speaker recognition database. The weight was tuned by testing different values of w, starting from equal weight (w =0.5) to a comparable higher weight (w =0.8) for each incremental value of 0.05, and using the best. Tuning is done on the NIST-2002 speaker recognition database and the GMM-UBM SV system. The main aim of the present work is to evaluate the SV

performance, assuming knowledge about the degradation is not available. Hence, the weight is fixed for all the experiments.

4.6 Experimental results and discussion

This section tabulates the various experimental results of the SV studies and also discusses the possible reasons for the trends in each case.

4.6.1 GMM-UBM based SV system

This section describes the effect of speech selection and mismatch between the sound units on a SV system. Also, demonstrates the significance of speaker specific information in VLRs, non-VLRs and merits of VLRs and non-VLRs conditioning for a SV task.

4.6.1.1 NIST-2003 Speaker recognition database

The performance of SV systems are evaluated on the complete NIST-2003 speaker recognition database to study the effectiveness of the proposed system on a large population speaker recognition database. The NIST-2003 speaker recognition database mainly contains degradation due to channel and sensor mismatch in some cases. The performance of VLRs and non-VLRs conditioned systems in Table 4.5 show the presence of a good amount of speaker discriminating information in both segment types. This initial result shows that speaker discriminating information in the VLRs is greater than the non-VLRs. But the detected VLRs account for more speech than the non-VLRs. To further investigate speaker specific information in VLRs and non-VLRs, the SV performance is evaluated by limiting the number of VLRs frames to the number of non-VLRs frames. The initial portion of the detected VLRs are used in the training and testing processes. By limiting the number of speech frames, the equal error rate (EER) of the SV system using VLRs is increased from 6.68% to 7.6%. This shows that the performance of the GMM-UBM based SV system depends on the number of frames used for training and testing of the SV system. However, the VLRs are relatively more speaker specific than the non-VLRs. A SV system that conditions on VLRs identified as 100-ms segments following VLROPs is also evaluated and found to not perform as well. This may be due to its using fewer frames of speech. This comparison shows that it is better to detect the complete VLRs and use for speaker recognition.

The performance improvement in the proposed system over the baseline system can be seen in

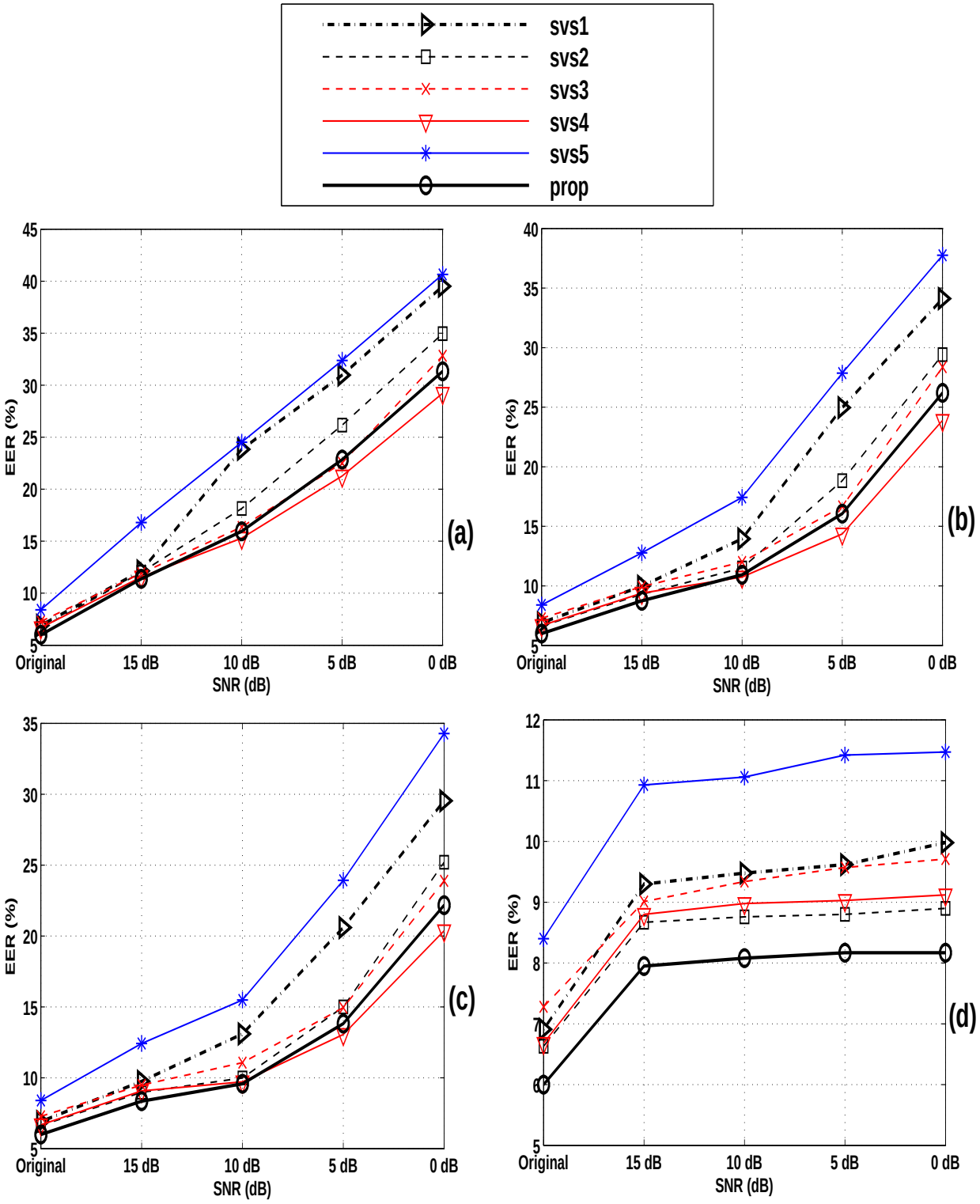

Figure 4.5: Summary of GMM-UBM based SV systems performance (**in EER**) for different experimental setup on NIST-2003 speaker recognition database. (a)-(d) Performance for different SNR level using: White, Factory1, Babble and Vehicle noise added test speech. The abbreviations **svs1-svs5** refer to the SV system using speech selection method as: **svs1** (energy based VAD), **svs2** (concatenation of VLRs and non-VLRs), **svs3** (100-ms segments following VLROPs as VLRs), **svs4** (VLRs using VLROPs and VLREPs) and **svs5** (non-VLRs using proposed method). The abbreviation **prop** refer to the proposed VLRs and non-VLRs combined system.

Table 4.5: Performance of GMM-UBM based SV systems in terms of EER using NIST-2003 speaker recognition database for original & noise added test speech.

		Without conditioning		Conditioning (Train Speech / Test Speech)					VLRs & non-VLRs score combination	
Noise	SNR (dB)	Baseline	Concat. VLRs & non-VLRs (VAD)	VLRs /VLRs (Selected using VLROP-100ms)	VLRs / VLRs	non-VLRs / non-VLRs	non-VLRs / VLRs	VLRs / non-VLRs	w =0.5	w =0.6
Original		6.91	6.63	7.28	6.68	8.4	12.73	13.23	6.09	**6**
White	**15**	12.15	12.01	11.92	11.56	16.8	18.87	22.67	11.47	**11.38**
	10	23.84	18.15	16.36	**15.30**	24.52	21.90	30.62	16.12	15.95
	5	30.98	26.16	22.53	**21.27**	32.38	26.37	36.72	22.99	22.85
	0	39.52	34.96	32.83	**29.22**	40.65	33.83	43.45	31.53	31.34
Factory1	**15**	9.98	9.3	9.94	9.39	12.78	17.84	17.07	8.89	**8.75**
	10	13.95	11.51	12.01	**10.75**	17.43	18.6	22.08	10.98	10.93
	5	24.97	18.83	16.67	**14.36**	27.86	22.17	32.02	16.21	16.07
	0	34.11	29.41	28.36	**23.89**	37.75	29.22	40.65	26.33	26.21
Babble	**15**	9.75	8.98	9.49	9.07	12.42	17.88	15.08	8.49	**8.35**
	10	13.09	9.99	11.06	9.71	15.49	19.06	19.10	9.62	**9.57**
	5	20.59	15	14.95	**13.05**	23.93	21.40	28.13	13.95	13.82
	0	29.53	25.2	23.89	**20.37**	34.28	28.18	37.17	22.35	22.17
Vehicle	**15**	9.30	8.67	9.01	8.8	10.93	17.57	12.28	8.1	**7.95**
	10	9.48	8.76	9.34	8.98	11.06	17.88	12.46	8.22	**8.08**
	5	9.62	8.8	9.57	9.03	11.42	18.02	12.87	8.26	**8.17**
	0	9.98	8.9	9.71	9.12	11.47	18.20	13.32	8.31	**8.17**

Table 4.5. The performance of the baseline system is nearly equal to that of the SV system using all detected VLRs and non-VLRs (robust VAD). The proposed SV system (6 %) performs better than all other SV systems. This shows that conditioning on segment type improves SV performance. In order to further investigate the significance of conditioning, the performance is evaluated for mismatched cases and is less compared to the matched cases. The performance of SV system for non-VLR train and VLR test (12.73%) is better than the VLR train and non-VLR test (13.23%). This result may be due to the smaller degradation of test features derived from VLRs.

4.6.1.2 Noise added NIST-2003 test speech

The performance of SV systems for noise added NIST-2003 test speech files is also given in Table 4.5. From the SV results given in Table 4.5 and Fig. 4.5, as the level of noise increases, the SV performance reduces for each system. But the performance reduction for the SV system using VLRs

Table 4.6: Performance of GMM-UBM based SV systems in terms of EER using IITG-MV speaker recognition database.

		Without conditioning		Conditioning (Train Speech / Test Speech)					VLRs & non-VLRs score combination	
Train	Test	Baseline	Concat. VLRs & non-VLRs (VAD)	VLRs /VLRs (Selected using VLROP-100ms)	VLRs / VLRs	non-VLRs / non-VLRs	non-VLRs / VLRs	VLRs / non-VLRs	w =0.5	w =0.6
Head phone	Head phone	3.71	3.61	3.81	**3.50**	4.81	6.11	10.70	3.70	3.60
	Tablet PC	11.01	10.71	10.61	10.21	12.61	15.90	17.21	10.11	**10.01**
	Mobile Phone	14.71	14.50	14.21	13.91	16.51	19.11	20.30	13.40	**13.30**
	Digital Voice recorder	19.31	16.40	15.31	**13.10**	23.41	18.91	26.91	14.40	14.31

is less than the SV system using the non-VLRs. The performance of the baseline is reduced compared to the SV using detected VLRs and non-VLRs without conditioning. This is mainly due to the failure of the energy based VAD to select the proper speech regions. This again confirms that the detection of VLRs and non-VLRs by the proposed method is robust to different noise conditions. The proposed SV for each noise and SNR gives better performance compared to the same selected speech frames without conditioning. These results show that, under degraded conditions, the SV performance can be improved by proper conditioning without estimating the degradation. Except for the vehicle noise, the performance of only VLRs is better than the proposed system for low SNR speech. This is mainly due to the severe degradation of low SNR non-VLRs. The features derived from these regions deviate from the trained model. As a result, it fails to take full benefit of the conditioning. But, for any real world SV system, this situation is less frequent than the channel and other high SNR conditions.

4.6.1.3 IITG-MV speaker recognition database

The performance of SV systems for the IITG-MV database is given in the Table 4.6. The training speech is clean and the test speech varies from clean sensor-matched to different degradations like noise, reverberation and sensor mismatches. The performance improvement in the proposed system can be seen from the Table 4.6 and Fig.4.6. This result indicates that the proposed approach also provides performance improvement for degradations like sensor mismatch, reverberation and real environmental noise.

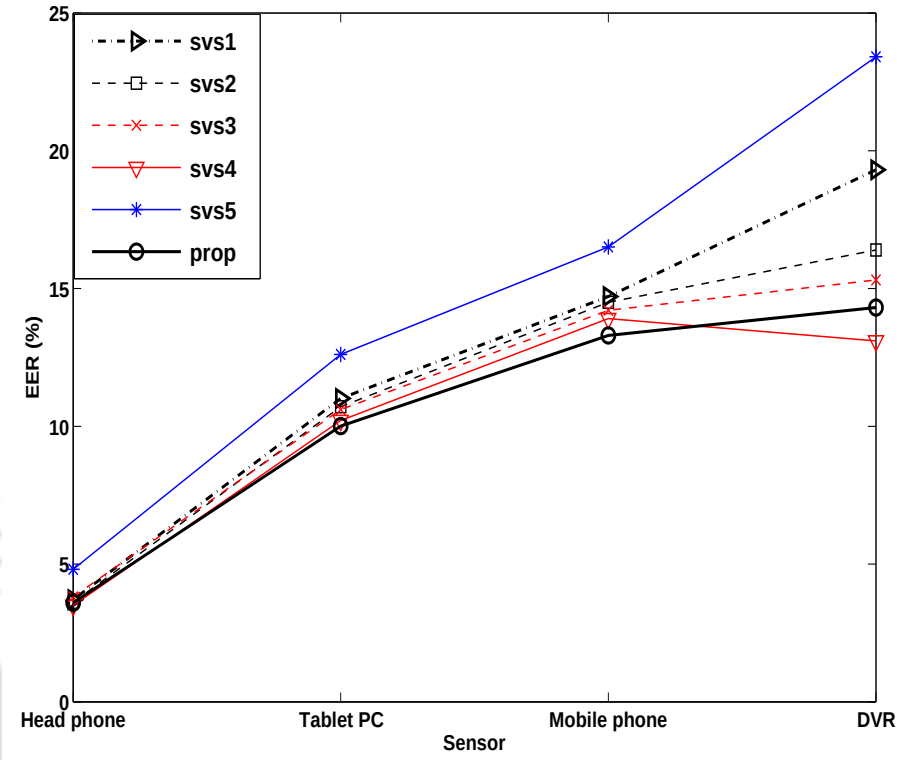

Figure 4.6: Summary of GMM-UBM based SV systems performance (**in EER**) for different experimental setup on IITG-MV speaker recognition database. The abbreviations **svs1-svs5** refer to the SV system using speech selection method as: **svs1** (energy based VAD), **svs2** (concatenation of VLRs and non-VLRs), **svs3** (100-ms segments following VLROPs as VLRs), **svs4** (VLRs using VLROPs and VLREPs) and **svs5** (non-VLRs using proposed method). The abbreviation **prop** refer to the proposed VLRs and non-VLRs combined system.

4.6.2 *i*-vector based SV system

The performance of *i*-vector based SV systems with LDA and WCCN for the original and noise added NIST-2003 speaker recognition database is summarized in Table 4.7. The EER and the decision cost function (DCF) of the *i*-vector based SV systems are also given in Fig. 4.7 and 4.8, respectively. For all experimental conditions considered, the EER and DCF for the proposed VLRs and non-VLRs conditioning is better than the speaker verification without conditioning. For instance, with the NIST-2003 speaker recognition database, the EER improves from 2.29% to 1.89% using the proposed VLRs and non-VLRs conditioning. Similarly, for the 0 dB factory noise case, the improvement is from 21.22% to 17.38%. The proposed SV system provides better performance than the SV systems using only VLRs up to 10, 5, 5 and 0 dB SNR for white, factory, babble and vehicle noise, respectively. For instance, with 0 dB SNR of vehicle noise, the improvement in EER is from 3.34% to 2.84%.

Table 4.7: Performance of *i*-vector based SV systems with LDA and WCCN using NIST-2003 speaker recognition database for original & noise added test speech. Performance is given for different VLRs, non-VLRs methods and score level combination of different VLRs and non-VLRs systems. Performance is given in terms of EER & DCF.

		Performance in terms of EER & (DCF)											
			Conditioning (Train Speech / Test Speech)						VLRs & non-VLRs score combination				
			VLRs train & test		non-VLRs train & test								
Noise	SNR (dB)	Without Cond.	VLROP-100 ms (vl_1)	Prop. Method (vl_2)	VLROP-100 ms & Eng. VAD (nvl_1)	Prop. VLRs & Eng. VAD (nvl_2)	Prop. VLRs & Sta. VAD (nvl_3)	Prop. Method (nvl_4)	vl_1 & nvl_1 w =**0.6**	vl_2 & nvl_2 w =**0.6**	vl_2 & nvl_3 w =**0.6**	VLRs & non-VLRs using Prop. Method w =**0.5**	VLRs & non-VLRs using Prop. Method w =**0.6**
Original		2.29 (0.0108)	2.72 (0.0120)	2.38 (0.0101)	3.16 (0.0129)	4.01 (0.0175)	4.24 0.0167	3.52 (0.0161)	2.12 (0.0092)	2.07 (0.0087)	1.95 (0.0086)	1.94 (0.0090)	**1.89 (0.0085)**
White	**15**	5.55 (0.0265)	5.34 (0.0245)	4.83 (0.0218)	9.43 (0.0426)	12.19 (0.0524)	12.42 (0.0530)	9.89 (0.0427)	5.05 (0.0236)	4.92 (0.0219)	4.75 (0.0218)	5.01 (0.0234)	**4.51 (0.0211)**
	10	8.81 (0.0421)	8.44 (0.0372)	7.04 (0.0338)	24.48 (0.0982)	28.50 (0.0998)	18.06 (0.0733)	16.98 (0.0730)	8.53 (0.0406)	7.81 (0.0382)	7.13 (0.0332)	7.94 (0.0404)	**6.91 (0.0321)**
	5	14.81 (0.0703)	12.64 (0.0598)	**10.98 (0.0533)**	32.56 (0.0998)	34.95 (0.0999)	26.55 (0.0923)	24.84 (0.0945)	13.41 (0.0656)	12.15 (0.0638)	11.74 (0.0562)	13.05 (0.0654)	11.33 (0.0575)
	0	24.29 (0.0981)	22.17 (0.0937)	**19.78 (0.0892)**	40.28 (0.0999)	40.96 (0.0999)	36.08 (0.0099)	35.68 (0.0999)	23.11 (0.0990)	21.68 (0.0992)	21.13 (0.0897)	22.76 (0.0993)	20.86 (0.0955)
Factory1	**15**	4.29 (0.0201)	4.61 (0.0195)	4.06 (0.0173)	7.27 (0.0297)	8.26 (0.0371)	8.94 (0.0399)	7.04 (0.0308)	3.62 (0.0175)	3.57 (0.0158)	3.68 (0.0161)	3.70 (0.0167)	**3.34 (0.0155)**
	10	5.96 (0.0287)	6.54 (0.0261)	5.19 (0.0228)	16.21 (0.0713)	19.19 (0.0839)	12.87 (0.0536)	10.88 (0.0489)	5.93 (0.0289)	5.51 (0.0286)	4.69 (0.0236)	5.32 (0.0254)	**4.56 (0.0225)**
	5	11.38 (0.0540)	10.93 (0.0485)	8.67 (0.0433)	28.13 (0.0998)	30.89 (0.0999)	19.37 (0.0767)	20.73 (0.0875)	11.17 (0.0530)	9.03 (0.0488)	**8.49 (0.0386)**	10.20 (0.0503)	8.53 (0.0390)
	0	21.22 (0.0945)	20.17 (0.0869)	17.02 (0.0736)	36.35 (0.0999)	38.57 (0.1000)	28.50 (0.0961)	32.20 (0.1000)	20.90 (0.0925)	18.06 (0.0859)	**16.53 (0.0718)**	19.42 (0.0887)	17.38 (0.0805)
Babble	**15**	3.97 (0.0183)	4.19 (0.0182)	3.65 (0.0168)	6.62 (0.0289)	8.80 (0.0362)	9.62 (0.0378)	6.36 (0.0285)	3.52 (0.0152)	3.24 (0.0158)	3.38 (0.0151)	3.20 (0.0153)	**3.11 (0.0148)**
	10	5.10 (0.0244)	5.64 (0.0241)	4.83 (0.0205)	14.49 (0.0660)	18.11 (0.0762)	12.91 (0.0529)	9.25 (0.0418)	5.51 (0.0240)	4.83 (0.0219)	4.47 (0.0204)	4.51 (0.0210)	**4.01 (0.0195)**
	5	8.99 (0.0461)	8.94 (0.0394)	7.58 (0.0334)	24.29 (0.0957)	28.45 (0.0999)	19.28 (0.0771)	17.02 (0.0753)	9.25 (0.0427)	8.17 (0.0392)	7.58 (0.0358)	8.22 (0.0393)	**7.42 (0.0331)**
	0	18.42 (0.0855)	17.97 (0.0766)	**14.90 (0.0637)**	33.06 (0.0999)	36.22 (0.0999)	29.04 (0.0960)	28.68 (0.0999)	18.47 (0.0802)	16.26 (0.0735)	15.49 (0.0698)	17.02 (0.0782)	15.17 (0.0696)
Vehicle	**15**	2.98 (0.0150)	3.52 (0.0147)	3.11 (0.0132)	4.89 (0.0201)	5.60 (0.0237)	6.05 (0.0253)	5.32 (0.0234)	2.93 (0.0124)	2.68 (0.0121)	2.80 (0.0122)	2.61 (0.0121)	**2.51 (0.0117)**
	10	2.98 (0.0153)	3.61 (0.0149)	3.21 (0.0133)	5.19 (0.0221)	5.78 (0.0243)	6.54 (0.0261)	5.32 (0.0242)	2.97 (0.0126)	2.73 (0.0120)	2.98 (0.0130)	2.71 (0.0123)	**2.57 (0.0118)**
	5	3.12 (0.0158)	3.76 (0.0157)	3.25 (0.0135)	5.28 (0.0228)	6.18 (0.0252)	7.04 (0.0293)	5.55 (0.0244)	3.18 (0.0143)	2.98 (0.0128)	3.07 (0.0138)	2.75 (0.0127)	**2.61 (0.0121)**
	0	3.53 (0.0168)	3.97 (0.0165)	3.34 (0.0150)	5.78 (0.0244)	6.68 (0.0289)	8.13 (0.0332)	5.55 (0.0258)	3.74 (0.0157)	3.19 (0.0143)	3.26 (0.0148)	2.98 (0.0139)	**2.84 (0.0132)**

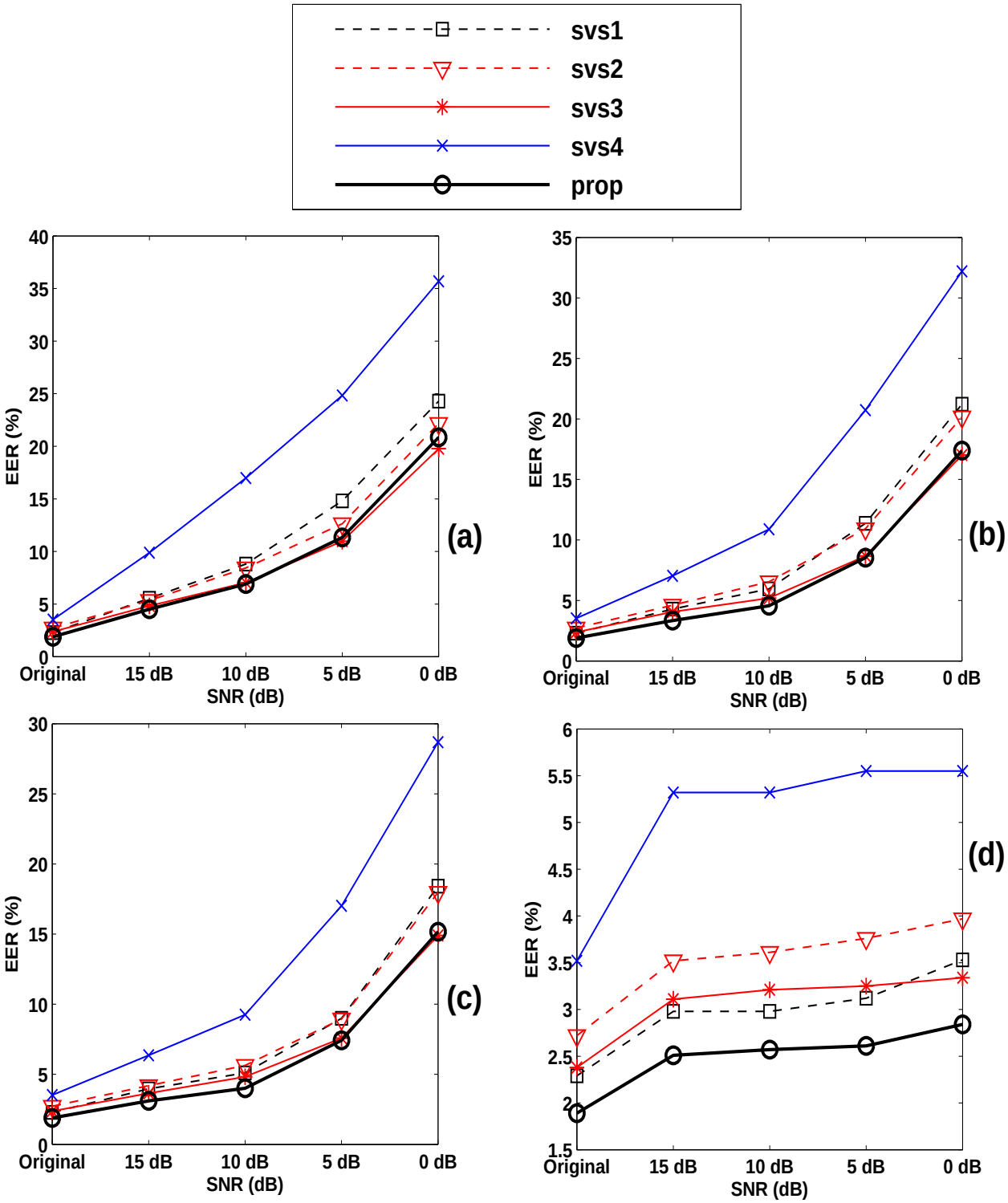

Figure 4.7: Summary of *i*- vector based SV systems performance (**in EER**) for different experimental setup on NIST-2003 speaker recognition database. (a)-(d) Performance for different SNR level using: White, Factory1, Babble and Vehicle noise added test speech. The abbreviations **svs1-svs4** refer to the SV system using speech selection method as: **svs1** (concatenation of VLRs and non-VLRs), **svs2** (100-ms segments following VLROPs as VLRs), **svs3** (VLRs using VLROPs and VLREPs) and **svs4** (non-VLRs using proposed method). The abbreviation **prop** refer to the proposed VLRs and non-VLRs combined system.

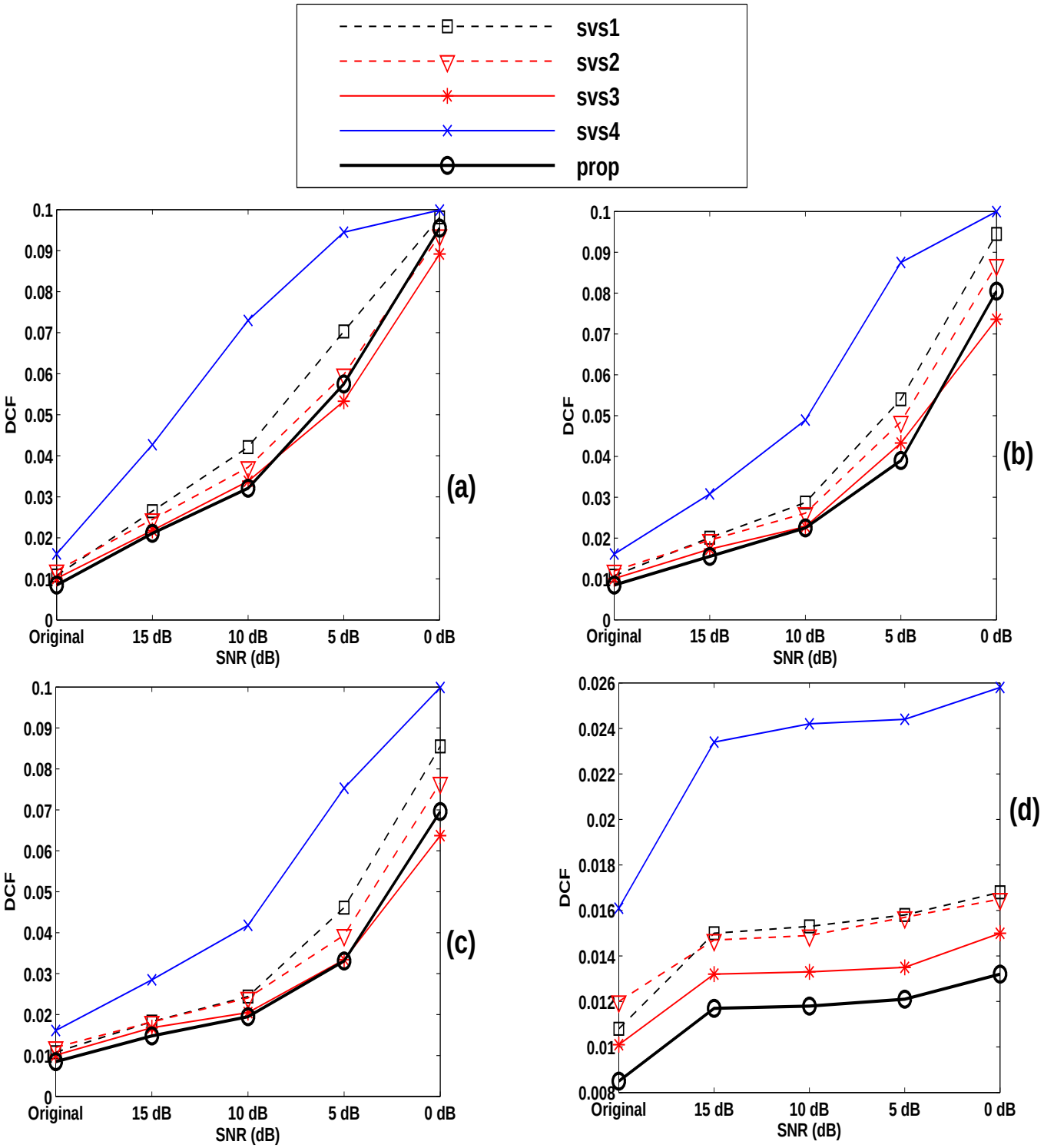

Figure 4.8: Summary of *i*- vector based SV systems performance (**in DCF**) for different experimental setup on NIST-2003 speaker recognition database. (a)-(d) Performance for different SNR level using: White, Factory1, Babble and Vehicle noise added test speech. The abbreviations **svs1-svs4** refer to the SV system using speech selection method as: **svs1** (concatenation of VLRs and non-VLRs), **svs2** (100-ms segments following VLROPs as VLRs), **svs3** (VLRs using VLROPs and VLREPs) and **svs4** (non-VLRs using proposed method). The abbreviation **prop** refer tothe proposed VLRs and non-VLRs combined system.

4.6.2.1 Effect of VLRs and non-VLRs detection on SV performance

For all experimental conditions considered, the GMM-UBM and the *i*-vector based SV systems provide better performance for the VLRs compared to the non-VLRs. As described in Chapter 3, using only VLRs provides better performance than using all speech regions for degraded and mismatched speech. But for clean and lower level degradation, neglecting the non-VLRs reduces SV performance. For instance, with the NIST-2003 speaker recognition database, the *i*-vector based SV system EER increases from 2.29% to 2.38%. Similarly, for the 15 dB vehicle noise case, EER increases from 2.98% to 3.11%. The SV system using the proposed VLRs and non-VLRs conditioning provides better performance for all experimental conditions. For instance, with the NIST-2003 speaker recognition database, the EER of the *i*-vector based SV system improves from 2.29% to 1.89% using the proposed VLRs and non-VLRs conditioning. Similarly, for the 0 dB white, factory, babble and vehicle noise, the improvement is from 24.29% to 20.86%, 21.22% to 17.38%, 18.42% to 15.17% and 3.53% to 2.84%, respectively. This shows that the SV performance for clean and degraded conditions can be improved by applying VLR and non-VLR conditioning.

From the SV results given in Table 4.5, 4.6 and 4.7, the performance of the SV system using VLRs depends on the detection accuracy of the VLRs. For instance, with the NIST 2003 speaker recognition database, the EER of the GMM-UBM based SV system is 6.63% and 7.28%, for VLRs detected by the proposed method and by considering 100-ms segments following VLROPs as VLRs, respectively. Similarly, for the *i*-vector based SV system, the EER is 2.38% and 2.72%.

The performance of the *i*-vector based SV system for different non-VLRs detection methods is given in Table 4.7. The performance of VLRs and non-VLRs combined system for different VLRs and non-VLRs detection methods is also given in the table. The SV performance of the proposed non-VLRs detection method in terms of EER and DCF is better than the non-VLRs selected using the VLRs and VAD. For instance, with NIST-2003 speaker recognition database, the EER is 3.52%, 4.01% and 4.24% for the non-VLRs selected by proposed method, using VLRs and energy VAD and using VLRs and statistical VAD [9]. As reported in [93] this result shows that for clean speech the SV performance of the non-VLRs selected by the energy based VAD is better than that of the statistical VAD. Alternatively for the degraded speech the non-VLRs selected using the statistical VAD performs significantly better compared to that of the energy VAD. The SV performance of the non-VLRs selected by using the VLRs (VLROP-100 ms) and energy VAD is better than the non-VLRs selected by using

the proposed VLRs and energy VAD. This may be due to the missing VLRs which are considered as non-VLRs in the former case. The EER and the DCF of the VLRs and non-VLRs combined system for different VLRs and non-VLRs detection methods are summarized in Fig 4.9 and 4.10, respectively. From Table 4.7, Fig 4.9 and 4.10, for all experimental conditions considered, the SV performance depends on the segmentation of the speech signal into VLRs and non-VLRs. Except for the 5 dB and 0 dB factory noise added test speech, the SV performance of the combined system in terms of EER and DCF is better for the proposed VLRs and non-VLRs detection method. For instance, with NIST-2003 speaker recognition database, the performance in EER improves from 1.94% to 1.89% only for the better classification of non-VLRs and non-speech regions. Similarly, the improvement is from 2.12% to 1.89% for better classification of VLRs and non-VLRs.

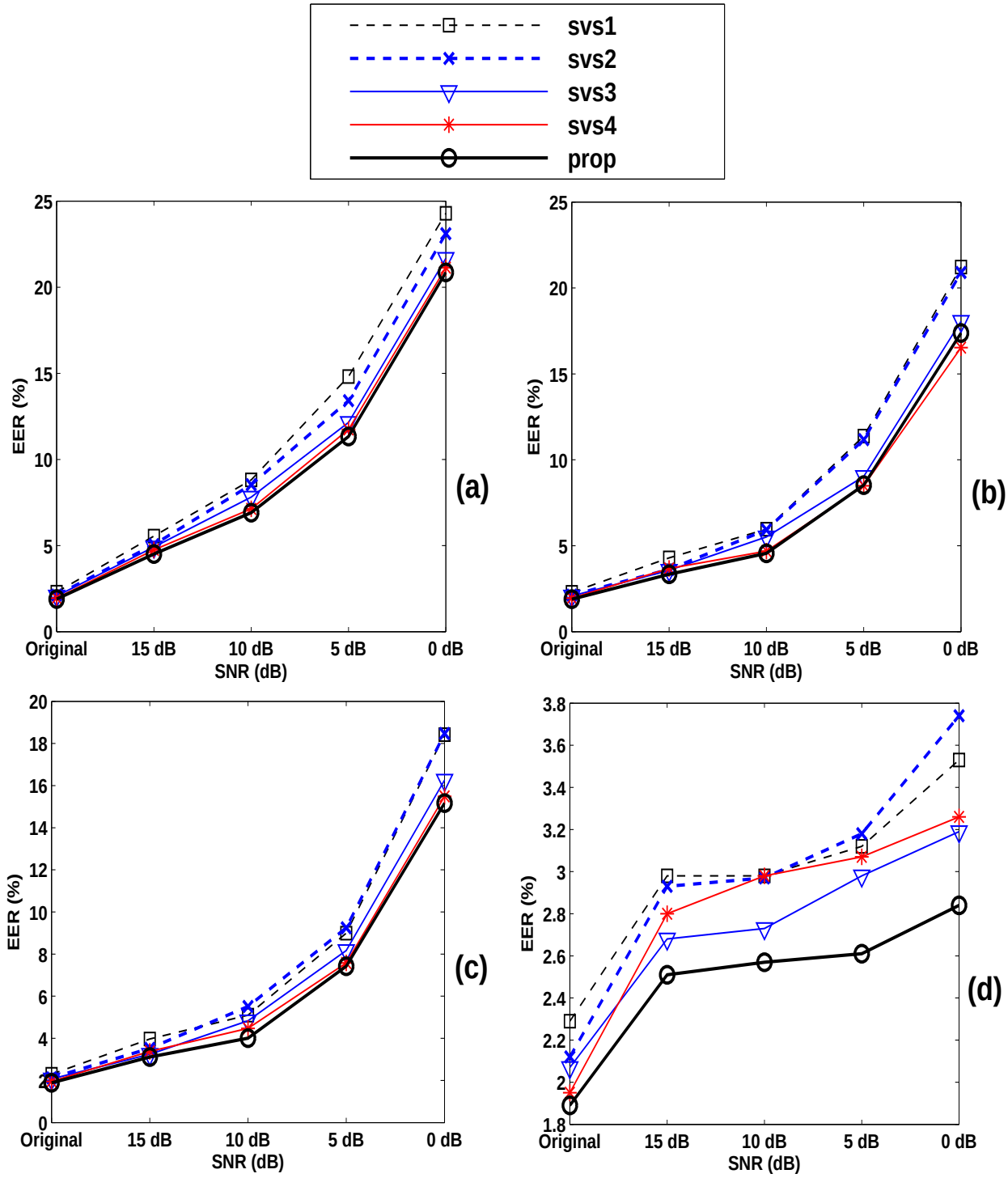

Figure 4.9: SV performance (**in EER**) of VLRs and non-VLRs combined i-vector based SV system for different VLRs and non-VLRs detection methods. (a)-(d) Performance for different SNR level using: White, Factory1, Babble and Vehicle noise added test speech. The abbreviations **svs1** refer to the SV system without conditioning (concatenation of VLRs and non-VLRs). The abbreviations **svs2-svs4** refer to the SV system using VLRs and non-VLRs detection method as: **svs2** (VLRs: 100-ms segments following VLROPs, non-VLRs: energy based VAD), **svs3** (VLRs: using VLROPs and VLREP, non-VLRs: energy based VAD) and **svs4** (VLRs: using VLROPs and VLREP, non-VLRs: statistical model based VAD). The abbreviation **prop** refer to the proposed VLRs and non-VLRs combined system.

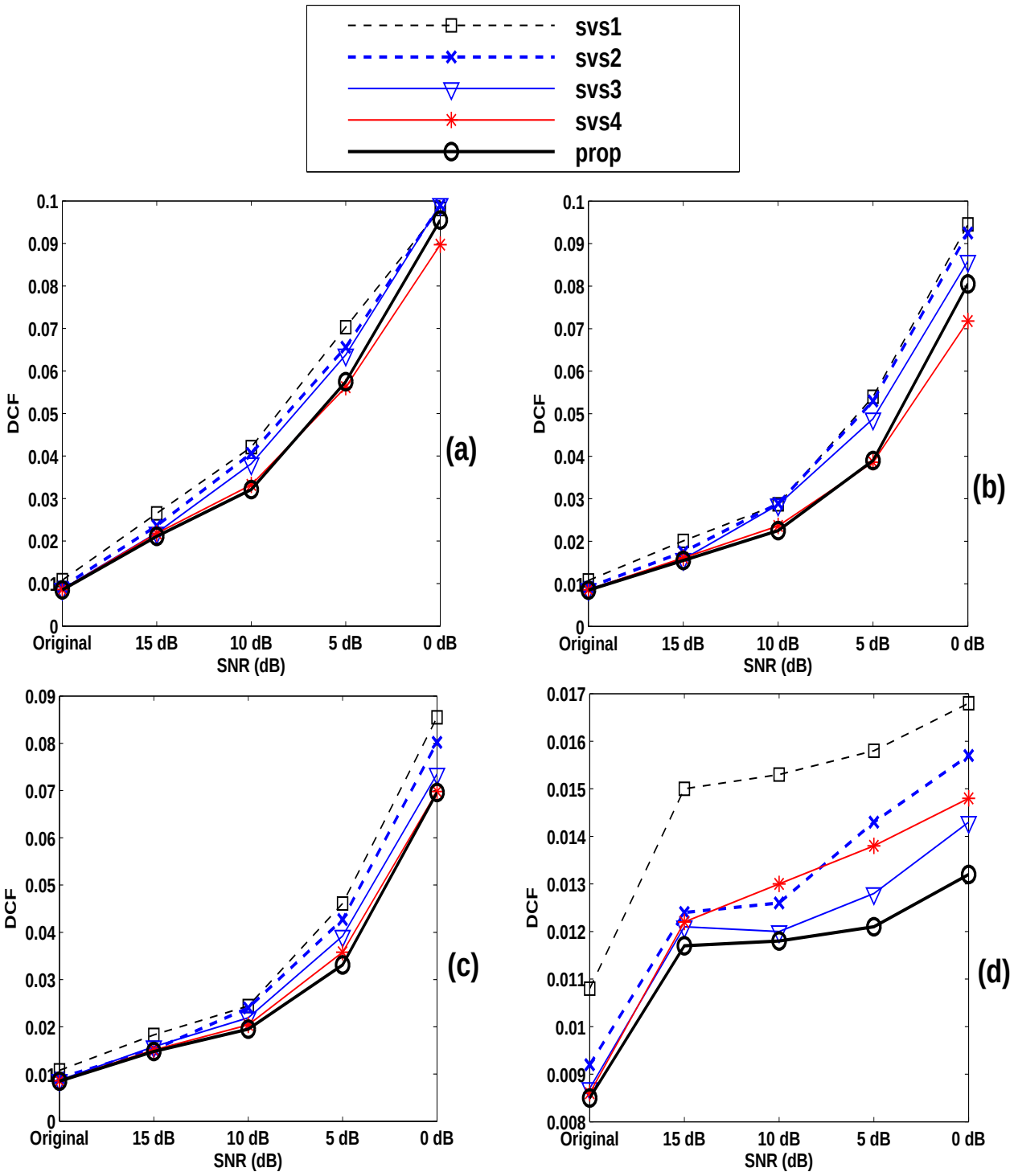

Figure 4.10: SV performance (**in DCF**) of VLRs and non-VLRs conditioned *i*-vector based SV system for different VLRs and non-VLRs detection methods. (a)-(d) Performance for different SNR level using: White, Factory1, Babble and Vehicle noise added test speech. The abbreviations **svs1** refer to the SV system without conditioning (concatenation of VLRs and non-VLRs). The abbreviations **svs2-svs4** refer to the SV system using VLRs and non-VLRs detection method as: **svs2** (VLRs: 100-ms segments following VLROPs, non-VLRs: energy based VAD), **svs3** (VLRs: using VLROPs and VLREP, non-VLRs: energy based VAD) and **svs4** (VLRs: using VLROPs and VLREP, non-VLRs: statistical model based VAD). The abbreviation **prop** refer to the proposed VLRs and non-VLRs combined system.

TH-1196_09610214

4.7 Summary

This work proposed the VLREP event and also methods for the detection of VLRs and non-VLRs. The VLRs are identified by processing the VLROPs and VLREPs through an iterative algorithm. After anchoring the detected VLRs, the LP residual samples are weighted to deemphasize the excitation strength within these regions. The weighted LP residual samples are used to excite the time varying all-pole filter to reconstruct a speech signal with enhanced non-VLRs. The non-VLRs are separated from silence/noise frames by processing the reconstructed speech signal. The VLRs and non-VLRs are used independently during training and testing of a SV system to improve matching between the training model and test features. Finally, the verification scores are combined with more weight to VLRs.

The experimental results presented in this work show that for both GMM-UBM and i-vector based SV systems, the proposed approach consistently performed better than a SV system using the same regions without conditioning. Explicit segmentation of speech into VLRs and non-VLRs is computationally expensive and makes the verification process slow. To avoid explicit segmentation, the next chapter presents a SV system by implicit modeling of VLRs and non-VLRs information in the i-vector.

5

Speaker verification by implicit modeling of vowel-like and non-vowel-like regions

Contents

Objective

As described in the previous chapter, conditioning of VLRs and non-VLRs provides improved performance for both GMM-UBM and *i*-vector based SV systems for clean and degraded speech. Conditioning by an explicit segmentation of speech during training and testing of a SV system is computationally expensive and makes the verification process slow. To avoid the explicit segmentation of speech during training and testing, in this chapter we have attempted a different approach to develop an *i*-vector based SV system by implicit modeling of VLRs and non-VLRs information in the i-vector. A method is proposed to learn VLRs and non-VLRs information in the total variability matrix representing both classes. The novelty of this work is it does not require explicit detection of VLRs and non-VLRs during training and testing process of the SV system. Only the total variability matrix is learned offline using the VLR and non-VLRs information.

5.1 Introduction

It is a well known fact that a SV system provides better accuracy, when the recording and phonetic condition of the training speech matches to that of the test speech [3, 39, 97]. Previous studies using different constraints for selecting speech data suggest that performance of a GMM based SV system can be improved by applying conditioning [66,67,70,81]. These works aim to reduce the variabilities by using similar speech segments for training and testing of a SV system. For example, [81], [70] and [66] uses conditioning based on frequently used words, syllable units and different phoneme units. Detection of constraint regions during training and testing process increases complexity of a SV system, and SV performance depends on the type of constraint and ability for the detection of constraint regions during training and testing process. To avoid complexity involved in multi-class segmentation of speech data, in the previous chapter, speech data is segmented into VLRs and non-VLRs, depending on the signal characteristic and speaker specific information. Experiments verify that independent processing of VLRs and non-VLRs provides significant performance improvement for both GMM-UBM and *i*-vector based SV systems for clean and degraded speech. Both for clean and degraded speech, the *i*-vector based SV system performed better compared to the GMM-UBM based SV system. To further reduce the complexity involved in an explicit segmentation of speech data during training and testing, this chapter presents a VLRs and non-VLRs conditioned *i*-vector based SV system by implicit modeling of VLRs and non-VLRs information in the *i*- vectors.

As described in the previous chapter (section 4.4.1), in an i-vector based SV system the GMM mean supervector for a speaker utterance is projected to a low rank total variability ($\boldsymbol{T}$) matrix to get a reduced dimension representation, which is called as i-vector. If a $\boldsymbol{T}$ matrix is learned in a supervised manner to represent the total variability in VLRs by some particular columns and the total variability in non-VLRs by other columns, then when the speech data supervector is projected to the modified $\boldsymbol{T}$ matrix, it provides flexibility for the observed data to select the most relevant subspace dimensions. As a result, particular dimensions of an i-vector will mostly represent variabilities present in the VLRs and other dimensions mostly the variabilities present in the non-VLRs. These dimensions depend on the number of columns corresponding to VLRs and non-VLRs in the $\boldsymbol{T}$ matrix. During computation of the cosine kernel score, VLRs and non-VLRs dimensions of the test i-vectors are automatically matched to the respective dimensions of the trained i-vectors. This type of implementation may help to develop a VLRs and non-VLRs conditioned SV system without explicit segmentation of VLRs and non-VLRs during training and testing of a SV system. Only a $\boldsymbol{T}$ matrix is learned offline by segmenting the development speech data into VLRs and non-VLRs. Since the development of $\boldsymbol{T}$ matrix is one time process, computational complexity of the VLRs and non-VLRs conditioned SV system will reduce significantly and SV performance will not depend on the detection of VLRs and non-VLRs during training and testing of the SV system.

The rest of the Chapter is organized as follows: SV by implicit modeling of VLRs and non-VLRs information in the i-vectors is described in Section 5.2. The experimental studies are described in Section 5.3. The experimental results are discussed in Section 5.4. The summary of the work are mentioned in Section 5.5.

5.2 Speaker verification by implicit modeling of VLRs and non-VLRs

The motivation behind this work is to implement and investigate an i-vector based SV system by implicit modeling of VLRs and non-VLRs information in the i-vectors. For performance comparison, first a baseline i-vector based SV system is developed without using VLRs and non-VLRs information. Then, two different approaches are considered for modifying the $\boldsymbol{T}$ matrix to represent VLRs and non-VLRs subspace dimensions. In the first approach, subspaces representing VLRs and non-VLRs in the total variability matrix are learned independently. In the second approach, first a total variability subspace is learned using only VLRs (non-VLRs), and then an additional set of subspace dimensions

(that are not represent by the first total variability subspace) are learned using only non-VLRs (VLRs) of the speech data. More specifically, in this implementation the difference between the variabilities present in VLRs and non-VLRs is learned by first finding the similarity between them.

5.2.1 Baseline total variability *i*-vector based speaker verification system

In the total variability *i*-vector based SV system [37], the dimensionality of the GMM mean supervector for a speaker utterance is reduced by projecting it to a lower dimensional subspace. The GMM mean supervector for a speaker utterance is created by concatenating the mean vectors of the adapted GMM. The reduced dimension representation is called the *i*-vector. The GMM mean supervector and *i*-vector representation can be related as,

$$\boldsymbol{M}_s = \boldsymbol{m} + \boldsymbol{T}\boldsymbol{w} \tag{5.1}$$

where $\boldsymbol{M}_s$ is the GMM mean supervector, $\boldsymbol{m}$ is the speaker and channel independent supervector (UBM mean supervector), $\boldsymbol{T}$ is the total variability matrix and $\boldsymbol{w}$ is the *i*-vector representation.

The learning of the $\boldsymbol{T}$ matrix from the development data and the extraction of *i*-vectors from the training and test speech data are done using the methodology described in [37]. It uses a variant of probabilistic principal component analysis modified to operate on the Baum-Welch statistics of the speech data computed using the UBM. The *i*-vector extracted from the speech data contains both speaker and channel variabilities. The performance of an *i*-vector based SV system can be improved by applying different session/ channel compensation methods. Combination of LDA and WCCN is used for the session/ channel compensation following the similar procedure given in [37], described in previous chapter (4.4.1). SV is done by comparing the *i*-vectors corresponding to the test speech and the claimed speaker's training speech using the cosine kernel [37] as,

$$\text{Score} = \frac{< \hat{\boldsymbol{y}}_{clm},\ \hat{\boldsymbol{y}}_{tst} >}{\|\hat{\boldsymbol{y}}_{clm}\|\ \|\hat{\boldsymbol{y}}_{tst}\|} \tag{5.2}$$

where $\hat{\boldsymbol{y}}_{clm}$ and $\hat{\boldsymbol{y}}_{tst}$ are the claimed and the test speaker's *i*-vector, respectively.

5.2.2 Total variability by independent learning of VLRs and non-VLRs subspace dimensions

The motivation behind this part of the work is to understand the dimensions required for VLRs and non-VLRs in the $\boldsymbol{T}$ matrix for better learning the variabilities present in these regions, and the effect of implicit modeling on the SV performance. This study also helps to understand the difference

between the variabilities present in VLRs and non-VLRs. Assuming the parameters representing the subspaces corresponding to VLRs and non-VLRs as different, the total variability corresponding to VLRs and non-VLRs are learned independently and the final $\boldsymbol{T}$ matrix is built by stacking them.

To represent the subspaces corresponding to the VLRs, a total variability matrix of columns $\boldsymbol{R}_{vl}$ is estimated from VLRs segments of the development data using Eq. 5.1 as,

$$\boldsymbol{M}_{vl} = \boldsymbol{m} + \boldsymbol{T}_{vl}\boldsymbol{w}_{vl} \tag{5.3}$$

where $\boldsymbol{M}_{vl}$ is the GMM mean supervector of the VLRs, $\boldsymbol{m}$ is the speaker and channel independent supervector, $\boldsymbol{T}_{vl}$ is the total variability matrix representing the VLRs and $\boldsymbol{w}_{vl}$ is the i-vector representation. The GMM mean supervector is created by adapting the UBM with VLRs portion of the speech data and concatenating the mean vectors of the adapted GMM.

Similarly, a total variability matrix of columns $\boldsymbol{R}_{nvl}$ representing the subspaces corresponding to the non-VLRs is estimated using the non-VLRs of the development data as,

$$\boldsymbol{M}_{nvl} = \boldsymbol{m} + \boldsymbol{T}_{nvl}\boldsymbol{w}_{nvl} \tag{5.4}$$

where $\boldsymbol{M}_{nvl}$ is the GMM mean supervector of the non-VLRs, $\boldsymbol{m}$ is the speaker and channel independent supervector, $\boldsymbol{T}_{nvl}$ is the total variability matrix representing the non-VLRs and $\boldsymbol{w}_{nvl}$ is the i-vector representation. The GMM mean supervector is created by adapting the UBM with non-VLRs portion of the speech data and concatenating the mean vectors of the adapted GMM.

Finally, the joint total variability matrix $\boldsymbol{T_c}$ ($\boldsymbol{T_c} = [\boldsymbol{T}_{vl}|\boldsymbol{T}_{nvl}]$) of columns $\boldsymbol{R}_c$ ($\boldsymbol{R}_c = \boldsymbol{R}_{vl} + \boldsymbol{R}_{nvl}$) is created by stacking the total variability matrices estimated from the VLRs and non-VLRs segments of the speech data, and used for the extraction of the VLRs and non-VLRs conditioned i-vectors as,

$$\boldsymbol{M}_s = \boldsymbol{m} + [\boldsymbol{T}_{vl}|\boldsymbol{T}_{nvl}]\boldsymbol{w}_c \tag{5.5}$$

where $\boldsymbol{M}_s$ is the GMM mean supervectors associated with a speech utterance, $\boldsymbol{m}$ is the speaker and channel independent supervector, $[\boldsymbol{T}_{vl}|\boldsymbol{T}_{nvl}]$ is the total variability matrix representing subspaces corresponding to VLRs and non-VLRs and $\boldsymbol{w}_c$ is the VLRs and non-VLRs conditioned i-vector of dimension $\boldsymbol{R}_c$. The GMM mean supervector ($\boldsymbol{M}_s$) is created without using the VLRs and non-VLRs information.

Similar to the baseline system, combination of LDA and WCCN is used for the session/ channel

compensation. SV is done by comparing the VLRs and non-VLRs conditioned i-vectors corresponding to the test speech and the claimed speaker's training speech using the cosine kernel.

5.2.3 Total variability by dependent learning of VLRs and non-VLRs subspace dimensions

Although signal characteristics of VLRs and non-VLRs are different, they are produced by the same speech production system. Therefore, in the dependent learning, VLRs and non-VLRs columns in the **T** matrix are learned by considering the similarity and difference between them in a manner similar to that in [189]. The difference between the variabilities present in VLRs and non-VLRs is learned by first finding the similarity. Here, the GMM mean supervector for a speech utterance is represented as,

$$\boldsymbol{M}_s = \boldsymbol{m} + \boldsymbol{T}_{vl}\boldsymbol{w}_{vl} + \boldsymbol{T}_{nvl}\boldsymbol{w}_{nvl} \tag{5.6}$$

where, $\boldsymbol{T}_{vl}$ and $\boldsymbol{T}_{nvl}$ are the total variability matrices representing the subspaces corresponding to the VLRs and non-VLRs, respectively and $\boldsymbol{w}_{vl}$ and $\boldsymbol{w}_{nvl}$ are the corresponding i-vector representations.

We basically consider two approaches for dependent learning of VLRs and non-VLRs columns in the $\boldsymbol{T}$ matrix, namely constraint-1 and constraint-2. For the constraint-1 system, we first estimate the $\boldsymbol{T}_{vl}$ matrix of column $\boldsymbol{R}_{vl}$ using the supervectors created using VLRs only, assuming the $\boldsymbol{T}_{nvl}$ as zero matrix. Then the supervectors created using non-VLRs only are projected to the $\boldsymbol{T}_{vl}$ and the residues are used to estimate the $\boldsymbol{T}_{nvl}$ matrix of column $\boldsymbol{R}_{nvl}$. Alternatively for the constraint-2 system, the estimation of $\boldsymbol{T}_{nvl}$ is followed by that of the $\boldsymbol{T}_{vl}$ using the supervectors created using the non-VLRs and VLRs, respectively. Finally, the joint total variability matrix $\boldsymbol{T_c}$ ($\boldsymbol{T_c} = [\boldsymbol{T}_{vl}|\boldsymbol{T}_{nvl}]$) of column $\boldsymbol{R}_c$ ($\boldsymbol{R}_c = \boldsymbol{R}_{vl} + \boldsymbol{R}_{nvl}$) is created by stacking the total variability matrices.

The i-vector representing training and test speech utterances are estimated by finding the projections of the corresponding supervectors to the joint total variability matrix (Eq. 5.5). SV is done by comparing the VLRs and non-VLRs conditioned i-vectors corresponding to the test speech and the claimed speaker's training speech using the cosine kernel.

5.2.3.1 Session/ channel compensation

The session/ channel compensation is done using two different methods. In the first method, session/ channel compensation is performed similar to the baseline system. In the second method, first the within-class and between-class covariance matrices of LDA and within-class covariance matrix

of the WCCN are estimated using i-vectors extracted from the constraint-1 and constraint-2 systems. Then, the LDA and the WCCN matrices are computed by the weighted average of these matrices. More specifically, within-class and between-class covariance matrices of the LDA are computed as,

$$\boldsymbol{W}_c = k_1 \boldsymbol{W}_{k_{vl}} + k_2 \boldsymbol{W}_{k_{nvl}} \tag{5.7a}$$

$$\boldsymbol{B}_c = k_1 \boldsymbol{B}_{k_{vl}} + k_2 \boldsymbol{B}_{k_{nvl}} \tag{5.7b}$$

Where $\boldsymbol{W}_{k_{vl}}$ and $\boldsymbol{B}_{k_{vl}}$ are within-class and between-class covariance matrices estimated using i-vectors of the constraint-1 system. $\boldsymbol{W}_{k_{nvl}}$ and $\boldsymbol{B}_{k_{nvl}}$ are within-class and between-class covariance matrices estimated using i-vectors of the constraint-2 system. k_1 and k_2 are the weights associated with constraint-1 and constraint-2 systems.

Similarly, within-class covariance matrix of the WCCN is computed by the weighted average of the within-class covariance matrices estimated from constraint-1 and constraint-2 systems as,

$$\boldsymbol{W} = k_1 \boldsymbol{W}_{vl} + k_2 \boldsymbol{W}_{nvl} \tag{5.8}$$

Where $\boldsymbol{W}_{vl}$ and $\boldsymbol{W}_{nvl}$ are the within-class covariance matrices estimated using i-vectors of the constraint-1 and constraint-2 system, respectively.

5.3 Experimental Studies

This section describes the speech data and the experimental setup used for different experiments.

5.3.1 Speech data

We have used NIST 2003 [182] and NIST 2012 speaker recognition databases [190] for SV experiments. The Switchboard cellular part 2 Audio database [184] is used as the development database for SV experiments on NIST 2003 speaker recognition database. In NIST 2012 speaker recognition database, most of the target speakers (speakers for whom model will be created) training data is taken from the training and test data of NIST 06, NIST 08 and NIST 10 speaker recognition evaluation (SRE) databases. Therefore, the training data of each target speaker comprises multiple recording sessions. Unlike NIST 2003 SRE, in NIST 2012 SRE knowledge of all target speakers is allowed in computing each trial verification score. To analyze the effect of this knowledge on the system performance, test data are also taken from the new non-target speakers (hypothesized speaker of a test segment is not

in fact the actual speaker). Therefore, it forms two test conditions for the non-target speakers: known (speakers having model) and unknown (speakers not having model). Since, the knowledge of the target speakers is allowed and each speaker have multiple sessions recorded data, we have used training data set as the development data for experiments on NIST 2012 speaker recognition database.

First, performance of each SV system is evaluated on NIST 2003 speaker recognition database to understand the effect of implicit conditioning on the SV performance. Then, the test speech files of NIST-2003 speaker recognition database are mixed with four different noises from the NOISEX-92 database [181]: white, factory, babble and vehicle noise to study the effect of implicit conditioning under mismatched conditions. The energy level of the noise is scaled such that the overall SNR of the noise added speech is maintained at 15, 10, 5 and 0 dB, respectively. Finally, the performance is evaluated on NIST 2012 speaker recognition database for the five test conditions of the core task.

5.3.2 Experimental setup for NIST 2003 speaker recognition database

This section describes the experimental setup for SV experiments on the NIST 2003 speaker recognition database.

5.3.2.1 Processing of speech data

For learning VLRs and non-VLRs subspaces in the **T** matrix, the development data is segmented into VLRs and non-VLRs using the method proposed in the previous chapter (Sections 4.2 and 4.3). SV experiment presented in the previous chapter (Section 4.6) shows that voice activity detection (VAD) by concatenation of VLRs and non-VLRs (robust VAD) provides improved performance compared to the energy based VAD. For all experiments on NIST 2003 speaker recognition database, the VAD is done by concatenating VLRs and non-VLRs. It is to be noted that VLRs and non-VLRs information are not used for processing the training and the test speech data, VLRs and non-VLRs are detected only for a robust VAD.

5.3.2.2 Feature extraction and normalization

The feature extraction (39 dimensional MFCC) and feature normalization (CMS followed by CVN) remained same as described in Chapter 3.

5.3.2.3 Universal background model

For simplicity and better understanding the effect of implicit modeling on the SV performance, the UBM remained same for baseline and all VLRs and non-VLRs condition systems. The same UBM is also used for noise added test conditions. The UBM is developed using approximately thirty hours of speech data, taken from the randomly selected 250 male and 250 female speech utterances of the development database. Using each gender speech, two gender dependent 512 mixture size GMMs are built. Finally, a 1024 mixture size gender independent UBM is built by pooling the two models and normalizing the weights [35]. The UBM is developed without using the knowledge of VLRs and non-VLRs.

5.3.2.4 Total variability matrix

For all experiments on NIST 2003 speaker recognition database, a $\boldsymbol{T}$ matrix of 500 columns is created using 1872 speech utterances taken from the development database. First, to understand the effect of VLRs and non-VLRs subspace dimensions on the SV performance, the VLRs and non-VLRs columns in the $\boldsymbol{T}$ matrix are varied starting from a lower value (50) to a comparable higher value (450) for each incremental value of 50, by preserving the total number of column as 500. This experiment is conducted by independent learning of VLRs and non-VLRs subspace dimensions in the $\boldsymbol{T}_c$ matrix (section 5.2.2). Then, the $\boldsymbol{T}_c$ matrix is constructed by dependent learning of variabilities corresponding to the VLRs and the non-VLRs. As described in section 5.2.3, for the constraint-1 system, the $\boldsymbol{T}_c$ matrix is created by stacking $\boldsymbol{T}_{nvl}$ of 200 columns to $\boldsymbol{T}_{vl}$ of 300 columns. Similarly for the constraint-2 system, $\boldsymbol{T}_c$ matrix is created by stacking $\boldsymbol{T}_{vl}$ of 200 columns to $\boldsymbol{T}_{nvl}$ of 300 columns. For the baseline system, a $\boldsymbol{T}$ matrix having same number of columns (500) is created without using VLRs and non-VLRs information.

5.3.2.5 Session/channel compensation

The LDA and WCCN matrices are created using the same development data which is used for learning the $\boldsymbol{T}$ matrix. For all SV systems, the LDA dimension is fixed at 200. The choice of this dimension is based on the best performance of the baseline system. In case of the dependent learning of VLRs and non-VLRs subspace dimensions, the LDA and WCCN matrices are created by weighted average of these matrices generated from constraint-1 and constraint-2 systems, as described in the previous section. The weight for the SV system under trial is given as 0.8 and for other system as

0.2. For instance, for constraint-1 system, the weights k_1 and k_2 are fixed at 0.8 and 0.2, respectively. Similarly for the constraint-2 system the weights k_1 and k_2 are given at 0.2 and 0.8, respectively. These weights remained same for original and noise added test conditions. It is experimentally verified that any weight between 0.7 and 0.9 provides approximately the same performance.

5.3.2.6 Combination of systems

The combination of different subsystems is done linearly with equal weight to each sub system. Let for a particular speaker, the confidence scores obtained from $\boldsymbol{N}$ subsystems are $\boldsymbol{s}_1$, $\boldsymbol{s}_2$, ... $\boldsymbol{s}_N$, respectively. Then, the combined scores ($\boldsymbol{S}_c$) is calculated as,

$$\boldsymbol{S}_c = \frac{\boldsymbol{s}_1 + \boldsymbol{s}_2 + ... + \boldsymbol{s}_N}{\boldsymbol{N}} \tag{5.9}$$

5.3.3 Experimental setup for NIST 2012 speaker recognition database

This section describes the experimental setup on the NIST 2012 speaker recognition database. The baseline and the VLRs and non-VLRs conditioned systems are subsystems of the IITG Speaker Verification Systems submitted for the NIST 2012 SRE.

5.3.3.1 Detection of speech regions

Processing of very large set of training and test speech data involved in NIST 2012 SRE through VLRs and non-VLRs detection algorithm for robust VAD requires relatively more time compared to an energy based VAD. Therefore, the VAD is done by simple energy based thresholding. For learning VLRs and non-VLRs subspaces in the total variability matrix, the development data is segmented into VLRs and non-VLRs by identifying VLRs within the detected speech regions. After identifying the VLRs, the remaining speech regions are treated as non-VLRs.

5.3.3.2 Feature extraction and normalization

The feature extraction (39 dimensional MFCC) and feature normalization (CMS followed by CVN) remained same as in the case of the NIST 2003 speaker recognition database.

5.3.3.3 Universal background model

For all experiments on NIST 2012 speaker recognition database, we have used gender dependent SV systems. Unlike NIST 2003 SRE, in NIST 2012 SRE the training and test speech contains both telephone and microphone channel recordings. The speech data are recorded in three different recording

conditions: telephone recorded phone call data, microphone recoded interview data and microphone recorded phone call data. The training speech available for the development of speaker models contains telephone channel recording for all the speakers and microphone channel recording for some speakers. For this reason, two gender dependent UBMs are built using telephone channel recorded speech data. The male UBM is built using approximately 40 hours speech data (from 725 male speakers), and the female UBM is built using approximately 50 hours speech data (from 1099 female speakers), taken from the development data. Same UBMs are used for baseline and VLRs and non-VLRs conditioned SV systems.

5.3.3.4 Total variability matrix

For each speaker, multiple speech segments are available for the development of speaker model. The duration of speech segments varies from few seconds to minutes. To maintain uniformity across all i-vectors, for a given speaker and channel, the MFCC features are redistributed to have approximately equal number of feature vectors for each segment. Using the redistributed feature vectors, a $\boldsymbol{T}$ matrix is learned suitable for both telephone and microphone speech in a manner similar to that in [189]. The GMM mean supervector ($\boldsymbol{M_s}$) and i-vector representation ($\boldsymbol{w}$) are related as,

$$\boldsymbol{M_s} = \boldsymbol{m} + [\boldsymbol{T}_p|\boldsymbol{T}_m]\boldsymbol{w} \tag{5.10}$$

where, $\boldsymbol{m}$, $\boldsymbol{T}_p$ and $\boldsymbol{T}_m$ are the UBM mean supervector, total variability matrix using telephone speech and total variability matrix using microphone speech, respectively. First, $\boldsymbol{T}_p$ of 600 columns is learned using the telephone speech. Second, $\boldsymbol{T}_m$ of 200 columns is learned using the microphone speech. Finally, a total variability matrix of 800 columns is built by stacking $\boldsymbol{T}_p$ and $\boldsymbol{T}_m$.

Similar to NIST 2003 speaker recognition database, two VLRs and non-VLRs conditioned total variability matrices are learned using dependent learning. The total variability matrices are learned in a manner to represent the subspaces corresponding to VLRs and non-VLRs and suitable for both telephone and microphone speech. For this, the GMM mean supervector ($\boldsymbol{M_s}$) and i-vector representation ($\boldsymbol{w}$) are related as,

$$\boldsymbol{M}_s = \boldsymbol{m} + [\boldsymbol{T}_{p,vl}|\boldsymbol{T}_{p,nvl}|\boldsymbol{T}_{m,vl}|\boldsymbol{T}_{m,nvl}]\boldsymbol{w} \tag{5.11}$$

where, $\boldsymbol{T}_{p,vl}$ and $\boldsymbol{T}_{p,nvl}$ represent the subspaces corresponding to VLRs and non-VLRs learned using

telephone speech, and $\boldsymbol{T}_{m,vl}$ and $\boldsymbol{T}_{m,nvl}$ represent the subspaces corresponding to VLRs and non-VLRs learned using microphone speech. In constraint-1 system, first the VLRs subspaces corresponding to telephone speech ($\boldsymbol{T}_{p,vl}$) is learned using the supervectors created using VLRs of telephone speech. Then, the supervectors created using VLRs of microphone speech are projected to the $\boldsymbol{T}_{p,vl}$ and the residues are used to estimate the $\boldsymbol{T}_{m,vl}$. The VLRs subspaces corresponding to telephone and microphone speech is represented by stacking $\boldsymbol{T}_{p,vl}$ and $\boldsymbol{T}_{m,vl}$ ($\boldsymbol{T}_{vl} = [\boldsymbol{T}_{p,vl}|\boldsymbol{T}_{m,vl}]$). The supervectors created using non-VLRs of telephone speech are projected to the $\boldsymbol{T}_{vl}$ and the residues are used to estimate the $\boldsymbol{T}_{p,nvl}$. Finally, the supervectors created using non-VLRs of microphone speech are projected to the $\boldsymbol{T}_1$ ($\boldsymbol{T}_1 = [\boldsymbol{T}_{p,vl}|\boldsymbol{T}_{p,nvl}|\boldsymbol{T}_{m,vl}]$) and the residues are used to estimate the $\boldsymbol{T}_{m,nvl}$. We have learned 350, 250, 100 and 100 subspace dimensions corresponding to the $\boldsymbol{T}_{p,vl}$, $\boldsymbol{T}_{p,nvl}$, $\boldsymbol{T}_{m,vl}$ and $\boldsymbol{T}_{m,nvl}$, respectively. For constraint-2 system, the estimation of non-VLRs subspaces is followed by that of the VLRs subspaces. We have learned 250, 350, 100 and 100 subspace dimensions corresponding to the $\boldsymbol{T}_{p,vl}$, $\boldsymbol{T}_{p,nvl}$, $\boldsymbol{T}_{m,vl}$ and $\boldsymbol{T}_{m,nvl}$, respectively.

5.3.3.5 Session/channel compensation

The LDA and WCCN matrices are created using the same development data which is used for learning the total variability matrix. For each SV system, the LDA dimension is fixed at 200. In case of the VLRs and non-VLRs conditioned systems, LDA and WCCN matrices are created by weighted average of within-class and between-class covariance matrices of the constraint-1 and constraint-2 systems, as in the case of the NIST 2003 speaker recognition database.

5.3.3.6 Scoring and Combination of systems

In NIST 2012 SRE database, for all the speakers at least one telephone channel recorded speech segment is available and the microphone channel recorded speech data is available for some speakers. Depending on the availability of speech data, the telephone and microphone models for each speaker is created by taking the average of i-vectors generated from each segment (after redistribution of feature vectors) of the respective channel. During testing, the telephone test segments are scored against telephone models. The microphone test segments, are scored against the microphone models, if the microphone model is available otherwise scored against telephone model. For NIST 2012 SRE, the final score is required to represent the systems estimate of the log-likelihood ratio (i.e., the natural logarithm of the target/non-target likelihood ratio). For this, the cosine kernel scores generated by

each system is mapped to log-likelihood ratios and the combination of system is performed using the BOSARIS toolkit [191], which uses the linear logistic regression method for the same. Tunning of the system parameters and calibration is done using a development data set, derived from the actual training data of NIST 2012 SRE. A part of training data is used for the development of speaker models and rest are used for testing. Using 4000 test segments, 80000 test trials are performed for different test conditions.

5.4 Experimental results and discussion

This section tabulates the various experimental results and also discusses the possible reasons for the trends in each case.

5.4.1 NIST 2003 speaker recognition database

This section describes experimental results on NIST 2003 speaker recognition database for dependent and independent learning of VLRs and non-VLRs subspace dimensions.

5.4.1.1 Independent learning of VLRs and non-VLRs subspace dimensions

The Table 5.1 shows the performance of the VLRs and non-VLRs conditioned SV system in terms of equal error rate (EER) and minimum detection costs (DCF). In this experiment, the implicit modeling of VLRs and non-VLRs dimensions in the i-vectors is done by independent learning of VLRs and non-VLRs subspaces in the total variability matrix. For a comparison, the performance of the baseline system for the same dimension of i-vectors is also given in the table. This result shows that learning of VLRs and non-VLRs subspaces separately within the total variability matrix provides improved performance compared to the baseline system. As expected, the performance of the VLRs and non-VLRs conditioned system depends on the subspace dimension of these regions within the total variability matrix. The EER and DCF of the conditioned system is relatively more for higher values of VLRs/ non-VLRs subspace dimensions. This may be due to dominance of one subspace in the total variability matrix. But, for a higher subspace dimension of VLRs, performance is better compared to the same higher subspace dimension of non-VLRs. For instance, for VLRs and non-VLRs subspace dimensions 450 and 50, EER and DCF obtained are 5.42% and 0.0208, respectively. Alternatively, for VLRs and non-VLRs subspace dimensions 50 and 450, EER and DCF obtained are 6.82% and 0.0251, respectively. From the Table 5.1, it may be observed that the best EER and DCF are obtained

Table 5.1: Performance of the VLRs and non-VLRs conditioned SV system using independent learning of VLRs and non-VLRs subspace dimensions. Performance is given in terms of EER and minimum DCF on NIST 2003 speaker recognition database without session/channel compensation. For comparison, the performance of the baseline system is also given.

Dim. of T_{vl}	Dim. of T_{nvl}	EER	DCF
Baseline		5.69	0.0231
450	50	5.42	0.0208
400	100	5.05	0.0194
350	150	4.96	0.0189
300	200	**4.83**	**0.0188**
250	250	4.98	0.0207
200	300	5.12	0.0211
150	350	6.09	0.0225
100	400	6.45	0.0247
50	450	6.82	0.0251

for 300 and 200 subspace dimensions of VLRs and non-VLRs, respectively. The EER (4.83%) and DCF (0.0188) obtained for these subspace dimensions of VLRs and non-VLRs is better compared to the EER (5.69%) and DCF (0.0231) of the baseline system. This first result shows that a VLRs and non-VLRs conditioned SV system is possible by implicit learning of VLRs and non-VLRs subspace dimensions.

The performance of SV systems with session/ channel compensation is given in the Table 5.2. Depending on the SV performance without session/ channel compensation, for the VLRs and non-VLRs conditioned SV system performance is given for the four best subspace dimensions of VLRs and non-VLRs. With application of the LDA, EER and DCF of the baseline system are improved from 5.69% to 2.84% and 0.0231 to 0.0123, respectively. The combination of LAD and WCCN further improved EER and DCF to 2.34% and 0.0105, respectively. In case of the VLRs and non-VLRs conditioned system, similar to the performance before session/ channel compensation, the best EER and DCF are obtained for 300 and 200 subspace dimensions of VLRs and non-VLRs, respectively. With application of LDA, EER and DCF of the conditioned system (for best subspace dimension of VLRs and non-VLRs) are improved from 4.83% to 4.15% and 0.0188 to 0.0171, respectively. The combination of LDA and WCCN, improved EER and DCF to 3.34% and 0.0144, respectively. For

Table 5.2: Performance of the VLRs and non-VLRs conditioned SV system using independent learning of VLRs and non-VLRs subspace dimensions. Performance is given in terms of EER and minimum DCF on NIST 2003 speaker recognition database using LDA and combination of LDA and WCCN as session/channel compensation methods. For comparison, the performance of the baseline system is also given.

		LDA		**LDA+WCCN**	
Size of T_v	**Size of T_{nv}**	**EER**	**DCF**	**EER**	**DCF**
Baseline		**2.84**	**0.0123**	**2.34**	**0.0105**
350	150	4.24	0.0173	3.56	0.0148
300	200	4.15	0.0171	3.34	0.0144
250	250	4.29	0.0184	3.45	0.0152
200	300	4.37	0.0189	3.61	0.0156

each of the session/ channel compensation methods, the relative performance improvement in case of the baseline system is more compared to that of the VLRs and non-VLRs conditioned SV system. As a result, the performance of VLRs and non-VLRs conditioned SV system in terms of EER and DCF is poorer compared to the baseline system. This results infers that matching of the VLRs and non-VLRs dimensions of the test i-vectors to the respective dimensions of the trained i-vectors during cosine kernel scoring fails due to rotation of i-vector space by LDA and WCCN.

5.4.1.2 Dependent learning of VLRs and non-VLRs subspace dimensions

In this experiment the implicit modeling of VLRs and non-VLRs dimensions in the i-vector is done by dependent learning of VLRs and non-VLRs subspaces in the total variability matrix. The Table 5.3 shows the performance of VLRs and non-VLRs conditioned SV systems in terms of EER and DCF. For comparison, the performance of the baseline system is also given in the table. As in case of the independent learning, the performance of the VLRs and non-VLRs conditioned SV systems in terms of EER and DCF is better compared to the baseline without session/ channel compensation. With application of LDA, EER of the constraint-1 and constraint-2 systems are improved from 4.67% to 2.98% and 4.56% to 2.91%, respectively. Similarly, the DCF is improved from 0.0191 to 0.0133 and 0.0180 to 0.0123 for constraint-1 and constraint-2 systems, respectively. The combination of LDA and WCCN, further improved EER and DCF for both the constrained systems. By comparing Tables 5.2 and 5.3, it can be observed that without and with session/ compensation the VLRs and non-VLRs

Table 5.3: Performance of the VLRs and non-VLRs conditioned SV system using dependent learning of VLRs and non-VLRs subspace dimensions. Performance is given in terms of EER and minimum DCF on NIST 2003 speaker recognition database without and with session/ channel compensation. For comparison performance of the baseline system is also given.

			LDA				LDA+WCCN			
	i-vector		General		weighted		General		weighted	
SV System	**EER**	**DCF**	**EER**	**DCF**	**EER**	**DCF**	**EER**	**DCF**	**EER**	**DCF**
Baseline	5.69	0.0231	2.84	0.0123	-	-	2.34	0.0105	-	-
constraint-1	4.67	0.0191	2.98	0.0133	2.30	0.0103	2.89	0.0133	2.21	0.0102
constraint-2	**4.56**	**0.0180**	2.91	0.0123	**2.27**	**0.0098**	2.80	0.0127	**2.16**	**0.0097**

conditioned system outperformed for dependent learning compared to independent learning. Although the independent rotation of VLRs and non-VLRs subspace by LDA and WCCN is reduced up to an extent in the dependent learning, the performance of the baseline system is still better compared to the VLRs and non-VLRs condition systems.

Using LDA and WCCN matrices computed by the weighted average approach, the performance in terms of EER and DCF is improved for both constraint-1 and constraint-2 systems. For instance, using LDA computed by the weighted average method, EER and DCF of constraint-1 system is improved to 2.30% and 0.0103, respectively. Similarly, EER and DCF of constraint-2 system is improved to 2.27% and 0.0098, which are better compared to the EER (2.84%) and DCF (0.0123) of the baseline system. From the Table 5.3, it can be observed that using combination of LDA and WCCN as session/ channel compensation method, the VLRs and non-VLRs conditioned systems are also outperformed compared to the baseline system.

The performance of VLRs and non-VLRs condition systems and baseline system for different noise added test conditions of NIST 2003 speaker recognition database are summarized in Table 5.4. In this experiment, the session/ channel compensation is performed by weighted average approach. As described in the previous section, in constraint-1 system the first 350 dimensions represents the variabilities present in the VLRs and the variabilities in the non-VLRs similar to that of the VLRs. The remaining 250 dimensions represents the the variabilities present only in the non-VLRs. Alternatively, in constraint-2 system, the first 350 dimensions represents the variabilities present in the non-VLRs and the variabilities in the VLRs similar to that of the non-VLRs. The remaining 250 dimensions

Table 5.4: Summary of the SV performance for noise added test speech of NIST 2003 speaker recognition database. Performance is given in terms of EER and minimum DCF without and with session/ channel compensation.

		Performance in terms of EER & (min.DCF)									
		i-vector					i-vector with LDA & WCCN				
Noise	**SNR**	**Baseline** (S_1)	**const.-1** (S_2)	**const.-2** (S_3)	S_1&S_2 **Comb.**	S_1, S_2&S_3 **Comb.**	**Baseline** (S_1)	**const.-1** (S_2)	**const.-2** (S_3)	S_1&S_2 **Comb.**	S_1, S_2&S_3 **Comb.**
Original		5.69 (0.0231)	4.67 (0.0191)	4.56 (0.0180)	4.15 (0.0175)	4.29 (0.0179)	2.34 (0.0105)	2.21 (0.0102)	2.16 (0.0097)	1.94 (0.0091)	1.71 (0.0088)
White	**15**	9.16 (0.0433)	8.85 (0.0423)	8.94 (0.0386)	8.26 (0.0385)	7.67 (0.0386)	5.51 (0.0256)	5.01 (0.0235)	4.83 (0.0240)	4.38 (0.0215)	4.06 (0.0209)
	10	13.36 (0.0632)	12.55 (0.0606)	12.05 (0.0542)	11.56 (0.0546)	10.93 (0.0552)	8.6 (0.0403)	7.85 (0.0350)	7.54 (0.0345)	6.72 (0.0318)	6.41 (0.0313)
	5	20.05 (0.0871)	18.42 (0.0838)	17.34 (0.0784)	16.75 (0.0794)	16.89 (0.0810)	14.67 (0.0669)	13.14 (0.0615)	12.10 (0.0575)	11.74 (0.0565)	11.33 (0.0556)
	0	29.72 (0.0987)	27.91 (0.0989)	27.64 (0.0985)	26.64 (0.0986)	26.91 (0.0985)	25.06 (0.0983)	23.21 (0.0972)	22.44 (0.0949)	21.72 (0.0954)	21.40 (0.0966)
Factory-1	**15**	8.13 (0.0344)	7.27 (0.0318)	6.82 (0.0297)	6.36 (0.0288)	6.23 (0.0282)	4.38 (0.0197)	3.65 (0.0174)	3.70 (0.0175)	3.29 (0.0160)	3.16 (0.0151)
	10	9.84 (0.0435)	9.34 (0.0416)	8.49 (0.0383)	8.08 (0.0378)	8.08 (0.0375)	5.96 (0.0284)	5.23 (0.0235)	5.28 (0.0248)	4.78 (0.0222)	4.51 (0.0210)
	5	16.07 (0.0703)	15.04 (0.0680)	13.55 (0.0613)	13.27 (0.0612)	13.09 (0.0624)	11.51 (0.0518)	9.53 (0.0452)	9.89 (0.0450)	8.89 (0.0413)	8.67 (0.0413)
	0	26.91 (0.0980)	24.20 (0.0980)	24.20 (0.0970)	22.76 (0.0971)	22.99 (0.0970)	21.81 (0.0913)	19.19 (0.0844)	19.42 (0.0817)	17.88 (0.0805)	17.88 (0.0814)
Babble	**15**	7.36 (0.0322)	6.95 (0.0295)	6.36 (0.0271)	6.18 (0.0261)	6.01 (0.0257)	3.83 (0.0181)	3.61 (0.0164)	3.52 (0.0159)	3.07 (0.0148)	3.02 (0.0137)
	10	9.30 (0.0392)	8.71 (0.0373)	8.03 (0.0343)	7.76 (0.0330)	7.45 (0.0324)	5.10 (0.0241)	4.69 (0.0212)	4.51 (0.0215)	3.88 (0.0192)	3.56 (0.0183)
	5	13.73 (0.0609)	13.23 (0.0591)	12.42 (0.0557)	11.92 (0.0543)	11.69 (0.0544)	8.67 (0.0427)	8.08 (0.0374)	7.67 (0.0373)	6.48 (0.0344)	6.09 (0.0329)
	0	22.94 (0.0910)	22.08 (0.0897)	22.22 (0.0896)	20.95 (0.0883)	20.58 (0.0884)	18.11 (0.0814)	16.03 (0.0730)	15.85 (0.0728)	14.54 (0.0685)	14.45 (0.0694)
Vehicle	**15**	6.59 (0.0289)	6.09 (0.0243)	5.19 (0.0241)	4.831 (0.0222)	5.01 (0.0221)	2.93 (0.0150)	2.93 (0.0134)	2.93 (0.0136)	2.71 (0.0126)	2.52 (0.0120)
	10	6.77 (0.0286)	5.82 (0.0247)	5.46 (0.0243)	4.92 (0.0222)	5.05 (0.0222)	2.98 (0.0155)	3.11 (0.0136)	3.02 (0.0140)	2.66 (0.0128)	2.43 (0.0124)
	5	6.86 (0.0288)	5.88 (0.0258)	5.51 (0.0250)	5.23 (0.0232)	5.14 (0.0228)	3.07 (0.0162)	3.11 (0.0145)	3.25 (0.0148)	2.75 (0.0132)	2.48 (0.0128)
	0	7.22 (0.0306)	6.14 (0.0266)	6.09 (0.0261)	5.55 (0.0241)	5.60 (0.0240)	3.29 (0.0175)	3.29 (0.0153)	3.43 (0.0154)	2.89 (0.0140)	2.61 (0.0134)

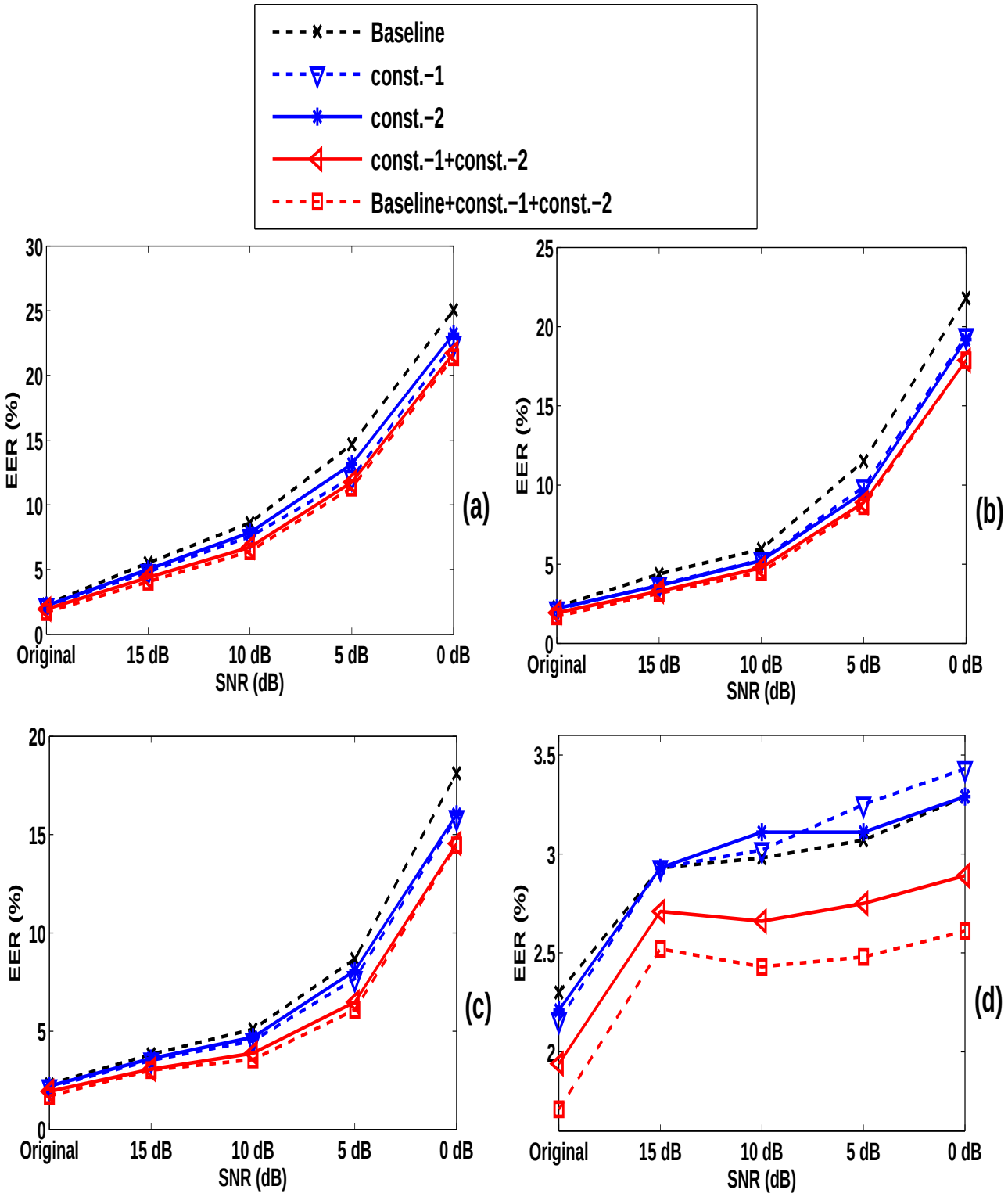

Figure 5.1: Summary of the SV performance (**in EER**) with session/ channel compensation for different experimental setup on NIST-2003 speaker recognition database. (a)-(d) Performance for different SNR level using: White, Factory1, Babble and Vehicle noise added test speech.

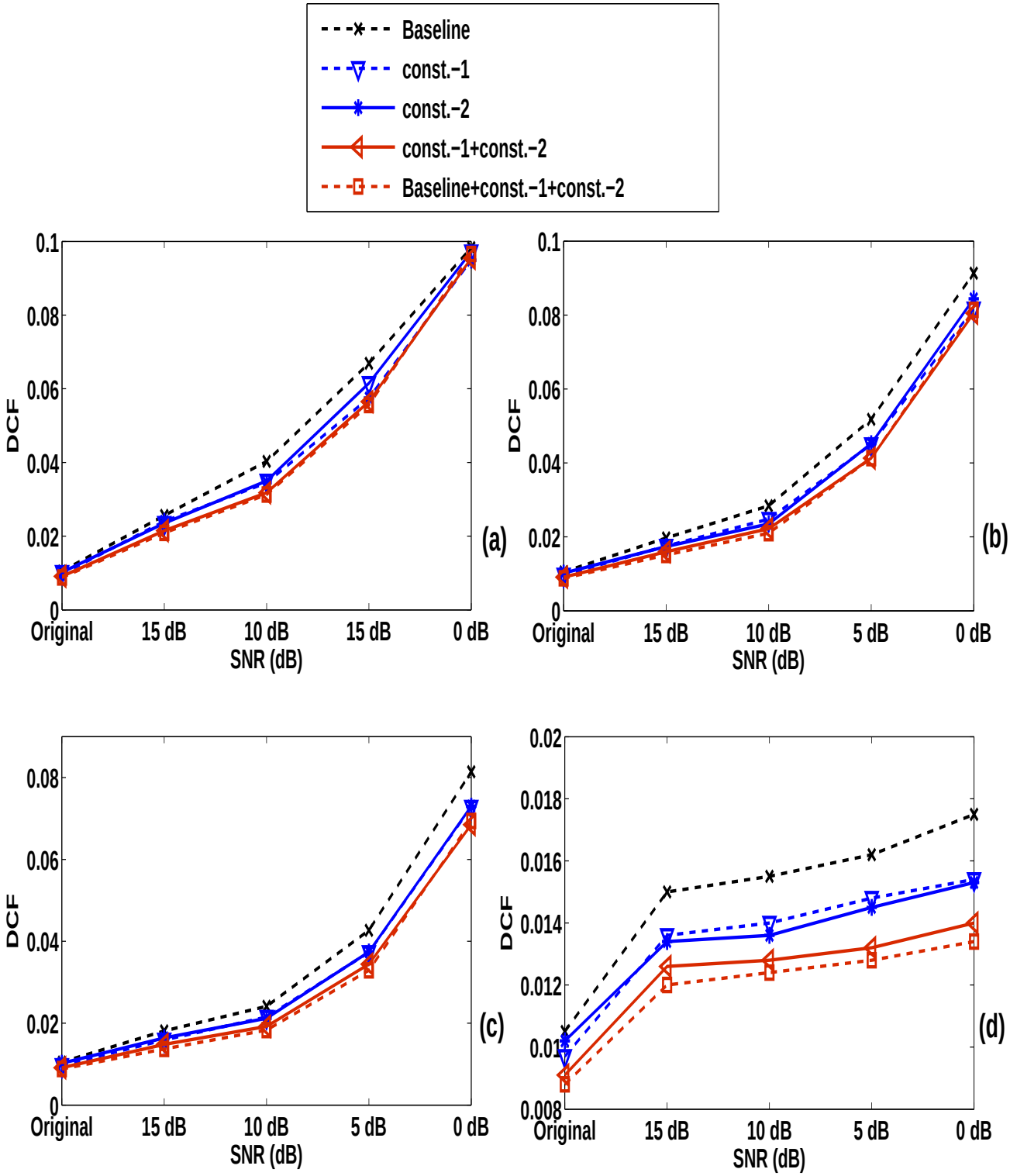

Figure 5.2: Summary of the SV performance (**in DCF**) with session/ channel compensation for different experimental setup on NIST-2003 speaker recognition database. (a)-(d) Performance for different SNR level using: White, Factory1, Babble and Vehicle noise added test speech.

represents the the variabilities present only in the VLRs. In the baseline system, the i-vectors are extracted without using VLRs and non-VLRs information. Therefore the i-vectors space of the baseline, constraint-1 and constraint-2 systems may model different speaker information. Motivated by this, these systems are combined. To observe performance at different levels of combination, first the scores obtained from the two constrained systems are linearly combined with equal weight to each system. Then, the score obtained from the baseline is linearly combined with the constrained systems. The performance of the combined systems are given in the Table 5.4. For a better comparison, the performance of different systems in terms of EER and DCF are summarized in Figs. 5.2 and 5.2, respectively.

The speaker verification results given in the Table 5.4 shows that for all noise added test conditions, constraint-1 and constraint-2 systems provide improved performance compared to the baseline system without and with session/ channel compensation. For instance, for 0 dB white noise added test speech EER is 25.06%, 23.21% and 22.44% for baseline, constraint-1 and constraint-2 systems, respectively. Similarly, DCF obtained for baseline, constraint-1 and constraint-2 system is 0.0983, 0.0972 and 0.0949, respectively. The combination of constraint-1 and constraint-2 systems provides improved performance compared to each constraint system. For all experimental conditions considered, best performance is observed for the combination of the baseline system with the constrained systems. For instance, for NIST 2003 speaker recognition database with LDA and WCCN, EER and DCF are improved from 2.34% to 1.71% and 0.0105 to 0.0088, respectively. Similarly for 0dB factory noise added test speech with LDA and WCCN, EER and DCF are improved from 21.81% to 17.88% and 0.0913 to 0.0814, respectively.

5.4.2 NIST 2012 speaker recognition database

The performance of different SV systems on the actual evaluation trials of the NIST SRE 2012 for the five common conditions of the core task are summarized in the Table 5.5. In NIST SRE 2012, the actual DCF is used as the primary metric for the performance evaluation. For this the system calibration and fusion is done based on the actual DCF of each sub system. For each test condition, the best DCF is obtained for the constraint-2 system. By comparing different systems, it can be observed that the combination of baseline and constraint systems provides better performance in terms of EER and DCF compared to the baseline system.

Table 5.5: Performance of the SV systems on NIST 2012 speaker recognition database for the common evaluation conditions of the core task. Performance is given in terms of EER and actual DCF

		Performance of SV system in terms of EER and Act. DCF									
		Baseline (S_1)		const.-1 (S_2)		const.-2 (S_3)		S_2&S_3 Comb.		S_1, S_2&S_3 Comb.	
Evaluation condition		EER	DCF	EER	DCF	EER	DCF	EER	DCF	EER	DCF
Phone call	No noise	3.52	0.60	4.65	0.57	4.40	0.55	3.97	0.55	3.22	0.56
	Added noise	4.15	0.72	5.50	0.71	5.32	0.67	4.86	0.68	3.95	0.68
	Noisy env.	4.04	0.64	5.45	0.61	5.07	0.59	4.58	0.60	3.63	0.60
Interview	No noise	7.35	0.56	8.11	0.55	7.31	0.52	7.17	0.53	6.66	0.53
	Added noise	5.89	0.54	6.28	0.64	6.18	0.53	5.76	0.57	5.53	0.53

5.5 Summary

This work proposed a VLRs and non-VLRs conditioned SV system by implicit modeling of VLRs and non-VLRs dimensions in the *i*-vectors. The implicit modeling is done by learning the total variability matrix in a supervised manner to represent variabilities present in VLRs and non-VLRs segments of the speech data. Two different methods are presented for learning VLRs and non-VLRs subspaces in the total variability matrix by considering the similarity and difference between them. The novelty of this work is unlike most of the conditioned SV systems, it does not require explicit segmentation of speech data during training and testing of the SV system. Only a total variability matrix is learned offline using the VLRs and non-VLRs segments of the development data.

The performance of the proposed SV system is evaluated on NIST 2003 and NIST 2012 speaker recognition databases. The experimental results presented in this work show that for both clean and degraded speech, the proposed SV system performed better than a SV system using the same regions without conditioning.

The SV experiment on NIST 2003 (Tables 4.7 and 5.4) shows that the implicit conditioning of VLRs and non-VLRs provides poorer performance compared to explicit conditioning of these regions. Alternatively the implicit conditioning makes the verification process fast by avoiding the explicit segmentation of speech data during training and testing of the SV system.

6

Summary and Conclusions

Contents

Objective

In this chapter, contributions of this thesis towards development of a speaker verification system under degraded conditions are summarized. Future research directions made possible by the present work are also outlined.

6.1 Summary

The objective of this thesis work is to develop a SV system by independent processing of VLRs and non-VLRs for achieving better SV performance under clean and degraded conditions. To achieve this, a method is proposed for VLRs and non-VLRs segmentation, and three different methods are proposed for processing VLRs and non-VLRs for a SV task. First, a SV system is developed by using only the VLRs to demonstrate the significance of the VLRs for SV under degraded conditions. Then, the VLRs and non-VLRs are used independently during training and testing of a SV system, and the scores are combined with higher weight on VLRs, for a better SV system under clean and degraded conditions. Finally, a SV system is developed by implicit modeling of VLRs and non-VLRs information to reduce the computational complexity involved in the explicit segmentation of these regions.

(i) **SV using VLRs:** For any practical application of a *text-independent* SV system, along with phonetic variability the speech signal may be affected by background noise, sensor mismatch and channel mismatch. In such a scenario, without knowledge of testing environment a better SV system can be developed by selecting only those speech regions, that are relatively more speaker discriminative and less affected by various degradations. This can be achieved using the knowledge of VLROP. VLROP helps in identifying VLRs, that are more speaker discriminative and less degradation affected speech regions. To demonstrate the significance of VLRs for SV under degraded conditions, a robust VLROPs method is required.

A VLROP detection method is proposed by utilizing the advantages of ZFFS and HE of LP residual. Two VLROP evidences are obtained by convolving excitation contour derived from the HE of LP residual and second order difference of ZFFS with a FOGD. Finally, the *VLROP evidence using the excitation source information* is obtained by adding the two evidences and normalizing by 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. Peak locations in the combined evidence are hypothesized VLROPs. The performance of proposed method is evaluated using a 60-speaker subset of the TIMIT database for clean as well as noise mixed speech signal. It is observed that the proposed *VLROP evidence using the excitation source information* is robust to noise degradation. The proposed VLROP detection algorithm performed better compared to the VLROPs detected by each individual feature and the vowel onset point detection method proposed in [79].

Using each hypothesized VLROP as the anchor point, 100 ms regions right to the VLROPs are marked as VLRs. SV systems are developed using VLRs and the speech regions detected by an energy based VAD (baseline). For both the systems, 39 dimensional MFCC (13 MFCC, 13 ΔMFCC and 13 $\Delta\Delta$MFCC) is used as the speaker features and GMM-UBM is used as the modeling technique. Degradation effect is compensated in the cepstral domain using CMVN and in the score level using T-norm. The performance of the SV systems are evaluated on NIST-2003 speaker recognition database. Then, two different noises are taken from NOISEX-92 database to create noise mixed NIST-2003 speech. Performance of the SV systems are evaluated for different degraded conditions on noise mixed NIST-2003 speaker recognition database. Finally, performance of the SV systems are evaluated on the IITG MV speaker recognition database for clean and real environmental degraded speech.

The experiments show that the performance of a SV system depends on the detection accuracy of the VAD and without knowledge of testing environment, the score normalization method like T-norm could not help much to remove the degradation effect on the verification scores. For clean speech with less number of VLRs frames, the proposed SV system gives comparable performance to the baseline system. Alternatively, under degraded conditions, the proposed system provides significantly improved performance. Therefore for severely corrupted speech signal, a better SV system can be developed by selecting only the VLRs.

(ii) **SV by explicit segmentation of VLRs and non-VLRs:** Detection VLRs by using a fixed length after VLROPs is not accurate. The other speech regions due to the non-VLRs account for a large number of frames and also contain good speaker information. For better

SV performance under clean and degraded conditions, a SV system is proposed by explicit segmentation of speech into VLRs and non-VLRs. To achieve this, methods are proposed for detecting complete VLRs and non-VLRs.

The VLREP event is defined and a method is proposed for detecting VLREPs using the excitation source information. The VLROPs and VLREPs are hypothesized and used in an iterative algorithm for detecting the VLRs. Next, for detection of non-VLRs, the LP residual samples in the VLRs are attenuated significantly to indirectly emphasize the residual samples in the non-VLRs. The modified LP residual samples excite the time varying all pole filter to reconstruct non-VLRs enhanced speech. The non-VLRs are separated from silence/noise frames by processing the reconstructed speech signal. The performance of proposed VLRs and non-VLRs methods are evaluated using a 60-speaker subset of the TIMIT database for clean as well as noise mixed speech signal. The experiments show that the detection accuracy of VLRs is improved significantly by using VLROPs and VLREPs compared to that detected by using only VLROPs. The proposed non-VLRs detection method performed better compared to the detection of non-VLRs by using VLRs and energy VAD.

The VLRs and non-VLRs are used independently during training and testing of a SV system. Finally, the scores are combined with higher weight on VLRs, which are more speaker specific and less degradation affected speech regions. GMM-UBM and *i*-vector based SV systems are developed using 39 dimensional MFCC. The feature vectors are normalized using CMVN. For the *i*-vector based SV system the session/channel variability compensation is performed using LDA and WCCN. The SV experiments are conducted on NIST-2003 speaker recognition database for original and noise mixed test speech, and on the IITG MV speaker recognition database.

It is observed that for both GMM-UBM and *i*-vector based SV systems, proposed approach consistently performed better than a SV system using the same regions without conditioning. The performance of the SV system using VLRs and non-VLRs depends on the detection accuracy of these regions.

(iii) **SV system by implicit modeling of VLRs and non-VLRs information:** To reduce the complexity involved in an explicit segmentation of speech data during training and testing, a

different approach is proposed to developed an i-vector based SV system by implicit modeling of VLRs and non-VLRs information.

In this approach, a $\boldsymbol{T}$ matrix is learned offline in a supervised manner to represents the total variability in VLRs by some particular columns and the total variability in non-VLRs by other columns. The speech data supervectors are projected to the modified $\boldsymbol{T}$ matrix to provide flexibility for the observed data to select the most relevant subspace dimensions. As a result, particular dimensions of an i-vector will mostly represent variabilities present in the VLRs and other dimensions mostly the variabilities present in the non-VLRs. Two different methods are presented for learning VLRs and non-VLRs subspaces in the $\boldsymbol{T}$ matrix by considering the similarity and difference between them. The main novelty of this work it does not require explicit segmentation of speech data during training and testing of a SV system.

The SV experiments are conducted on NIST-2003 speaker recognition database for original and noise mixed test speech, and on the NIST-2012 speaker recognition database. The experimental results show that for both clean and noise degraded speech, the proposed SV performed better than a conventional i-vector system, and the performance is further improved by combing the SV systems.

The SV experiment on NIST 2003 shows that the implicit conditioning of VLRs and non-VLRs provides poorer performance compared to explicit conditioning of these regions. Alternatively the implicit conditioning makes the verification process fast by avoiding the explicit segmentation of speech data during training and testing of the SV system.

6.2 Contributions

The major contributions of the research work reported in this thesis includes,

(i) Method for the detection of VLROPs and end VLREPs using excitation source information.

(ii) An iterative algorithm for the detection of complete VLRs using VLROP and VLREP.

(iii) Method for the detection of non-VLRs by emphasizing excitation information of non-VLRs in the LP residual.

(iv) Demonstrating significance of VLRs for SV under degraded conditions.

(v) SV system using VLRs and non-VLRs conditioning and emphasizing the scores from VLRs for better SV under clean and degraded conditions.

(vi) SV by implicit modeling of VLRs and non-VLRs to avoid the computational complexity involved in explicit segmentation of training and testing speech data.

6.3 Directions for future work

Based on the outcome of this thesis work, this section provides some of the possible future directions for research.

(i) To provide robustness, the excitation source features are used for the detection of VLROPs and VLREPs. The performance of detection algorithms may be improved by combining different spectral features with the excitation source features. Specifically, a thorough analysis is required for the detection of VLREP event.

(ii) For the detection of VLRs we have used VLROPs and VLREPs. The detection accuracy of VLRs may be improved by using statistical methods with the proposed signal processing methods.

(iii) The reconstructed non-VLRs enhanced speech provides better discrimination between non-VLRs and non-speech regions. We have used excitation source features for the detection of non-VLRs. At this point a better non-VLR detection method may be possible either by using other spectral feature along with the excitation source features or by constructing different classifiers for non-VLRs and non-speech regions using statistical models.

(iv) Unlike VLRs, the sound units in the non-VLRs vary among themselves. A better SV system may be developed by further classifying the non-VLRs, depending on the similarity in speaker specific information and signal characteristics. For example one such segmentation may be voiced and unvoiced. A thorough analysis is required for segmentation of these regions.

(v) The non-VLRs are relatively more affected under degraded conditions. The effect of environmental degradation on these regions may be reduced by applying a suitable speech enhancement method.

(vi) The MFCCs are used for the speaker specific features. The VLR and non-VLR segments are produced by different speech production mechanism, different features may be extracted from these regions. For example excitation source features may be extracted only from the VLRs for better representation of speaker specific excitation source. An investigation is required by considering different spectral and excitation source features.

(vii) Although the signal characteristics of VLRs and non-VLRs are different, they are produce by the same speech production system. For a better conditioned SV system, a method is required to find the similarity and difference in the speaker specific information present in these segments.

(viii) The implicit modeling proposed in this work need more investigation, specifically at the session/channel compensation stage to maintain conditioning.

Bibliography

[1] J. P. Campbell, "Speaker Recognition: A Tutorial," *Proc. IEEE*, vol. 85, no. 9, pp. 1437–1462, September 1997.

[2] F. Bimbot, J. Bonastreand, C. Fredouille, G. Gravier, I. Chagnolleau, S. Meignier, T. Merlin, J. Garcia, D. Delacretaz, and D. A. Reynolds, "A tutorial on text-independent speaker verification," *EURASIP Journal on Applied Signal Processing*, vol. 4, pp. 430–451, 2004.

[3] 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.

[4] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Trans. on speech and audio processing*, vol. 3, no. 1, pp. 72–83, January 1995.

[5] L. F. Lamel, L. R. Rabiner, A. E. Rosenberg, and J. G. Wilpon, "An improved endpoint detector for isolated word recognition," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. ASSP 29, pp. 777–785, August 1981.

[6] J. G. Wilpon, L. R. Rabiner, and T. Martin, "An improved word detection algorithm for telephone quality speech incorporating both syntactic and semantic constraints," *AT & T Bell Laboratories Technical Journal*, vol. 63, no. 3, pp. 479–498, July 1984.

[7] J. C. Junqua, B. Reaves, and B. Mak, "A study of endpoint detection algorithms in adverse conditions: Incidence on a DTW and HMM recognize," in *Proc. Eurospeech*, Genova, Italy, September 1991, pp. 1371–1374.

[8] J. Ramrez, J. C. Segura, C. Bentez, A. de la Torre, and A. Rubio, "Efficient voice activity detection algorithms using long-term speech information," *Speech Communication*, vol. 42, no. 3, pp. 271–287, 2004.

[9] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Processing Lett.*, vol. 6, no. 1, pp. 1–3, January 1999.

[10] Y. Cho and A. Kondoz, "Analysis and improvement of a statistical model based voice activity detector," *IEEE Signal Processing Lett.*, vol. 8, no. 10, pp. 276–279, October 2001.

[11] S. Gazor and W. Zhang, "A soft voice activity detector based on a Laplacian-Gaussian model," *IEEE Trans. Speech and Audio Processing*, vol. 11, no. 5, pp. 498–505, September 2003.

[12] J. H. Chang, N. S. Kim, and S. K. Mitra, "Voice activity detection based on multiple statistical models," *IEEE Trans. Signal Processing*, vol. 54, no. 6, pp. 1965–1976, June 2006.

[13] D. Ying, Y. Yan, J. Dang, and F. K. Soong, "Voice activity detection based on an unsupervised learning framework," *IEEE Trans. Audio, Speech and Lanuage Processing*, vol. 19, no. 8, pp. 2624–2633, November 2011.

[14] R. Saeidi, J. Pohjalainen, T. Kinnunen, and P. Alku, "Temporally weighted linear prediction features for tackling additive noise in speaker verification," *IEEE Signal Processing Lett.*, vol. 17, no. 6, pp. 599–602, June 2010.

[15] T. H. Falk and W. Chan, "Modulation spectral features for robust far-field speaker identification," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 18, no. 1, pp. 90–100, August 2010.

[16] N. Wang, P. C. Ching, N. Zheng, and T. Lee, "Robust speaker recognition using denoised vocal source and vocal tract features," *IEEE Trans. Audio, Speech and Lanuage Processing*, vol. 19, no. 1, pp. 196–205, January 2011.

[17] J. Wolf, "Efficient acoustic parameters for speaker recognition," *Journal of the Acoustical Society of America*, vol. 51, no. 6 (Part2), pp. 2044–2056, 1972.

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

[19] S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. ASSP-29, no. 2, pp. 254–272, April 1981.

[20] S. R. M. Prasanna, C. S. Gupta, and B. Yegnanarayana, "Extraction of speaker-specific excitation information from linear prediction residual of speech," *Speech Communication*, vol. 48, pp. 1243–1261, june 2006.

[21] K. S. R. Murty and B. Yegnanarayana, "Combining evidence from residual phase and mfcc features for speaker recognition," *IEEE Signal Processing Lett.*, vol. 13, no. 1, pp. 52–55, January 2006.

[22] B. Yegnanarayana, K. S. Reddy, and S. P. Kishore, "Source and systsem feature for speaker recognition using aann models," in *Proc. Int. Conf. Acoust. Speech and Signal Process.*, Salt Lake City, UT, USA, 2001, pp. 409–412.

[23] A. Adami, R. Mihaescu, D. Reynolds, and J. Godfrey, "Modeling prosodic dynamics for speaker recognition," in *Proc. Int. Conf. Acoust. Speech and Signal Process.*, Hong Kong, China, April 2003.

[24] D. Reynolds, W. Andrews, J. Campbell, J. Navratil, B. Peskin, A. Adomi, Q. Jin, D. Kluracek, J. Abramson, R. Mihaescu, J. Godfrey, D. Jones, and S. Xiang, "The SuperSID Project: Exploiting high-level information for high-accuracy speaker recognition," in *Proc. Int. Conf. Acoust. Speech and Signal Process.*, Hong Kong, China, 2003, pp. 784–787.

[25] Z. Chen, Y. F. Liao, and Y. T. Juang, "Eigen-prosody analysis for robust speaker recognition under mismatch handset environment," in *Proc. Int. Conf. on Spoken Language Process.*, Jeju, South Korea, October 2004, pp. 1421–1424.

[26] E. Shriberg, L. Ferrer, S. Kajarekar, A. Venkataraman, and A. Stocke, "Modeling prosodic feature sequences for speaker recognition," *Speech Communication*, vol. 46, no. 3-4, pp. 455–472, 2005.

[27] N. Dehak, P. Dumouchel, and P. Kenny, "Modeling prosodic features with joint factor analysis for speaker verification," *IEEE Trans. Audio, Speech and Lanuage Processing*, vol. 15, no. 7, pp. 2095–2103, September 2007.

[28] T. Kinnunen, B. Zhang, J. Zhu, and Y. Wang, "Speaker verification with adaptive spectral subband centroids," in *Proc. Internat. Conf. on Biometrics*, Seoul, Korea, August 2007, pp. 58–66.

[29] R. Mammone, X. Zhang, and R. Ramachandran, "Robust speaker recognition: a feature based approach," *IEEE Signal Processing Magazine*, vol. 13, no. 5, pp. 58–71, September 1996.

[30] H. Hermansky, "Perceptual linear predictive (plp) analysis of speech," *Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, April 1990.

[31] 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.

[32] L. Mary and B. Yegnanarayana, "Extraction and representation of prosodic features for language and speaker recognition," *Speech Communication*, vol. 50, no. 10, pp. 782 –796, April 2008.

[33] D. A. Reynolds, "Experimental evaluation of features for robust speaker identification," *IEEE Trans. Speech Audio Processing*, vol. 2, no. 4, pp. 639–643, October 1994.

[34] W. N. Chan, N. Zheng, and T. Lee, "Discrimination power of vocal source and vocal tract related features for speaker segmentations," *IEEE Trans. Audio, Speech and Signal Processing*, vol. 15, no. 6, pp. 1884–1892, August 2007.

[35] 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.

[36] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," *IEEE Trans. Audio Speech and Language Processing*, vol. 15, no. 4, pp. 1435–1447, May 2007.

[37] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Trans. Audio, Speech and Lanuage Processing*, vol. 19, no. 4, pp. 788–798, May 2011.

[38] S. Kim, T. Eriksson, H. G. Kang, and D. H. Youn, "A pitch synchronous feature extraction method for speaker recognition," in *Proc. Int. Conf. Acoust. Speech and Signal Process.*, vol. I, Quebec, Canada, May 2004, pp. 405–408.

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

[40] D. A. Reynolds, "Channel robust speaker verification via feature mapping," in *Proc. Int. Conf. Acoust. Speech and Signal Process.*, vol. 2, Hong Kong, April 2003, pp. II–53–56.

[41] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," in *Proc. Odyssey: The Speaker and Language Recognition Workshop*, Crete, Greece, June 2001, pp. 1–6.

[42] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, "A study of inter-speaker variability in speaker verification," *IEEE Trans. Audio, Speech and Lanuage Processing*, vol. 16, no. 5, pp. 980–988, July 2008.

[43] 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 and Audio Processing*, vol. 11, no. 5, pp. 435–446, September 2003.

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

[45] B. Xiang, U. V. Chaudhari, J. Navratil, G. N. Ramaswamy, and R. A. Gopinath, "Short-time Gaussianization for robust speaker verification," in *Proc. Int. Conf. Acoust. Speech and Signal Process.*, vol. 1, Orlando, Florida, USA, May 2002, pp. 681–684.

[46] A. Stolcke, S. Kajarekar, and L. Ferrer, "Nonparametric feature normalization for SVM-based speaker verification," in *Proc. Int. Conf. Acoust. Speech and Signal Process.*, Las Vegas, Nevada, April 2008, pp. 1577–1580.

[47] V. Wan and S. Renals, "Speaker verification using sequence discriminant support vector machines," *IEEE Trans. Speech and Audio Processing*, vol. 13, no. 2, pp. 203–210, March 2005.

[48] A. Solomonoff, W. M. Campbell, and I. Boardman, "Advances in channel compensation for SVM speaker recognition," in *Proc. Int. Conf. Acoust. Speech and Signal Process.*, vol. 1, Philadelphia, USA, March 2005, pp. 629–632.

[49] N. Brummer, L. Burget, J. H. Cernocky, O. Glembek, F. Grezl, M. Karafiat, D. A. van Leeuwen, P. Matejka, P. Schwarz, and A. Strasheim, "Fusion of heterogeneous speaker recognition systems in the STBU submission for the NIST speaker recognition evaluation 2006," *IEEE Trans. Audio, Speech and Lanuage Processing*, vol. 15, no. 7, pp. 2072–2084, September 2007.

[50] A. Hatch, S. Kajarekar, and A. Stolcke, "Within-class covariance normalization for SVM based speaker recognition," in *Proc. Int. Conf. on Spoken Language Process.*, Pittsburgh, PA, USA, September 2006, pp. 1471–1474.

[51] A. Stolcke, S. Kajarekar, L. Ferrer, and E. Shriberg, "Speaker recognition with session variability normalization based on MLLR adaptation transforms," *IEEE Trans. Audio, Speech and Lanuage Processing*, vol. 15, no. 7, pp. 1987–1998, September 2007.

[52] Q. Jin and A. Waibel, "Application of LDA to speaker recognition," in *Proc. Int. Conf. on Spoken Language Process.*, Beijing, China, October 2000.

[53] D. Burton, "Text-dependent speaker verification using vector quantization source coding," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 35, no. 2, pp. 133–143, February 1987.

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

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

[56] K. Farrell, R. Mammone, and K. Assaleh, "Speaker recognition using neural networks and conventional classifiers," *IEEE Trans. Speech Audio Processing*, vol. 2, no. 1, pp. 194–205, January 1994.

[57] B. Yegnanarayana and S. P. Kishore, "AANN: an alternative to GMM for pattern recognition," *Neural Networks*, vol. 15, pp. 459–469, January 2002.

[58] W. M. Campbell, J. P. Campbell, D. A. Reynolds, E. Singer, and P. A. Torres-Carrasquillo, "Support vector machines for speaker and language recognition," *Computer Speech and Language*, vol. 20, no. 2-3, pp. 210–229, 2006.

[59] D. Ramos-Castro, J. Fierrez-Aguilar, J. Gonzalez-Rodriguez, and J. Ortega-Garcia, "Speaker verification using speaker- and test-dependent fast score normalization," in *Pattern Recognition Lett.*, vol. 28, no. 1, 2007, pp. 90–98.

[60] 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.

[61] D. A. Reynolds, "The effect of handset variability on speaker recognition performance: experiments on the Switchboard corpus," in *Proc. Int. Conf. Acoust. Speech and Signal Process.*, vol. 1, Atlanta, Ga, May 1996, pp. 113–116.

[62] R. Dunn, T. Quatieri, D. A. Reynolds, and J. Campbell, "Speaker recognition from coded speech and the effects of score normalization," in *35th Asilomar Conf. on Signals, Systems and Computers*, vol. 2, Pacific Grove, California, USA, November 2001, pp. 1562–1567.

[63] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The DET curve in asssessment of detection task performance," in *Proc. Eur. Conf. Speech Communication Technology*, Rhodes, Greece, 1997, pp. 1895–1898.

[64] D. A. van Leeuwen, A. F. Martin, M. A. Przybocki, and J. S. Bouten, "NIST and NFI-TNO evaluations of automatic speaker recognition," *Computer Speech and Language*, vol. 20, no. 1, pp. 128–158, January 2006.

[65] "The NIST Year 2003 Speaker Recognition Evaluation Plan," in *http://www.itl.nist.gov/iad/mig/tests /spk/2003/2003-spkrec-evalplan-v2.2.pdf*.

[66] S. Kajarekar, "Phone-based cepstral polynomial SVM system for speaker recognition," in *Proc. INTERSPEECH*, 2008.

[67] F. Castaldo, D. Colibro, E. Dalmasso, P. Laface, and C. Vair, "Compensation of nuisance factors for speaker and language recognition," *IEEE Trans. Audio, Speech and Lanuage Processing*, vol. 15, no. 15, pp. 1969–1978, Sept. 2007.

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

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

[70] T. Bocklet and E. Shriberg, "Speaker recognition using syllable based constraints for cepstral frame selection," in *Proc. Int. Conf. Acoust. Speech and Signal Process.*, 2009, pp. 4525–4528.

[71] M. Tatham and K. Morton, *A Guide to Speech Production and Perception.* Edinburgh University Press, Edinburgh, 2011.

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

[73] TIMIT, *"Timit Acoustic-Phonetic Continuous Speech Corpus"*, NIST Order PB91-505065, National Institute of Standards and Technology, Gaithersburg, MD, USA, 1990, Speech Disc 1-1.1., 1990.

[74] F. H. Liu, R. M.Stern, A. Acero, and P. J. Moreno, "Environment normalization for robust speech recognition using direct cepstral comparison," in *Proc. Int. Conf. Acoust. Speech and Signal Process.*, vol. II, Australia, April 1994, pp. 61–64.

[75] D. J. Hermes, "Vowel onset detection," *Journal of the Acoustical Society of America*, vol. 87, pp. 866–873, 1990.

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

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

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

[79] 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.

[80] W. Campbell, D. Sturim, and D. Reynolds, "SVM based speaker verification using a GMM supervector kernel and NAP variability compensation," in *Proc. Int. Conf. Acoust. Speech and Signal Process.*, Philadelphia, USA, March 2005, pp. 637–640.

[81] D. E. Sturim, D. A. Reynolds, R. B. Dunn, and T. F. Quatieri, "Speaker verification using text-constrained Gaussian mixture models," in *Proc. Int. Conf. Acoust. Speech and Signal Process.*, 2002, pp. 677–680.

[82] J. Ortega-Garca and J. Gonzlez-Rodrguez, "Overview of speech enhancement techniques for automatic speaker recognition," in *Proc. Int. Conf. Spoken Language*, Philadelphia, PA, October 1996, pp. 929–932.

[83] A. Drygajlo and M. El-Maliki, "Speaker verification in noisy environment with combined spectral subtraction and missing data theory," in *Proc. Int. Conf. Acoust. Speech and Signal Process.*, vol. 1, Seattle, Washington, USA, May 1998, pp. 121–124.

[84] P. Krishnamoorthy and S. R. M. prasanna, "Application of combined temporal and spectral processing methods for speaker recognition under noisy, reverberant or multi-speaker environments," *Sadhana*, vol. 34, no. 5, pp. 729–754, October 2009.

[85] Q. Li, J. Zheng, A. Tsai, and Q. Zhou, "Robust endpoint detection and energy normalization for real-time speech and speaker recognition," *IEEE Trans. Speech and Audio Process.*, vol. 10, no. 3, pp. 146–157, March 2002.

[86] H. Yu and M. W. Mak, "Comparison of voice activity detectors for interview speech in NIST speaker recognition evaluation," in *Proc. Interspeech*, Florence, Italy, Aug. 2011, pp. 2353–2356.

[87] V. Hautamaki, M. Tuononen, T. Niemi-Laitinen, and P. Franti, "Improving speaker verification by periodicity based voice activity detection," in *Proc. 12th Int. Conf. Speech and Computer*, vol. 2, Moscow, October 2007, pp. 645–650.

[88] R. Chengalvarayan, "Robust energy normalization using speech/non-speech discriminator for german connected digit recognition," in *Proc. Eurospeech*, Budapest, Hungary, Sept. 1999, pp. 61–64.

[89] E. Nemer, R. Goubran, and S. Mahmoud, "Robust voice activity detection using higher-order statistics in the lpc residual domain," *IEEE Trans. Speech and Audio Process.*, vol. 9, no. 3, pp. 217–231, March 2001.

[90] J. W. Shin, J. H. Chang, H. S. Yun, and N. S. Kim, "Voice activity detection based on generalized gamma distribution," in *Proc. Int. Conf. Acoust. Speech and Signal Process.*, vol. 1, Philadelphia, USA, March 2005, pp. 781–784.

[91] B. Fauve, N. Evans, and J. Mason, "Improving the performance of text-independent short-duration SVM- and GMM- based speaker verification," in *Proc. Odyssey: the Speaker and Language Recognition Workshop*, Stellenbosch, South Africa, Jan. 2008.

[92] F. Beritelli and A. Spadaccini, "The role of voice activity detection in forensic speaker verification," in *Pro. Int. Conf. on Digital Signal Process.*, Corfu, Greece, July 2011, pp. 1–6.

[93] M. Sahidullah and G. Saha, "Comparison of speech activity detection techniques for speaker recognition," *Cornell University Library (arXiv:1210.0297)*, pp. 1–7, October 2012.

[94] L. Besacier, J. F. Bonastre, and C. Fredouille, "Localization and selection of speaker-specific information with statistical modelling," *Speech Communication*, vol. 31, no. 2-3, p. 89106, June 2000.

[95] J. P. Eatock and J. S. Mason, "A quantitative assesment of the relative speaker discriminating properties of phonemes," in *Proc. Int. Conf. Acoust. Speech and Signal Process.*, vol. 1, Adelaide, Australia, April 1994, pp. 133–136.

[96] A. Park and T. J. Hazen, "Asr dependent techniques for speaker identification," in *Proc. Int. Conf. on Spoken Language Process.*, Denver, Colorado, USA, Sept. 2002, pp. 1337–134.

[97] C. S. Jung, M. Y. Kim, and H. G. Kang, "Selecting feature frames for automatic speaker recognition using mutual information," *IEEE Trans. Audio, Speech and Lanuage Processing*, vol. 18, no. 6, pp. 1332–1340, August 2010.

[98] D. Baum, D. Schneider, T. Mertens, and J. Kohler, "Constrained subword units for speaker recognition," in *Proc. Odyssey: The Speaker and Language Recognition Workshop*, Brno, Czech Republic, June 2010, pp. 1–6.

[99] E. Shriberg and A. Stolcke, "Language-independent constrained cepstral features for speaker recognition," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, Prague, Czech, May 2011, pp. 5296–5299.

[100] Aronowitz, D. Burshtein, and A. Amir, "Text independent speaker recognition using speaker dependent word spotting," in *Proc. Interspeech*, Jeju Island, Korea, Oct. 2004, pp. 1789–1792.

[101] B. Baker, R. Vogt, and S. Sridharan, "Gaussian mixture modelling of broad phonetic and syllabic events for text-independent speaker verification," in *INTERSPEECH*, Lisbon, Portugal, Sept. 2005, pp. 2429–2432.

[102] H. Lei and N. Mirghafori, "Comparisons of recent speaker recognition approaches based on word-conditioning," in *Proc. Odyssey: The Speaker and Language Recognition Workshop*, Stellenbosch, South Africa, Jan. 2008.

[103] ——, "Word-conditioned HMM supervectors for speaker recognition," in *Interspeech*, Antwerp, Belgium, August 2007, pp. 746–749.

[104] S.Nakagawa, W.Zhang, and M.Takahashi, "Text-independent speaker recognition by combining speaker specific gmm with speaker adapted syllable-based hmm," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, vol. 1, Montreal, Quebec, Canada, May 2004, pp. 81–84.

[105] R. Faltlhauser and G. Ruske, "Improving speaker recognition performance using phonetically structured Gaussian mixture models," in *Eurospeech*, Aalborg, Denmark, Sept. 2001, pp. 751–754.

[106] E. G. Hansen, R. E. Slyh, and T. R. Anderson, "Speaker recognition using phoneme-specific gmms," in *Proc. Odyssey: The Speaker and Language Recognition Workshop*, Toledo, Spain, May 2004.

[107] U. V. Chaudhari, J. Navratil, and S. H. Maes, "Multigrained modeling with pattern specific maximum likelihood transformations for text-independent speaker recognition," *IEEE Trans. Speech Audio and Process.*, vol. 11, no. 1, pp. 61–69, Jan. 2003.

[108] M. Hebert and L. Heck, "Phonetic class-based speaker verification," in *Eurospeech*, Geneva, Switzerland, Sept. 2003, pp. 1665–1668.

[109] K. J. Han, J. Pelecanos, and M. K. Omar, "Keyword-conditioned phone n-gram modeling with contextual information for speaker verification," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, Kyoto, Japan, March 2012, pp. 4797–4800.

[110] N. Scheffer, R. Vogt, S. Kajarekar, and J. Pelecanos, "Combination strategies for a factor analysis phone-conditioned speaker verification system," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, Taipei, Taiwan, April 2009, pp. 4053–4056.

[111] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. ASSP-27, no. 2, pp. 113–120, April 1979.

[112] 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.

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

[114] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, vol. 4, Washington, USA, April 1979, pp. 208–211.

[115] O. Cappe, "Elimination of the musical noise phenomenon with the ephraim and malah noise suppressor," *IEEE Trans. Speech Audio and Process.*, vol. 2, no. 2, pp. 345–349, April 1994.

[116] S. Kamath and P. Loizou, "A multi-band spectral subtraction method for enhancing speech corrupted by colored noise," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, vol. 4, Orlando, USA, May 2002.

[117] H. Gustafsson, S. Nordholm, and I. Claesson, "Spectral subtraction using reduced delay convolution and adaptive averaging," *IEEE Trans. Speech and Audio Process.*, vol. 9, pp. 799–807, Nov 2001.

[118] P. Renevey and A. Drygajlo, "Statistical estimation of unreliable features for robust speech recognition," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, vol. 3, Istanbul, Turkey, 2000, pp. 1731–1734.

[119] K. Yamashita and T. Shimamura, "Nonstationary noise estimation using low-frequency regions for spectral subtraction," *IEEE Signal Process. Letters*, vol. 2, pp. 465–468, June 2005.

[120] D. L. Donoho, "De-noising by soft-thresholding," *IEEE Trans. Information Theory*, vol. 41, no. 3, pp. 613–627, May 1995.

[121] M. Bahoura and J. Rouat, "Wavelet speech enhancement based on the teager energy operator," *IEEE Signal Process. Letters*, vol. 8, no. 1, pp. 10–12, Jan. 2001.

[122] M. T. Johnson, X. Yuan, and Y. Ren, "Speech signal enhancement through adaptive wavelet thresholding," *Speech Communication*, vol. 49, pp. 123–133, Feb. 2007.

[123] Y. Ephraim, "A Bayesian estimation approach for speech enhancement using hidden Markov models," *IEEE Trans. Signal Process.*, vol. 40, no. 4, pp. 725–735, April 1992.

[124] R. Martin, "Speech enhancement based on minimum mean-square error estimation and supergaussian priors," *IEEE Trans. Speech and Audio Process.*, vol. 13, no. 5, pp. 845–856, Sept. 2005.

[125] B. Chen and P. Loizou, "Speech enhancement using a MMSE short time spectral amplitude estimator with Laplacian speech modeling," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, vol. 1, Philadelphia, PA, USA, March 2005, pp. 1097–1100.

[126] D. Y. Zhao, W. B. Kleijn, A. Ypma, and B. de Vries, "Online noise estimation using stochastic-gain HMM for speech enhancement," *IEEE Trans. Audio, Speech and Lanuage Process.*, vol. 16, no. 4, pp. 835–846, May 2008.

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

[128] P. Lockwood and J. Boudy, "Experiments with a nonlinear spectral subtractor (NSS), Hidden Markov Models and the projection, for robust speech recognition in cars," *Speech Communication*, vol. 11, no. 2-3, pp. 215–228, June 1992.

[129] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech and Signal Process.*, vol. ASSP-32, no. 6, pp. 1109–1121, Dec. 1984.

[130] ——, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech and Signal Process.*, vol. ASSP-33, no. 2, pp. 443–445, April 1985.

[131] T. May, S. van de Par, and A. Kohlrausch, "Noise-robust speaker recognition combining missing data techniques and universal background modeling," *IEEE Trans. Audio, Speech and Lanuage Process.*, vol. 20, no. 1, pp. 108–121, January 2012.

[132] A. Panda, N. Tripathi, and T. Srikanthan, "Improved spectral subtraction technique for text-independent speaker varification," in *Proc. Int. Conf. on Digital Signal Processing*, Cardiff, Wales, UK, July 2007.

[133] 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. on Communication (NCC)*, Bangalore, India, January 2011.

[134] L. Burget, P. Matejka, P. Schwarz, O. Glembek, and J. Cernocky, "Analysis of feature extraction and channel compensation in a GMM speaker recognition system," *IEEE Trans. Audio, Speech and Lanuage Processing*, vol. 15, no. 7, pp. 1979–1986, Sept. 2007.

[135] C. Barras and J. L. Gauvain, "Feature and score normalization for speaker verification of cellular data," in *Proc. Int. Conf. Acoust. Speech and Signal Process.*, Hong Kong, China, 2003, pp. 49–53.

[136] M. Alam, P. Ouellet, P. Kenny, and D. O'Shaughnessy, "Comparative evaluation of feature normalization techniques for speaker verification," in *Proc. Int. Conf. Advances in nonlinear speech process.*, Berlin, Heidelberg, 2011, pp. 246–253.

[137] J. Han and W. Gao, "Robust telephone speech recognition based on channel compensation," *Pattern Recognition*, vol. 32, no. 6, pp. 1061–1067, June 1999.

[138] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. Speech and Audio Processing*, vol. 2, no. 4, pp. 578–589, October 1994.

[139] M. Mason, B. B. R. Vogt, and S. Sridharan, "Data-driven clustering for blind feature mapping in speaker verification," in *INTERSPEECH*, Lisbon, Portugal, Sept 2005, pp. 3109–3112.

[140] D. ZHU, B. MA, H. LI, and Q. HUO, "A generalized feature transformation approach for channel robust speaker verification," in *Proc. Int. Conf. Acoust. Speech and Signal Process.*, vol. 4, Honolulu, Hawaii, USA, April 2007, pp. 19–41.

[141] E. G. Hansen, R. E. Slyh, and T. R. Anderson, "Supervised and unsupervised speaker adaptation in the nist 2005 speaker recognition evaluation," in *Proc. Odyssey the Speaker and Language Recognition Workshop*, San Juan, Puerto Rico, 2006.

[142] W. M. Campbell, D. E. Sturim, W. Shen, D. A. Reynolds, and J. Navratil, "The MIT-LL/IBM 2006 speaker recognition system: High-performance reduced-complexity recognition," in *Proc. Int. Conf. Acoust. Speech and Signal Process.*, vol. 4, Honolulu, Hawaii, USA, April 2007, pp. 217–220.

[143] M. J. F. Gales, "Semi-tied covariance matrices for hidden Markov models," *IEEE Trans. Speech and Audio Process.*, vol. 7, no. 3, pp. 272–281, May 1999.

[144] W. Campbell, D. Sturim, D. Reynolds, and A. Solomonoff, "SVM based speaker verification using a GMM supervector kernel and NAP variability compensation," in *Int. Conf. Acoust., Speech, Signal Process.*, vol. 1, Toulouse, France, May 2006, pp. 97–100.

[145] W. M. Campbell, D. E. Sturim, and D. A. Reynolds, "Support vector machines using GMM supervectors for speaker verification," *IEEE Signal Processing Lett.*, vol. 13, no. 5, pp. 308–311, May 2006.

[146] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Speaker and session variability in GMM-based speaker verification," *IEEE Trans. Audio, Speech and Lang. Process.*, vol. 15, no. 4, pp. 1448–1460, May 2007.

[147] O. Glembek, L. Burget, N. Brummer, and P. Kenny, "Comparison of scoring methods used in speaker recognition with joint factor analysis," in *Proc. Int. Conf. Acoust. Speech and Signal Process.*, Taipei, Taiwan, April 2009, pp. 4057–4060.

[148] N. Dehak, R. Dehak, P. Kenny, and P. Dumouchel, "Comparison between factor analysis and GMM support vector machines for speaker verification," in *Proc. Odyssey: The Speaker and Language Recognition Workshop*, Stellenbosch, South Africa, 2008, p. January.

[149] R. Vogt, S. Kajarekar, and S. Sridharan, "Discriminant NAP for SVM speaker recognition," in *Proc. Odyssey: The Speaker and Language Recognition Workshop*, Stellenbosch, South Africa, January 2008.

[150] N. Dehak, R. Dehak, P. Kenny, N. Brummer, P. Ouellet, , and P. Dumouchel, "Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification," in *Proc. INTERSPEECH*, Brigthon, U.K, Sept. 2009, pp. 1559–1562.

[151] B. C. Haris and R. Sinha, "Sparse representation over learned and discriminatively learned dictionaries for speaker verification," in *Proc. Int. Conf. Acoust. Speech and Signal Process.*, Kyoto, Japan, March 2012.

[152] A. Hatch and A. Stolcke, "Generalized linear kernels for one-versus-all classification: application to speaker recognition," in *Proc. Int. Conf. Acoust. Speech and Signal Process.*, Toulouse, Franc, May 2006, pp. 585–588.

[153] M. J. F. Gales and S. Young, "Hmm recognition in noise using parallel model combination," in *Proc. Eurospeech*, Berlin, Germany, 1993, pp. 837–840.

[154] M. J. F. Gales and S. J. Young, "Robust continuous speech recognition using parallel model combination," *IEEE Trans. Speech and Audio Process.*, vol. 4, no. 5, pp. 352–359, Sept. 1996.

[155] W. Wu, T. F. Zheng, M.-X. Xu, and F. K. Soong, "A cohort-based speaker model synthesis for mismatched channels in speaker verification," *IEEE Trans. Audio, Speech and Lanuage Process.*, vol. 15, no. 6, pp. 1893–1903, Aug. 2007.

[156] S. Sagayama, Y. Yamaguchi, S. Takahashi, and J. Takahashi, "Jacobian approach to fast acoustic model adaptation," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, Munich, Germany, April 1997, pp. 835–838.

[157] C. Cerisara, L. Rigaziob, and J. C. Junqua, "A-Jacobian environmental adaptation," *Speech Commun.*, vol. 42, no. 1, pp. 25–41, Jan. 2004.

[158] L. Zao and R. Coelho, "Colored noise based multicondition training technique for robust speaker identification," *IEEE Signal Processing Lett.*, vol. 18, no. 11, pp. 675–678, Nov. 2011.

[159] L. P. Heck and M. Weintraub, "Handset-dependent background models for robust text-independent speaker recognition," in *Proc. Int. Conf. Acoust. Speech and Signal Process.*, vol. 2, Munich, Germany, April 1997, pp. 1071–1074.

[160] K. P. Li and J. E. Porter, "Normalizations and selection of speech segments for speaker recognition scoring," in *Proc. Int. Conf. Acoust. Speech and Signal Process.*, vol. 1, New York, NY, USA, April 1988, pp. 595–598.

[161] A. Higgins, L. Bahler, and J. Porter, "Speaker verification using randomized phrase prompting," *Digital Signal Processing*, vol. 1, no. 2, pp. 89–106, 1991.

[162] T. Matsui and S. Furui, "Concatenated phoneme models for text-variable speaker recognition," in *Proc. Int. Conf. Acoust. Speech and Signal Process.*, vol. 1, Minneapolis, Minn, USA, April 1993, pp. 391–394.

[163] D. A. Reynolds, "Comparison of background normalization methods for text-independent speaker verification," in *Proc. European Conf. on Speech Comm. and Tech.*, vol. 2, Rhodes, Greece, Sept. 1997, pp. 963–966.

[164] M. Carey, E. Parris, and J. Bridle, "A speaker verification system using alpha-nets," in *Proc. Int. Conf. Acoust. Speech and Signal Process.*, vol. 1, Toronto, Canada, May 1991, pp. 397–400.

[165] G. Gravier and G. Chollet, "Comparison of normalization techniques for speaker recognition," in *Workshop on Speaker Recognition and its Commercial and Forensic Applications*, Avignon, France, April 1998, pp. 97–100.

[166] C. Fredouille, J. Bonastre, and T. Merlin, "Similarity normalization method based on world model and a posteriori probability for speaker verification," in *Proc. European Conf. on Speech Comm. and Tech.*, Budpest, Hungary, Sept. 1999, pp. 983–986.

[167] R. Vogt, B. Baker, and S. Sridharan, "Modeling session variability in text-independent speaker verification," in *Proc. INTERSPEECH*, Lisbon, Portugal, Sept. 2005, pp. 3117–3120.

[168] D. E. Sturim and D. A. Reynolds, "Speaker adaptive cohort selection for tnorm in text-independent speaker verification," in *Proc. Int. Conf. Acoust. Speech and Signal Process.*, vol. 1, Philadelphia, USA, March 2005, pp. 741–744.

[169] S. Yin, R. Rose, and P. Kenny, "Adaptive score normalization for progressive model adaptation in text independent speaker verification," in *Proc. Int. Conf. Acoust. Speech and Signal Process.*, Las Vegas, Nevada, April 2008, pp. 4857–4860.

[170] J. Navratil and G. N. Ramaswamy, "The awe and mystery a of T-norm," in *Proc. Eurospeech*, Geneva, Switzerland, September 2003, pp. 2009–2012.

[171] M. Ben, R. Blouet, and F. Bimbot, "A monte-carlo method for score normalization in automatic speaker verification using Kullback-Leibler distances," in *Proc. Int. Conf. Acoust. Speech and Signal Process.*, vol. 1, Orlando, Fla, USA, May 2002, pp. 689–692.

[172] S.-C. Yin, R. Rose, and P. Kenny, "A joint factor analysis approach to progressive model adaptation in text-independent speaker verification," *IEEE Trans. Audio, Speech and Lanuage Processing*, vol. 15, no. 7, pp. 1999–2010, Sept. 2007.

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

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

[175] B. Yegnanarayana and P. S. Murthy, "Enhancement of reverberant speech using LP residual signal," *IEEE Trans. Speech and Audio Processing*, vol. 8, no. 3, pp. 267–281, May 2000.

[176] A. N. Khan and B.Yegnanarayana, "Vowel onset point based variable frame rate analysis for speech recognition," in *Pro. Int. Conf. Intelligent Sensing and Information Process.*, January 2005, pp. 392–394.

[177] S. R. M. Prasanna and B. Yegnanarayana, "Extraction of pitch in adverse conditions," in *Proc. Int. Conf. Acoust. Speech and Signal Process.*, vol. 1, May 2004, pp. 109–112.

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

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

[180] 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 Lanuage Processing*, vol. 15, no. 1, pp. 34–43, January 2007.

[181] 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.

[182] NIST, *"NIST-Speaker Recognition Evaluations." in [On-line]*, Available: http://www.nist.gov/speech/tests/spk., 2003.

[183] 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 and Signal Process.*, vol. 4, Honolulu, HI, April 2007, pp. 281–284.

[184] "Linguistic data consortium, "Switchboard cellular part 2 audio"," in http: // www.ldc.upenn.edu/ Catalog/ CatalogEntry.jsp catalogId = LDC 2004S07, 2004.

[185] J. Navratil, G. N. Ramaswamy, and R. D. Zilca, "Statistical model migration in speaker recognition," in *Proc. INTERSPEECH*, Jeju Island, Korea, October 2004.

[186] M. K. Omar, J. Navratil, and G. Ramsawamy, "Maximum conditional mutual information modeling for speaker verification," in *Proc. INTERSPEECH*, Lisbon, Portugal, September 2005, pp. 2169–2172.

[187] N. Fakotakis and J. Sirigos, "A high performance text independent speaker recognition system based on vowel spotting and neural nets," in *Proc. Int. Conf. Acoust. Speech and Signal Process.*, vol. 2, Atlanta, Georgia, USA, May 1996, pp. 661–664.

[188] B. C. Haris, G. Pradhan, A. Misra, S. R. M. Prasanna, R. K. Das, and R. Sinha, "Multivariability speaker recognition database in Indian scenario," *Int. Journal of Speech Technology (Springer)*, vol. 15, no. 4, pp. 441–453, March 2012.

[189] M. Senoussaoui, P. Kenny, N. Dehak, and P. Dumouchel, "An i-vector extractor suitable for speaker recognition with both microphone and telephone speech," in *Odyssey: The Speaker and Language Recognition Workshop*, 2010.

[190] *The NIST Year 2012 Speaker Recognition Evaluation Plan, www.nist.gov/itl/iad/mig/upload/NIST SRE12 evalplan-v17-r1.pdf.*

[191] *The BOSARIS Toolkit, www.https://sites.google.com/site/bosaristoolkit/.*

List of Publications

Journal Publications

- Published Papers:
  1. G. Pradhan and S. R. M. Prasanna, "Speaker verification by vowel and nonvowel like segmentation", **IEEE Trans. Audio, Speech, and Language Process.**, vol. 21, no. 4, pp. 854-867, April 2013.
  2. G. Pradhan, Haris. B.C, S. R. M. Prasanna and R. Sinha, "Speaker verification in sensor and acoustic environment mismatch conditions", **International Journal of Speech Technology (Springer)**, vol. 15, pp. 381-392, June 2012.
  3. Haris. B. C, G. Pradhan, A. Misra, S. R. M. Prasanna, R. K. Das and R. Sinha, "Multi-variability speaker recognition database in Indian scenario," **International Journal of Speech Technology (Springer)**, vol. 15, pp. 441-453, June 2012.
  4. G. Pradhan and S. R. M. Prasanna, "Speaker verification under degraded condition: a perceptual study", **International Journal of Speech Technology (Springer)**, vol. 14, no.4, pp. 405-417, Oct. 2011.
  5. S. R. M. Prasanna and G. Pradhan, "Significance of Vowel-Like Regions for Speaker Verification under Degraded Condition", **IEEE Trans. Audio, Speech, and Language Process.**, vol. 19, no. 8, pp. 2552-2565, May 2011.
  6. G. Pradhan and S. R. M. Prasanna, "Significance of Vowel Onset Point Information for Speaker Verification", **International Journal of computer & Communication Technology** , vol.2, pp. 56-61, Feb. 2011.
- Manuscripts to be Communicated
  1. G. Pradhan, Haris B. C, S. R. M. Prasanna and R. Sinha, "Speaker verification by implicit modeling of vowel-like and non-vowel-like information".
  2. G. Pradhan and S. R. M. Prasanna, "Speaker verification under degraded conditions: A review"

Conference and Workshop Publications

1. Haris B. C, G. Pradhan, R. Sinha and S. R. M. Prasanna, "The IITG Speaker Verification System for NIST SRE 2012," *in IEEE Int. Conf. Acoust., Speech, Signal process.,* Vancouver, Canada, May, 2013
2. G. Pradhan and S. R. M. Prasanna, "Speaker verification under degraded condition using vowel and non-vowel like regions," *in Proc. Centenary Conf. Electrical Engg.,* IISc, Bangalore, India, 2011.
3. G. Pradhan and S. R. M. Prasanna, " Significance of Speaker Information in Wideband Speech," *in Proc. National Conf. on Communication (NCC)*, Bangalore, India, January 2011.
4. 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. on Communication (NCC)*, Bangalore, India, January 2011.