# SPEAKER VERIFICATION USING SUFFICIENT TRAIN AND LIMITED TEST DATA




*Rohan Kumar Das*

# SPEAKER VERIFICATION USING SUFFICIENT TRAIN AND LIMITED TEST DATA

*A*

*Thesis submitted*

*for the award of the degree of*

**DOCTOR OF PHILOSOPHY**

By

**ROHAN KUMAR DAS**


DEPARTMENT OF ELECTRONICS AND ELECTRICAL ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY GUWAHATI

GUWAHATI - 781 039, ASSAM, INDIA

September 2017

# Certificate

This is to certify that the thesis entitled "**Speaker Verification using Sufficient Train and Limited Test Data**", submitted by **Rohan Kumar Das** (126102026), 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

My Father

**Late Parag Kumar Das**

for whom I could not do anything

&

My Mother

**Purabi Das**

for whom whatever I do will fall short

# Acknowledgements

The thesis is incomplete if I fall short of acknowledging the few people behind this. The people who stand beneath, I would like to acknowledge by few of my words.

I am thankful to my research supervisor Prof. S. R. Mahadeva Prasanna for believing in me and allowing to work under his guidance. His passion towards research in terms of regular meetings, stimulating discussions has been an inspiration for all of us. There are many things which we have to learn from him, vision for a research problem, time management, etc. that are very much important for conducting research. During the course of PhD thesis several collaborative works took place with his suggestions, which gave me a direction in my research plan. Also I am indebted to him for arranging funding during the PhD period, without which it would have been difficult in different stages.

I would like to express my gratitude towards my Doctoral Committee (DC) members Prof. Rohit Sinha, Prof. S. Dandapat and Dr. S. Sundaram for their suggestions and insightful comments during the seminars. A special acknowledgment towards the Chairman of the DC, Prof. Rohit Sinha sir for his regular interactions and suggestions whenever needed. I must admit, I am able to learn many things by interacting with him, for which I shall be grateful forever. Further, Dr. S. Sundaram sir always encouraged with his discussions and suggestions whenever I meet him, whether it's in the corridor or in any official meeting.

During the days when I was working in the project, the review meetings of the projects provided a platform to interact with Prof. B. Yegnanarayana as the Chairman of the project review committee, who is an institution himself. His immense passion and zeal towards research can motivate anyone in the world. His suggestions regarding various studies during the project meeting discussions in the vibrant voice still echoes in my ears. Further, I would like to convey my gratitude towards the funding agency of the project e-Security Division, Department of Information Technology (DeitY), New Delhi for providing enormous funding to build advance computational facility in the EMST Laboratory.

Prof. S. Nandi from CSE Department has been always encouraging and suggesting in various activities, which are always fruitful to me. I am thankful to him for allowing me to work on a project for few months during PhD period. Dr. L. N. Sharma, Technical Officer EMST Laboratory maintained all sorts of facilities in the laboratory and helped whenever there is some issue for which I acknowledge. Also, I would convey my gratitude to the Head of the Department, EEE Prof. C. Mahanta and the Office Staff of EEE department, especially Mukut Da for timely forwarding different applications.

Among department faculties my sincere gratitude towards Prof. A. Mahanta, Prof. A. K. Gogoi, Prof. P. K. Bora and Dr. S. R. Ahmed for discussions and suggestions whenever needed.

During the initial days my seniors Dr. Haris B C and Dr. G. Pradhan helped me to understand the details of the system development, which made a lot of things easier later. I shall be thankful to them forever. Further, I would express my acknowledgment towards Dr. Debadatta Pati for several discussions during different stages. Apart from them, I am thankful to the other senior members of the EMST laboratory Dr. Govind D, Dr. S. Shukla, Dr. Deepak K. T., O. P. sir, Ramesh sir, Dr. M. K. Nath, Dr. Sunil Y and Dr. S. Shahnawazuddin, who directly/indirectly helped at different times. A special thanks to Dr. Biswajit Dev Sarma for making me understand different concepts whenever needed during initial period. I shall remain grateful to my close friends Bidisha and Sarfaraz for all those technical/non-technical discussions, which made me alive as a human during tough periods. Also, I am thankful to Sarfaraz for the joint efforts towards different works, especially for various speaker recognition challenges that we participated. I cannot forget the discussions and frustration level shared from time to time with Nagaraj and Banri during PhD period. A thankful note to past/present members of our group Padhy, Anurag, Jiss, Subhadeep, Biju, Bhukya, Subhasis, Himakshi, Mawsumi, Sishir, Parishmita, Akhilesh, Vikram, Protima, Abhishek, Anupama, Deepshikha, Chafidul, Iranee, Pallavi, Soumar, Nazibur, Tilendra, Shikha, Sreeram, Nagendra, Mrinmoy, Moa, Sandeep, Saswati, Vineeta, Prabhakar, Alex, Ato, Ashutosh, Akhil, Salil and the rest for their direct/indirect contributions.

During the PhD period, I was funded by International Speech Communication Association (ISCA), Microsoft Research India, Xerox Research Center India (XRCI) and Science & Engineering Research Board (SERB), Govt. of India for attending Interspeech conference in abroad, for which I would like to acknowledge them. Further, I would offer my sincere thanks to MHRD, Govt. of India for providing the fellowship during final stages of my PhD thesis work. I would again like to thank Sarfaraz for reading the thesis patiently and carefully for making corrections related to writing.

The stay at IITG campus would not have been comfortable without the facilities provided by the institute. The campus life remained lively in presence of my friends, Arghya, Subhra, etc.

Finally, it comes to the family and the relatives who continuously cared for me for my well being since childhood. Especially, my mother for making me strong enough to fight back in life during my struggles and my father who is watching and protecting me from the evil somewhere from the infinity.

*Rohan Kumar Das*

# Abstract

The thesis focuses on speaker verification (SV) from the perspective of application oriented systems and identifies a framework of sufficient train with limited test data as the favorable one. Three different directions are highlighted that have scope towards improving the system performance with limited test data based scenario. These directions are investigated in detail and a combined system is proposed including the conducted explorations.

The source features provide information about the glottal excitation in the form of pitch period, strength of excitation, glottal signal shape, etc. Since the glottis and associated muscle structure are unique for each individual, the information represented by the source features is expected to be distinct for each speaker and can be utilized for SV. Three source features namely mel power difference of spectrum in subbands (MPDSS), residual mel frequency cepstral coefficient (RMFCC) and discrete cosine transform of integrated linear prediction residual (DCTILPR) are explored and their significances for limited test data cases are demonstrated. The source features are found to capture different attributes of source information, which on fusion provides comparable performance to the conventional mel frequency cepstral coefficient (MFCC) based vocal tract features.

The lexical match between train and test sessions is found to be beneficial for having an improved performance. With this motivation a text-constrained model based SV framework has been proposed. In this framework, the explicit utilization of speech content is made in order to have a better performance with limited test data based SV. The work is extended to include speaker-specific phonetic information in terms of the vocal tract constriction (VTC) feature. It captures the level of constriction produced in the vocal tract while producing different sound units, which has definite speaker information.

In the back-end of the SV system, it is necessary to have a suitable pattern recognition approach to handle limited test data. In this regard, kernel discriminant analysis (KDA) has been explored for the discussed SV framework. It maps the data points into higher

dimensional space and performs discriminant analysis. The KDA has successfully demonstrated its importance, especially for limited test data condition. Finally, a combined SV framework is proposed including the stated explorations aiming towards improving the performance in such a scenario.

Further, as the limited test data based SV framework is from the view of practical systems, there comes a lot of issues while going for deployment. Concerning this, few issues are investigated and attempted to address, which are beneficial for improving the SV performance. Mismatch in speaking rate, session variability and template aging issues are considered in this work.

The major contributions of this thesis are as follows.

- Bringing out an SV framework having sufficient train with limited test data suitable for application oriented systems.
- Exploring different attributes of source information and demonstrating their significance for SV with limited test data on fusion.
- Proposal of a text-constrained model based framework and explicit/implicit utilization of speaker-specific phonetic information for speaker modeling.
- Exploring kernel based discriminant analysis and its scope for SV with limited test data.
- A common framework involving different attributes of excitation source features, speaker-specific phonetic information in terms of VTC feature and KDA as suitable pattern recognition technique at the back-end for dealing with the limited test data scenario.
- Investigating some issues related to SV with limited test data from the perspective of practical systems: mismatch speech tempo, session variability and template aging.

**Keywords:** speaker verification, limited data, short utterances, source information, text-constrained model, kernel discriminant analysis, practical systems.

# Contents

# List of Figures

# List of Tables

# List of Acronyms

| | |
|---|---|
| AFA | Acoustic Factor Analysis |
| AM | Arithmetic Mean |
| CCA | Canonical Correlation Analysis |
| CMN | Cepstral Mean Normalization |
| CMS | Cepstral Mean Subtraction |
| CMVN | Cepstral Mean and Variance Normalization |
| CVN | Cepstral Variance Normalization |
| DCF | Detection Cost Function |
| DCT | Discrete Cosine Transform |
| DCTILPR | Discrete Cosine Transform of Integrated Linear Prediction Residual |
| DET | Detection Error Tradeoff |
| DNN | Deep Neural Network |
| DTFT | Discrete Time Fourier Transform |
| EER | Equal Error Rate |
| EFR | Eigen Factor Radial |
| EM | Expectation Maximization |
| EMD | Empirical Mode Decomposition |
| FAR | False Acceptance Rate |
| FDLP | Frequency Domain Linear Prediction |
| FRR | False Rejection Rate |
| GCI | Glottal Closure Instant |
| GFD | Glottal Flow Derivative |
| GM | Geometric Mean |
| GMM | Gaussian Mixture Model |

| | |
|---|---|
| GPLDA | Gaussian Probabilistic Linear Discriminant Analysis |
| HE | Hilbert Envelope |
| HMM | Hidden Markov Model |
| H-norm | Handset Dependent Score Normalization |
| HT-norm | Handset Dependent Test Score Normalization |
| IDFT | Inverse Discrete Fourier Transform |
| IF | Instantaneous Frequency |
| IFCC | Instantaneous Frequency Cosine Coefficient |
| ILPR | Integrated Linear Prediction Residual |
| IVR | Interactive Voice Response |
| JFA | Joint Factor Analysis |
| KDA | Kernel Discriminant Analysis |
| LAR | Log Area Ratios |
| LDA | Linear Discriminant Analysis |
| LP | Linear Prediction |
| LPC | Linear Predictive Coefficient |
| LPCC | Linear Predictive Cepstral Coefficient |
| LSF | Line Spectral Frequencies |
| MAP | Maximum a Posteriori |
| MFCC | Mel Frequency Cepstral Coefficient |
| MHEC | Mean Hilbert Envelope Coefficient |
| ML | Maximum Likelihood |
| MLLR | Maximum Likelihood Linear Regression |
| MPDSS | Mel Power Difference of Spectrum in Subbands |
| NAP | Nuisance Attribute Projection |
| NSWEC | Normalized Singular Value Weighted Eigenvector Coefficient |
| OLA | Overlap and Add |
| PARCOR | Partial Correlation Coefficient |
| PCA | Principal Component Analysis |
| PLDA | Probabilistic Linear Discriminant Analysis |

| | |
|---|---|
| PLP | Perceptual Linear Prediction |
| PNCC | Power-Normalized Cepstral Coefficient |
| PSOLA | Pitch Synchronous Overlap and Add |
| QMF | Quality Measure Function |
| RBF | Radial Basis Function |
| RMFCC | Residual Mel Frequency Cepstral Coefficient |
| SAD | Speech Activity Detection |
| SI | Speaker Identification |
| SNR | Signal to Noise Ratio |
| SOLA | Synchronous Overlap and Add |
| SRE | Speaker Recognition Evaluation |
| SUV | Short Utterance Variance |
| SUVN | Short Utterance Variance Normalization |
| SV | Speaker Verification |
| SVD | Singular Value Decomposition |
| SVM | Support Vector Machine |
| T-norm | Test Score Normalization |
| UBM | Universal Background Model |
| V/UV | Voiced/Unvoiced |
| VAD | Voice Activity Detection |
| VLR | Vowel-like Region |
| VQ | Vector Quantization |
| VTC | Vocal Tract Constriction |
| WCCN | Within Class Covariance Normalization |
| ZFF | Zero Frequency Filtering |
| ZFFS | Zero Frequency Filtered Signal |
| Z-norm | Zero Score Normalization |

# 1

# Introduction

Contents

### Objective of the thesis

*The recent advancements in the area of speaker verification (SV) have shown possibilities for application oriented systems. These systems, however, have to fulfill several requirements to fit under practical scenario. One such prime requirement can be seen as involvement of a limited amount of data for verification of a trial. This is required to provide user comfort and effective decision delivery for regular use. However, the performance degrades with involvement of less data and thus showcases it as an interesting problem statement. The thesis motivates from this and suggests a framework for text-independent* ***speaker verification using sufficient train and limited test data****. Three different potential directions are identified for improving the performance of SV systems in such a scenario. The first one deals with use of voice source features that carry complementary information from that carried by the conventional vocal tract features. Another direction is based on proposal of text-constrained models and vocal tract constriction (VTC) information for speaker modeling. The third one corresponds to suitable pattern recognition approaches for limited test data scenario. It is expected that these three different directions when explored and finally combined to a common platform may be able to handle SV with limited test data. The exploration of the same is made and reported in this thesis. Additionally, some important issues from the practical system perspective are investigated that have significance towards a limited data based SV framework.*

## 1.1 Introduction to speaker recognition

Human beings can be recognized using speech as a biometric feature as each speaker has different style of speech delivery, vocabulary usage and physiological structure of their speech production system. The physiological structure that includes shape and size of the vocal tract, size of the larynx causes difference between the speakers in speech production. Speaker modeling is essential for many tasks, such as speaker recognition, speaker diarization, speaker change detection and speaker clustering. Speaker recognition refers to recognition of a person based on voice samples of that person. Speaker diarization deals with finding who spoke when and is useful to find the speech of a speaker from a conversation of multiple speakers. Speaker change detection refers to the task of finding the regions, where the change of a speaker occurs in a speech recording of conversation containing multiple speakers. Similarly, speaker clustering groups a set of speakers on a similarity basis.

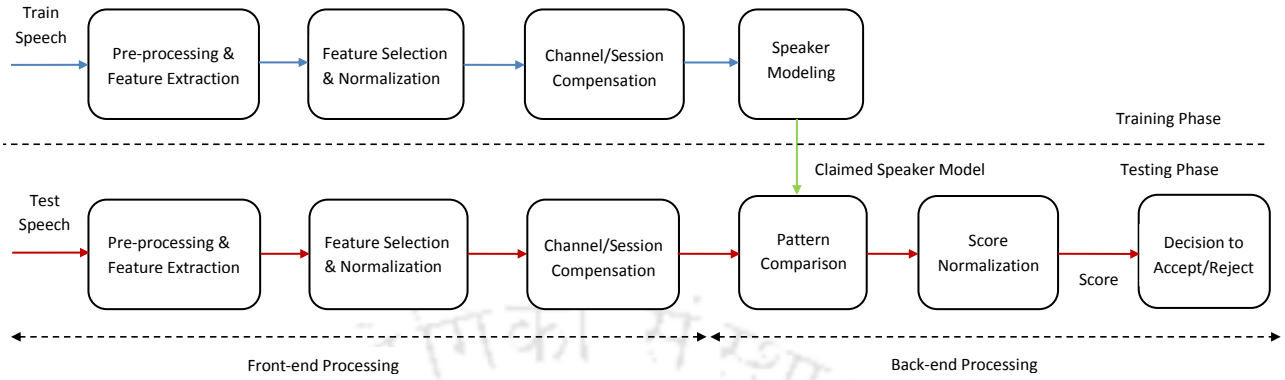


**Figure 1.1:** Generic block diagram of a speaker verification system.

This work focuses on the speaker knowledge utilized for authenticating a person from the speech signal, which is referred to as speaker recognition [1, 2]. Based on its task objective, it can be categorized as SV and speaker identification (SI). In SV, a claim is associated for each test example and the scenario is more often cooperative. On the contrary, in SI, the speaker does not make any explicit claim. It rather attempts to find the best match of the test speech against the set of available trained models of the speakers. The SI systems are broadly used in forensic application. On the other hand, SV systems are practically more acclaimed for deployment purpose. Based on the constraint upon the text content of the speech data, SV systems can be classified into text-dependent SV systems and text-independent SV systems [3, 4]. In a text-dependent SV system, the text is fixed and the speakers have to utter the same text during training and testing. Therefore, a very small amount of data involving a sentence of 3-4 seconds (s) is used in this case. On the contrary, text-independent SV does not put any restriction on text content of the speech data during training as well as testing. Typically, it requires 2-3 minutes of speech data for training and 20-30 s for testing.

## 1.2 A glance towards text-independent speaker verification

Text-independent SV has undergone an evolution in past several decades. A generic block diagram of the SV process is shown in Figure 1.1. It shows different modules that are involved for training a speaker model and then verification of a claim by a test speech with respect to the claimed model. Based on the type of processing involved in SV systems, the modules can be categorized into front-end processing modules and back-end processing modules.

## 1. Introduction

The front-end processing includes extracting relevant features from the speech signal, which carry speaker-specific information, and then applying suitable techniques for feature selection. There are various types of spectral features that are used for extracting speaker information in terms of the vocal tract characteristics [5]. The most widely used among them are mel-frequency cepstral coefficients (MFCCs) [6]. Some other spectral features having potential for practical use are linear prediction cepstral coefficients (LPCCs), line spectral frequencies (LSFs) and perceptual linear prediction (PLP) [7–9]. The voice source features that characterize the glottal source of a speaker carry significant speaker-specific information [10–13]. The literature shows that even though the voice source features are not as much discriminative as vocal tract features, the fusion of the two helps in improvement of performance [4,10–13]. The different types of voice source features are based on linear prediction (LP) residual, parametric glottal flow model parameters, wavelet analysis, residual phase and modulation spectrum. Modulation frequency that contains the information regarding the rate and style of speaking is also used as a feature for speaker recognition [14,15]. Prosodic features are also used in speaker recognition, of which the most important one is the fundamental frequency ($F_0$) [16]. It is found that $F_0$ related feature in combination with spectral features is effective, especially in noisy conditions [17]. Some other prosodic features like duration, speaking rate and energy distributions/modulations are also used for speaker recognition [16–18].

Feature selection techniques are applied on the extracted features to identify the speech regions. Conventionally energy based voice activity detection (VAD) [19] is performed to get the features for the regions of interest, which contain speech from the speaker, discarding silence regions. However, in noisy conditions this technique fails, which degrades the SV performance. Therefore, in order to handle these degraded conditions of speech, especially in noisy environments, different feature selection techniques are proposed. Some of the different robust VAD methods are as follows: periodicity based VAD [20], statistical VAD [21], vowel & non-vowel like region selection [22, 23] and self-adaptive VAD [24]. After selecting the features from region of interest, it is necessary to normalize the features in order to nullify the common offset for channel/session compensation, which gives better speaker discrimination. With MFCC and LPCC features, the normalization is termed as cepstral mean subtraction (CMS) or cepstral mean normalization (CMN) as performed in the cepstral domain [25]. Moreover, cepstral variance normalization (CVN) is performed on top of CMS or CMN to normalize the features to fit zero mean unit variance distribution [26].

The extracted speaker-specific features can be either directly used or further processed by various algorithms for speaker modeling. In last few decades, speaker modeling has evolved from vector quantization (VQ) [27], Gaussian mixture models (GMMs) [28], GMM with universal background model (GMM-UBM) [29], support vector machines (SVMs) [30], joint factor analysis (JFA) [31] to the state-of-the-art i-vector modeling [32]. The i-vector based speaker modeling represents the large-size GMM supervector of each utterance into a low dimensional vector that possesses dominant speaker characteristics. Further, to have channel/session compensation over the speaker models, various techniques like linear discriminant analysis (LDA) [33], within class covariance normalization (WCCN) [34], nuisance attribute projection (NAP) [35] and probabilistic linear discriminant analysis (PLDA) [36] are used. Recently proposed acoustic factor analysis (AFA) based framework for SV, which hypothesizes that the conventional acoustic factors reside in a lower dimensional subspace, has been found to work well over the standard baseline framework [37]. The work of [37] is based on the i-vector based modeling structure, where the UBM is trained using the acoustic factors instead of the feature dimensions. The authors extended this work to propose maximum likelihood AFA (ML-AFA), which performs better under degraded conditions [38]. This is achieved as the acoustic factors are updated in each iteration during training the UBM model [38]. Motivated by the recent developments with respect to deep neural network (DNN), it has been implemented for SV to extract Baum-Welch statistics in the i-vector based modeling [39]. The success achieved by using DNN has led to many other works developed using DNN i-vector based framework [40]. The channel/session compensation techniques for this approach remain similar to that of those in the i-vector based modeling. The channel/session compensated speaker models are used for validation of a claim by testing via approaches such as distance computation in terms of various measures like Mahalanobis distance, Euclidean distance, log likelihood calculation and cosine kernel scoring. These different scoring methodologies yield scores, that can be further normalized based upon the task and the conditions of the trials. Some of the widely used score normalization techniques are as follows: zero normalization (Z-norm) [41], test normalization (T-norm) [42], handset normalization (H-norm) [43] and handset dependent T-norm (HT-norm) [44]. It has been seen from the literature that combining multiple classifiers provides significant improvement. There are different toolkits available such as Bosaris, Fusion & Calibration (FoCal) to fuse the scores obtained from multiple classifiers. They use the method of logistic regression for fusing the scores that helps in achieving improved SV performance [45].

## 1.3 Motivation of the current work

This section highlights the motivation of this thesis in detail. It discusses regarding the emerging application based systems using speech as a biometric measure. Further, the recent attention towards limited data based SV involving short utterances is reported. The importance of the limited data based SV framework is also discussed from the perspective of field deployable systems.

### 1.3.1 Emerging application oriented systems and insights

The area of SV has witnessed significant advances with the development of various techniques suitable for modeling speaker characteristics. These advances in the field of SV have opened doors towards field deployable systems for person authentication. There have been several attempts made for the development of person authentication systems in a practical scenario. Remote person authentication has been carried out using SV, that shows the potential for deployable systems [46]. The authors of [47] have developed a smart home security application for controlling different household activities using short utterances with the knowledge of speaker and speech recognition. In [48], a speech biometric attendance system is developed for marking attendance over online telephone based network. A real-time hardware is implemented using field programmable gate array platform for SV with less time complexity involved as mentioned in [49]. Recently, a multi-level SV system has been proposed that uses different modalities of SV, which are combined in a sequential order for verification of a claim over a telephone based network [50]. All these works showcase SV as an emerging area for having deployable systems with real-world applications.

With the motivation of using speech as a biometric in application oriented services with a low security measures, a speech biometric attendance system is developed for regular student attendance. The system is implemented over an online telephone network for marking student attendance, which is handled by an interactive voice response (IVR) based system. The IVR system callflow is developed through ISDN-PRI (integrated services digital network-primary rate interface) line that can handle telephone channel calls through computer telephone interface card. It can handle up to 30 parallel calls over the telephone channel. The IVR is hosted on a voice-server, which runs the Asterisk software. It is a software implementation of a telephone private branch exchange, that allows the server to handle incoming calls and make outgoing calls to and from other public switched telephone network or voice over internet protocol services. This callflow is made for implementation of the attendance system of

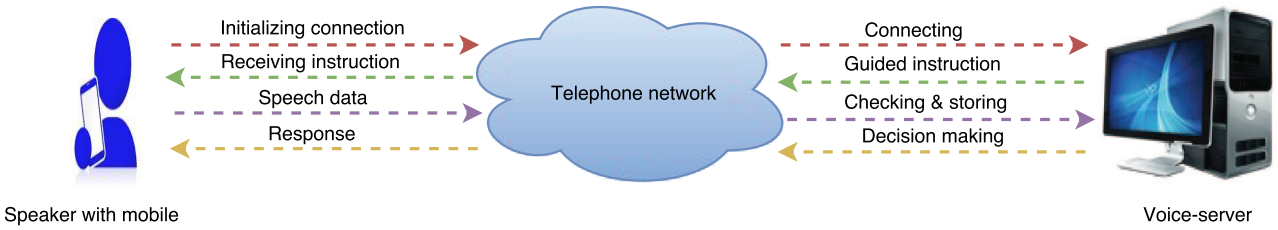


**Figure 1.2:** Overview of speech biometric based attendance marking system over the telephone network.

postgraduate students of EEE department at our institute. Students have to call to the designated toll-free number through which the IVR system callflow helps for enrollment as well as regular testings in a practical setting. Figure 1.2 shows the overview of the speech biometric attendance system over the telephone network developed for regular attendance marking. The system is based on i-vector based speaker modeling in a text-independent SV framework, which considers 3 minutes of speech for training and 20-30 s of speech input for testing.

The developed system is allowed to be used by the students for a period of two months on a trial basis for obtaining feedback. Most of the students had a problem with respect to the amount of speech data asked for verification of a trial. It is found that providing speech data of 20-30 s duration regularly to mark attendance burdened the users. Many of the students gave feedback to reduce the speech duration asked during testing. But when the speech duration asked during testing is reduced to around 10 s or less, the performance degraded by a large margin. However, it provided comfort to the users in terms of less speech input to be given during testing. Thus, this system implementation based experience also showed one basic need of having a short utterance based SV framework involving limited data, particularly in test sessions. This made our research proposal of SV with limited test data more prominent and interesting within the context of practical systems.

### 1.3.2 Attention towards short utterance based speaker verification

The NIST speaker recognition evaluations (SREs) provide the benchmark of the SV systems in terms of performance evaluation. These evaluations are held annually/biannually and the data provided for the evaluations vary from year to year. Based on the NIST SRE evaluation plans during 1999-2016 [51], a comparative study has been made to observe the trend in duration of data provided for training and testing sessions for the speakers. Table 1.1 shows the duration of train/test data for different year evaluation plans. It can be noticed that the duration of the segments involved for

**Table 1.1:** Duration of utterances from train/test session of NIST SRE databases in chronological order.

| Evaluation | Train Duration | Test Duration |
|---|---|---|
| SRE 1999 | 2 minutes | 15-45 s |
| SRE 2000 | 2 minutes | 15-45 s |
| SRE 2001 | 2 minutes | 15-45 s |
| SRE 2002 | 2 minutes | 15-45 s |
| SRE 2003 | 2 minutes | 15-45 s |
| SRE 2004 | 10 s, 30 s, 5 minutes | 10 s, 30 s, 5 minutes |
| SRE 2005 | 10 s, 5 minutes | 10 s, 5 minutes |
| SRE 2006 | 10 s, 5 minutes | 10 s, 5 minutes |
| SRE 2008 | 10 s, 5 minutes, 8 minutes | 10 s , 5 minutes, 8 minutes |
| SRE 2010 | 10 s, 5 minutes | 10 s, 5 minutes |
| SRE 2012 | SRE 2006-2010 | 30 s, 100 s, 300 s |
| SRE 2016 | 60 s | 10-60 s |

train and test sessions changed from NIST SRE 2004 onwards to include short segment of speech data having 10 s duration. The recent evaluations apart from concentrating on different scenarios like channel, environment, language and handset mismatch, are also found to focus on short train and test segments due to their significance towards practical systems for real-world applications. This altogether depicts short speech segment based SV as a prime need for application oriented systems.

Apart from the NIST SRE databases focusing towards short utterance based SV systems, there have been other recent databases released for making studies in such a scenario. The RSR2015 database has been designed with text-dependent SV modality, which has three different parts specific towards frameworks of short utterances [52]. The first part is based on conventional text-dependent based SV framework involving 30 fixed phrases, whereas the second part deals with short commands of 2-3 words that can be used for smart home application purposes. On the other hand, the third part of the database considers digit sequence based short utterances for studies. Another database has been collected with reference to the short utterance based studies in a collaborative effort of 21 different countries across the world over a period of one year. This database is named as RedDots database that has four different parts for short utterance SV based studies [53]. The first part considers conventional text-dependent SV based framework, where the short phrase is common across all the users. The second and third part contain fixed phrases, which are unique for every user that are given and user chosen, respectively. Finally, the fourth part involves free choice fixed phrases that are unique across the sessions for each speaker. This subset has two different enrollment conditions,

which are named as text-dependent and text-prompted. The latter refers to a text-independent based framework, where multiple unique short phrases are taken for training and a different phrase for testing. These databases project the significance of short utterance based SV in the present world and its scope for application oriented systems in the future. Different evaluations have been organized with these recent databases on short utterances with the possibilities of coming out with novel ideas and techniques. These may be useful for addressing different issues towards field deployable systems.

### 1.3.3 Significance of speaker verification with limited data

The limited data based scenario comes into the picture when the focus is towards person authentication for practical deployment. This is because the time involved for authenticating a person is very small with most of the other biometric attribute based systems used in practice. The use of short utterances for SV, not only reduces the required testing time, but also provides comfort to the speakers as they are less burdened with speaking. In this regard, the text-dependent SV is found as a favorable candidate for having a deployable system due to small amount of train/test time involved. However, there is a very high chance of spoofing by the unauthorized users as the fixed phrase is global across all the users of the system. Additionally, in case of text-dependent SV, less speaker variability is captured due to consideration of a small phrase for speaker modeling. On the contrary, the text-independent SV captures the speaker characteristics in a more generic manner and is less susceptible to the attackers. Further, as there is no constraint on the users on what is to be spoken, it is more convenient for the users. Because of all these factors, text-independent SV is considered as a better choice for implementation of a robust system.

The introduction of i-vector based speaker modeling gave a major breakthrough in the field of SV and provides a benchmark for text-independent SV studies [32]. Different studies show that as the duration of speech segments for verification of a trial becomes less, it affects the performance by a perceivable margin [54]. This drop in performance is mainly due to the poor speaker information extracted as very small amount of speech data is available. However, from the point of a practical system, short segment of speech is expected from the users. In this direction, there is a need to have alternative features capturing different speaker characteristics, robust modeling techniques and different classifiers for having a better characterization of speakers with limited data. This makes SV using limited data as an interesting domain for exploration, which can have investigations across different modules of SV.

The SV using limited data in terms of short utterances can be explored in different directions from the perspective of deployable systems. The thesis identifies some of these directions and then presents an attempt to achieve an improved speaker characterization using limited data. The different directions considered in this work include exploration of alternative/complementary features, utilization of speaker-specific phonetic information and suitable pattern recognition techniques useful from the perspective of limited data based SV.

## 1.4 Organization of the thesis

The problem statement for *speaker verification using sufficient train and limited test data* is addressed in the following manner.

- The **Chapter 2** provides a review in the area of SV from the perspective of limited data in terms of using short utterances. The review is based on possible directions for handling the limited data based SV framework. Three different directions are put forward that are categorized on the basis of similar aspects. These directions can be seen as use of alternative/complementary information, acoustic-phonetic information and suitable pattern recognition and normalization techniques. A sketch of a framework involving these stated directions is put forward as a favorable architecture for improving the performance of SV with limited data.
- The **Chapter 3** portrays the scope of SV using limited data for field deployable systems. It then identifies sufficient train with limited test data as a favorable framework. As a first step to deal with this scenario, different voice source features are explored. The three voice source features, mel power difference of spectrum in subbands (MPDSS), residual mel frequency cepstral coefficient (RMFCC) and discrete cosine transform of integrated linear prediction residual (DCTILPR) are found to have different attributes of excitation source information. These features on fusion helped in achieving an improved SV performance.
- The **Chapter 4** explores the significance of lexical match, which can be utilized for SV with limited test data. A text-constrained model based SV framework inspired from the text-dependent modality is proposed. It explicitly provides a lexical match between train and test sessions. This framework is first investigated for limited train and test condition followed by sufficient train with limited test data scenario. The scope of the source features in the context of text-constrained model is also studied that signifies their importance. Further, VTC feature is used for capturing

speaker-specific information in an implicit manner that improves the SV performance on fusion with conventional MFCC features.

- In **Chapter 5**, the prospect of kernel discriminant analysis (KDA) is discussed for SV and shown that the SV systems are benefited more for limited test data based scenario. The KDA based channel/session compensation used at the back-end of SV systems transforms the data points into a higher dimensional space. Then it performs discriminant analysis in the transformed domain, which is favorable for SV with limited test data. The chapter also proposes a combined framework using the explorations made in the previous chapters on different attributes of source information, VTC information along with KDA at the back-end. The proposed combined system is able to achieve a commendable improvement over the baseline system showing scope towards practical systems under limited test data.
- The **Chapter 6** investigates some issues towards practical systems and possible frameworks for dealing with those. One such issue is based on having a mismatch in speaking rate between train and test sessions. A framework using a prosody modification method is proposed to change the speaking rate according to the mismatch factor to compensate the mismatch in speech tempo. Session variability and template aging are another two issues that come into the picture for deployable systems over a period of time. Studies are carried out to get a better understanding of these issues and different frameworks are proposed to have a more reliable speaker model from long term practical system perspective.
- Finally, the **Chapter 7** discusses the summary of the thesis highlighting the prime contributions with possible future directions.

# 2

# Speaker Verification from Limited Data Perspective: A Review

Contents

### Overview

*This chapter portrays the review of different works related to speaker verification (SV) which are useful from the perspective of limited data based scenario. Three different directions are categorized based on their similar aspects that can be seen as, use of alternative/complementary features, utilization of acoustic-phonetic information and suitable pattern recognition and normalization techniques for handling limited data. The identification of these directions for addressing the problem statement has been made on the basis of their proven significance and impact towards dealing with SV using short utterances involving limited data. The review is followed by a discussion to bring out a framework involving the stated directions into a common platform for having a deployable system using limited data based SV.*

## 2.1 Introduction

The motivation of the thesis as discussed in Chapter 1 projects text-independent SV involving limited data as the prospective candidate from the view of field deployable systems. This work mainly focuses on reviewing text-independent SV with reference to limited data and projecting the scope towards future. Generally, limited data is referred to as short utterances in the literature. In this regard, different attempts which have been made or shown useful for the short utterances are reviewed. The work with respect to fixed phrase based short utterances (text-dependent) having correlation towards the limited data based framework of text-independent modality are also considered. A brief review of approaches to text-independent SV to show its evolution has already been presented in Chapter 1. This chapter reviews the directions that are suitable from the perspective of having a limited data based framework. These directions are mainly grouped into three categories based on their similar aspects. The first one is regarding exploration of different features having alternative/complementary information, which may be useful to capture speaker characteristics in a better way. The works based on acoustic-phonetic information and their explicit/implicit utilization for SV are grouped in another category that is having an impact towards handling limited data. Different pattern recognition approaches and normalization techniques from the perspective of handling limited data based SV are categorized under one category. These three different directions may be combined together to develop a common framework for having a practically realizable text-independent SV system under limited data based scenario.

The remaining part of the chapter is organized as follows. In Section 2.2, the different features having significance towards capturing speaker characteristics, which are useful for dealing with limited data are explored. Section 2.3 details about the acoustic-phonetic information utilization based work done for SV, which has importance towards limited data. The different modeling approaches, classifiers and normalization techniques from the view of limited data based framework are mentioned in Section 2.4. A discussion is made in Section 2.5, where the explored directions are attempted to fit into a common framework for the benefit of having an SV system with limited data based framework. Finally, Section 2.6 presents the scope of this review towards achieving the thesis objective.

## 2.2 Different features for speaker modeling

When there is very limited amount of speech available in the task of SV, it becomes highly challenging to recognize the intended person correctly. In this regard, features having complementary/alternative characteristics can be used to extract maximum information for speaker modeling. The different kinds of features categorized based on their origin, processing and nature, are presented in the following subsections.

### 2.2.1 Spectral features

The conventional SV systems widely use MFCC based vocal tract features for speaker modeling. They are obtained by computing the log magnitude spectrum of the speech signal, passing it through a non-linear filterbank and finally applying the discrete cosine transform (DCT). The MFCC based vocal tract features may not contain all the necessary information to perform speaker verification. To overcome this shortcoming, different features having alternative/complementary information are required to capture speaker information from the speech signal.

There are different spectral features available other than the conventional MFCC features that capture speaker characteristics. Linear prediction (LP) analysis provides an alternative way for estimating the spectrum of a speech signal [7, 55]. In LP analysis, the prediction of current sample is obtained from the linear combination of past $p$ samples, where $p$ is the order of prediction [7]. The difference between the actual sample and the predicted sample gives the predicted error and is termed as the LP residual signal. The predictor coefficients are transformed using different ways to obtain different types of feature representation. Some of the commonly used feature representations of the linear prediction coefficients are as follows: linear predictive cepstral coefficients (LPCCs), line

spectral frequencies, perceptual linear prediction (PLP) coefficients, partial correlation coefficients (PARCORs), log area ratios (LARs) and formant frequencies and bandwidths [7–9, 56]. The studies on fusion of systems built by using different spectral features shows that there exists complementary information and the fusion helps to improve the performance [57,58]. The authors in [59] have recently explored the fusion of the following three types of spectral features: frequency domain linear prediction (FDLP) [60], mean Hilbert envelope coefficients (MHECs) [61] and power-normalized cepstral coefficients (PNCCs) [62]. The authors then captured and characterized spectral variation by sliding a window over the spectral features to propose a feature namely normalized singular value weighted eigenvector coefficient (NSWEC). This feature being complementary to the existing features, yields an improvement on fusion with each of them. Additionally, the fusion of all the considered features also shows a trend of achieving an improved performance. The experiments are carried out using the standard NIST SRE database as well as the RSR2015 database designed for short utterance based SV studies. The results of these studies demonstrate that fusion of different spectral features can be helpful for SV with limited data.

### 2.2.2 Voice source features

The voice source features capture the glottal signal characteristics which are unique for every speaker. The LP residual signal contains the source information and is processed in different ways to extract the excitation source characteristics. Some of the widely used vocal tract features are the parametric glottal flow model parameters [10], LP residual phase [13], cepstral coefficients [12,63] and wavelet analysis based features [64]. The work of [65] presents a comparison of explicit and implicit modeling of the subsegmental information. The source features such as residual mel frequency cepstral coefficient (RMFCC) and mel power difference of spectrum in subbands (MPDSS) are used for the studies in this work [65, 66]. Some recent explorations in the direction of voice source features can be seen in [67]. In this work, the integrated LP residual (ILPR) signal that closely resembles the glottal flow derivative (GFD) signal is considered. It is processed in a pitch synchronous manner, on which DCT is taken and the feature is referred to as discrete cosine transform of ILPR (DCTILPR). The use of this feature is shown successfully into a speaker identification task depicting its capability to carry definite speaker characteristics. The authors of [68] proposed a novel feature called as the instantaneous frequency cosine coefficient (IFCC). The instantaneous frequency (IF) is computed with the motivation that it is free from the problem of phase warping. The narrow-band components of

speech are taken to compute the IF and then DCT applied on that to extract IFCC features as a compact representation. The IFCC features are compared with the MFCC and FDLP features that demonstrated it as a prospective voice source feature. Additionally, the fusion of IFCC features with the other two features gave an improved performance showing complementary nature of information. The literature shows that the voice source features, which represent excitation source characteristics are having complementary information than that captured by the vocal tract features. Further, the amount of speech required for train and test can be less while using the voice source features to that with the vocal tract features [11,69]. This is mainly due to the fact that the voice source features rely less on the phonetic content, which is different from the nature of the vocal tract features. Although the performance with the use of stand-alone voice source features is not significant, their fusion with the vocal tract features is helpful for speaker modeling to achieve an improved performance [13, 64]. Thus, the exploring the use of voice source features can be viewed as an important direction for SV with limited data.

### 2.2.3 Spectro-temporal features

Apart from the features discussed earlier, there are different attempts to capture the spectro-temporal signal information like energy modulation and formant transitions. Basically, it is captured in terms of the first and the second order derivatives commonly known as deltas ($\Delta$) and delta-deltas ($\Delta$-$\Delta$). These are added to the base coefficients for including temporal spectral characteristics [25,70]. Other types of spectro-temporal features are the time frequency principal components [71], data-driven temporal features [72] and modulation frequency [15]. As each of these types of features provides additional speaker characteristics, they can be useful for SV with limited data.

### 2.2.4 AM-FM analysis based features

From the AM-FM analysis of speech signal, there are features that can be used for capturing speaker-specific information. These features capture the non-linear and non-stationary aspects of speech signal and are found to be more useful for degraded condition [73–75]. Unlike the conventional MFCC features, which use non-linear mel filterbanks, these features use a parallel bank of overlapping band pass filters for the extraction of many AM-FM signals. The modern AM-FM based analysis techniques use empirical mode decomposition (EMD) based features. These features have complementary information to the existing vocal tract or source based features and their fusion helps in improving

the performance [76–78]. Therefore, there is scope for using the AM-FM analysis based features for limited data based SV studies along with conventional features.

### 2.2.5 Prosodic features

There are some more aspects of speech signal that contain speaker characteristics. Prosody is one such aspect that relates to speaking rate, intonation and rhythm. This information can be useful for speaker modeling when the amount of data available for SV is very small. The most widely used prosodic feature is the fundamental frequency ($F_0$) that captures the information about the physiological structure (larynx size) [79]. Its mean value has been used for SV task by many researchers [80,81]. The mean value of $F_0$ is also combined with other evidence that helps in achieving a better SV performance [18,80]. The authors in [82] have used different prosodic features and their representations for speaker and language identification. This shows their importance for the stated exploration. In [83], prosodic features like slope, curvature, duration, pitch contours, energy contours and their combinations are explored. The combination of duration with pitch and energy contours produced the best results among them. Further, significant improvement is achieved when the prosodic features are combined with baseline MFCC features in the JFA based modeling approach. Thus the prosodic features can also be used as an alternative representation to capture speaker characteristics. The prosodic features can be utilized along with the other features for SV under limited data scenario.

### 2.2.6 Speaker-specific high-level features

There are some high-level features that normally depend on the vocabulary of a speaker, which can be referred to as idiolect based features for SV [84]. These are considered as words, phones or in terms of some prosodic gestures as can be seen from different works [17,84,85]. The work of [86] proposed an $N$-gram based model for SV, which is based on grouping phonetic tokens. The grouping of two tokens leads to a bigram model, grouping of three tokens leads to a trigram model and so on. Tandem features are another way to use phoneme posterior probabilities to convert them as a feature representation. Different works have been reported with respect to the tandem features using Gaussian mixture model (GMM) and recently popular deep neural network (DNN) based frameworks [40, 87, 88]. Similar to the tandem features, there are bottleneck features, which are obtained by training a neural network with a bottleneck layer in the middle. These bottleneck features in fusion to the baseline setup help in achieving an improved SV performance as reported [89,90]. Thus this kind of information can also be

used along with the conventional features for extracting speaker characteristics in limited data based SV for field deployable systems.

In order to have a better speaker characterization using limited amount of data for SV, different features having complementary/additional information can be used to extract sufficient speaker characteristics. The other two directions for dealing with limited data condition are discussed in the following sections.

## 2.3 Acoustic-phonetic information for speaker modeling

Similar to the direction of using alternative features capturing speaker characteristics for limited data scenario, another direction can be seen as acoustic-phonetic information for speaker modeling. The text-dependent SV has an edge over the text-independent SV in terms of having the same lexical content during train and test sessions. With this motivation if the acoustic-phonetic information is used in an explicit or implicit manner, its advantage is expected. Different works have used acoustic-phonetic match between train and test sessions for building an improved version of SV system.

### 2.3.1 Phonetic class pronunciation modeling

The work in [91] uses phonetic-class pronunciation modeling for SV. The articulatory feature based conditional pronunciation models represent the pronunciation characteristics of the speakers. When the amount of speech data is limited, the phoneme dependent pronunciation based models lead to speaker models with less discrimination. The similar phonemes are grouped together based on their characteristics to represent background models and speaker models as phonetic-class dependent density functions. The phonemes are grouped using the following three methods: performing vector quantization (VQ) in the phoneme-dependent universal background model (UBM), employing phoneme properties from classical phoneme tree and combining VQ with phoneme properties. The SV studies performed using GMM-UBM based framework showed a better performance using the proposed phonetic-class pronunciation based modeling. This indicates that the phoneme-dependent UBM can be a better choice for speaker adaptation while dealing with limited data based SV.

### 2.3.2 Vowel-like region and non-vowel-like region based segmentation

The authors of [22] have explored SV under degraded condition and have proposed a method of modeling the speakers using only the vowel-like regions (VLRs) in the speech segments. The VLRs

belong to the high energy regions that include vowels, semi-vowels and diphthongs. These regions are found to be more robust towards noise. Additionally, the match of acoustic-phonetic information using the VLRs of train and test sessions, helps in achieving a better performance. This work has been further extended by the authors to have a two-class based segmentation of speech signals for processing. Based on the acoustic-phonetic information, the speech signal has been classified into two broad categories, namely, VLRs and non-VLRs as reported [23]. The authors have processed the speech signal with respect to these two categories separately for building two different SV systems over i-vector based framework. Finally, a score level fusion is done with the SV systems developed with two different segmented categories. The improved performance obtained with the fused system highlights the importance of the acoustic-phonetic match of the features from train and test sessions.

### 2.3.3 Vowel category based segmentation

The authors of [92] use the vowel category based information for SV using short utterances. The vowels of English and Chinese languages are considered and are detected using phoneme recognizer from the development database. Then using each category of vowels as a class, separate UBMs are trained. Given a train utterance, the phoneme recognizer is used for detecting the vowels and then maximum a posteriori (MAP) adaptation is done with respect to the UBM of the corresponding vowel category. The phoneme recognizer is then applied to the test speech for identifying different categories and tested against the adapted vowel category model of the claimed speaker. Finally, a score level fusion is made from the scores obtained for different vowel categories for verification of trial. This shows improvement with reference to the short utterance based cases considered by the authors.

### 2.3.4 Content matching

In order to have a content match of the train and the test sessions in a text-independent SV framework similar to that of the text-dependent modality, the authors of [93] proposed a novel approach. This method proposes to have a content match based speaker model by having a scaled version of the zeroth order statistics of the train data with respect to the test data in i-vector based speaker modeling. Further, it has been shown that use of DNN for extracting the posterior probability is helping more to have a better match of content than that of using a conventional UBM model. This method showed a significant improvement in the short utterance based scenario demonstrating the effectiveness of content matching in text-independent framework similar to a text-dependent SV system.

### 2.3.5 Acoustic factor analysis

With the dominance of factor analysis approaches in the field of SV in past several years, the authors of [37] introduced acoustic factor analysis (AFA) for SV. It is based on the hypothesis that the conventional acoustic factors reside in a lower dimensional subspace. With this approach, the UBM parameters are transformed into acoustic factors (say 32 acoustic factors) of a language from the given feature dimension. This transformed UBM parameters are then used for extraction of statistics from the train and the test data followed by i-vector based speaker modeling. The AFA based method is found to outperform the conventional approaches due to the phonetic compensation made by having transformed UBM. However, this approach fails in the presence of speech signal under noisy conditions. To overcome this drawback, the authors extended the work to propose maximum likelihood-acoustic factor analysis (ML-AFA) approach. In this method, the UBM parameters are transformed into the acoustic factors in an iterative way at the end of every iteration of expectation maximization (EM) algorithm during UBM training [38]. This modification in the AFA based framework is able to perform well in case of degraded speech conditions. The removal of the nuisance dimensions of the acoustic factors in every iteration makes the transformed UBM more robust to noise. The AFA and ML-AFA techniques exploit the acoustic factors of a given language and then perform SV with respect to train and test speech to represent them in terms of these acoustic factors. It thereby provides a better match of train and test sessions. Thus both AFA and ML-AFA approaches have the scope for improving the performance of SV with limited data.

### 2.3.6 Acoustic-phonetic information in background modeling

The acoustic-phonetic match of train and test sessions is found to be important as discussed. Apart from this, if the background models are built using the utterances having the same phonetic content, it can help in a better speaker characterization. In this regard, few works have been made to develop the background models with data having the same phonetic content to that of train and test from the perspective of text-dependent SV. The authors of [94] mention that even though the conventional i-vector modeling performs poorly under limited data, the phonetic constraints on background models may help to improve SV performance. It is shown that estimating within class covariance normalization (WCCN) and eigen factor radial (EFR) by the development data of the same phonetic content enhances the SV performance. Similar to this work, the authors in [95] investigate the significance

of a phonetically constrained probabilistic linear discriminant analysis (PLDA) model by training it with development data of the same phonetic content. The use of this kind of PLDA model gave an enhanced SV performance due to better estimation of PLDA model with the same phonetic content.

The works discussed throughout this section thus showcase the importance of the acoustic-phonetic information for speaker modeling. Different works reviewed in this regard highlight that an improved SV performance may be achieved when explicit/implicit consideration of phonetic content of a speech signal is made. This projects it as an insightful direction for dealing with the SV using limited data.

## 2.4 Pattern recognition approaches and normalizations

The previous sections focused on investigating complementary/alternative features, acoustic-phonetic match based work having importance for SV with limited data. This section highlights the favorable work for limited data based SV with reference to the pattern recognition and normalization techniques.

### 2.4.1 Singular value decomposition as a matching measure

In real-world based practical system conditions, noise may be present along with the speech input from the users. The work of [96] explored text-independent SV with short utterances in noisy conditions and proposed a method based on singular value decomposition (SVD) for matching. This technique considers the ratio of singular values of a matrix, which is obtained by the test feature and the average reference features from the constructed database. The proposed SVD based distance matching algorithm is compared with other conventional distance measures such as Euclidean, the weighted and the Mahalonobis distances showing better results more specifically towards the noisy conditions. The noisy conditioned examples are created by adding noise of 0 dB to 20 dB for the database involving short segments of around 3-6 s duration.

### 2.4.2 Gaussian PLDA

The duration of speech utterances plays a crucial role in evaluating the performance of an SV system. However, as discussed earlier, there arises a need to have limited data in a practical system. In this regard, it has been shown that a PLDA based framework can be more suitable for arbitrary duration utterances than the conventional approaches [97]. Additionally, the comparison between Gaussian vs. heavy-tailed PLDA is shown by the authors of [97], which projects Gaussian PLDA as a better option for i-vector based framework for achieving enhanced performance.

### 2.4.3 Spherical normalization

The authors of [98] have explored the effects of short and mismatched duration utterances on i-vector based SV framework. They then highlighted few directions to tackle such a scenario. The studies considered PLDA for channel/session compensation and scoring. When the PLDA model is trained using long segments of speech then the performance is found to be poorer for testing made with short duration utterances. In order to compensate this mismatch, the authors have mentioned regarding post processing of the i-vectors before building the PLDA models. Two kinds of normalizations are explored for the same. These can be seen as standardization and spherical normalization. The first one refers to have a zero mean unit variance normalization and then length normalization. In spherical normalization, the i-vectors are normalized by an iterative algorithm by which they are transformed to lie on the spherical surface. The two approaches are compared and it is shown that the latter works better with short utterances.

### 2.4.4 Minimax i-vector extractor

With the popularity of i-vector based SV in the recent years, it is considered as the benchmark for SV studies. However, as discussed already, in presence of limited speech data, the performance of such systems drops significantly. In order to achieve an improved performance with short duration utterances, the authors proposed a novel approach based on minimax strategy for extracting i-vectors [99]. They have used a minimax strategy for reestimation of Baum-Welch statistics that helped in modeling the speaker in a better way by creation of a more robust i-vector.

### 2.4.5 Short utterance variance normalization

In [100, 101], the authors have explored the short utterances based framework for i-vector based SV system. They have observed that the i-vectors of the short duration utterances from the same speakers vary from one another by a large margin. It is hypothesized that this variation is due to the difference in capturing of phonetic content by the i-vectors. The i-vectors generated from long duration utterances contain sufficient vocabulary of the language. On the other hand, the i-vectors of short utterances have very limited vocabulary of the language due to small amount of data available and thus capture limited speaker information. This directs towards a need for transforming the i-vectors of the long and short duration utterances into the same domain to have a better verification of a trial. The authors have proposed a technique called as the short utterance variance normalization (SUVN)

to compensate the mismatch in phonetic content. In this technique, the development data with long duration utterances and their truncated short duration utterances are taken to extract respective i-vectors and then normalization is performed. The short utterance variance (SUV) matrix $\boldsymbol{S}_{SUV}$ is calculated as,

$$\boldsymbol{S}_{SUV} = \frac{1}{N}\sum_{n=1}^{N}(\boldsymbol{w}_n^{full} - \boldsymbol{w}_n^{short})(\boldsymbol{w}_n^{full} - \boldsymbol{w}_n^{short})^T \tag{2.1}$$

where, $\boldsymbol{w}_n^{full}$ and $\boldsymbol{w}_n^{short}$ are the LDA transformed full and short utterance i-vectors of the truncated $n^{\text{th}}$ utterance, respectively and $N$ is the total number of utterances. Then the transformation matrix $\boldsymbol{G}$ is obtained using Cholesky decomposition of the inverse of the SUV matrix $\boldsymbol{S}_{SUV}$ as, $\boldsymbol{G}\boldsymbol{G}^T = {\boldsymbol{S}_{SUV}}^{-1}$. Using the obtained $\boldsymbol{G}$ matrix, the LDA transformed train and test i-vectors are further transformed as $\hat{\boldsymbol{w}}_s = \boldsymbol{G}\boldsymbol{w}$, where $\hat{\boldsymbol{w}}_s$ is the SUVN compensated i-vector. These normalized i-vectors are then considered in the SV framework for evaluation of a trial against the claimed speaker.

### 2.4.6 Duration mismatch compensation

The work of [102] proposed a novel method for compensating the duration mismatch of train and test sessions for SV studies. The work demonstrated that the number of detected phonemes in an utterance increases in an exponential manner with the increase of utterance duration. This makes duration variability as an additive noise for i-vector space, which is required to be compensated. Three different directions are proposed by the authors for removing the effect of duration mismatch. The first one refers to modeling PLDA with multi-duration training utterances, which is done by having different truncated versions of the long utterances so that the number of i-vectors for PLDA training increases and more speech content information can be utilized. Another approach deals with score domain calibration using quality measure function (QMF), which considers the amount of speech available after VAD. The third approach is based on creation of synthetic i-vectors of the truncated utterances by addition of some Gaussian random vectors having an intra-segment covariance matrix to the full duration i-vector. These three methods when combined altogether is found to perform well for short duration based test conditions.

The different pattern recognition approaches and normalization techniques suitable for SV framework with limited data involving short utterances are discussed in this section. These approaches can be used along with the other directions discussed in the previous sections for having an improved speaker characterization.

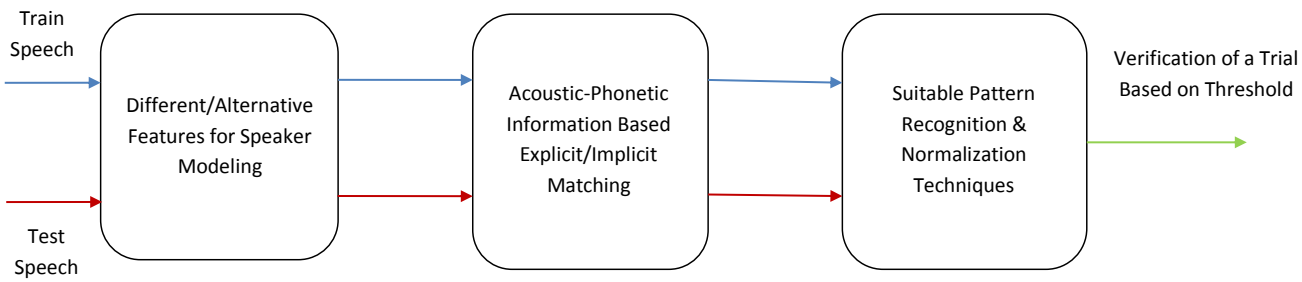


**Figure 2.1:** Block diagrammatic representation of a common framework involving three different directions.

## 2.5 Discussion

In this chapter, the three different directions from the perspective of SV with limited data are identified. These directions can be used in a common SV framework to have their impact on the overall system. In the front-end analysis, alternative/complementary feature representations can be considered for capturing maximum information of speaker characteristics using short speech segments. Then the acoustic-phonetic information based strategies and techniques can be applied on top of the extracted features before modeling to have a better matching under limited data scenario. Finally, the pattern recognition and normalization techniques showing significance for the limited data can be integrated at the back-end of the system to complete the overall framework. The block diagram of such a system utilizing the three different discussed directions is given in Figure 2.1. These explorations are expected to lead to the design of SV systems under limited data with an improved performance.

A few other issues may arise in building a practical SV system in limited data scenario. One such issue is session variability, which occurs when there is a large gap between the recording sessions of the speakers. Some commonly used methods for compensating session variability are JFA, LDA and NAP [31, 33, 35]. A comparison of different such methods has been made by the authors in [103]. A maximum-likelihood linear regression (MLLR) adaptation based method has been proposed in [104] to normalize session variability that helps in improving performance of SV systems.

Another issue for an application oriented SV system is related to identifying the reality of the users. For having an access to the intended service, the attackers of a system may use multiple ways such as the replay attacks, mimicking by the professionals, etc. The replay attacks being the most common among these. In this kind of attacks, the unauthorized users try to deceive the SV system with the use of recorded speech samples of the claimed speaker. The authors of [105,106] have proposed a novel method based on noise and reverberation levels for prevention of these kinds of attacks. In [107], a

study is conducted for replay attacks and then preventing them by the evidence of spectral bitmap based method. With the importance for detection and prevention of unauthorized users, the authors of [108] have developed a database for different studies related to spoofing.

There are some more explorations with reference to the limited data based SV by different researchers, where they adopt different strategies for achieving an improved SV system. The authors in [109] proposed a robust speech activity detection (SAD) for text-independent SV using short utterances, which involves a two way approach. A tri-Gaussian model is fitted to the energy component of a speech sample and then decision is taken for detecting it as a speech or non-speech region based on a threshold. The threshold is calculated by two different methodologies, which are weight-based SAD (wSAD) and mean-based SAD (mSAD). The performance obtained by this SAD method is found to be better than that obtained by the existing approaches. Thus it projects its suitability for text-independent SV using short utterances.

The work of [110] describes a method to have an improved SV performance under a limited data scenario with addition of controlled noise in the speech signal. It has been shown that due to the addition of controlled amount of noise, the distribution of features in the feature space changes. This results in a poorer performance under noisy conditions than that of the baseline case for limited data scenario. However, when the noisy speech features and the baseline features are considered as different versions of given speech data, their combination has been shown to give a better performance. This has been demonstrated using a GMM-UBM based SV framework using limited data utterances from the TIMIT database [111].

The issues discussed in this section may be helpful in building the field deployable systems for SV with limited data. The three different directions identified earlier can be considered along with the directions presented here.

## 2.6 Scope of the current work

This section discusses the scope of the work considered for addressing the problem statement of the thesis. The explorations in the domain of speaker verification have shown possibilities towards field deployable systems. However, minimal speech data is expected for verification of a trial from the perspective of practical systems. This acts as a hurdle in having a high performance compared to that achieved for sufficient data condition. Therefore, there is a definite need for putting effort

towards exploring different works for handling the SV under limited data based scenario. With this motivation, different useful directions are then put forward to utilize them into the SV framework under limited data. Three broad directions are highlighted by categorizing them with respect to their common ground. It is expected that an SV system with an improved performance can be achieved when the three directions together contribute towards having a common framework in case of a limited data scenario.

The directions reviewed for SV using limited data involving short utterances are taken as basic pillars for having an improved SV system. Before working along the identified directions, the different cases of limited data based framework are investigated in Chapter 3. Then sufficient train with limited test data case is put forward as the favorable one based on experimental studies and from the view of thesis objective. The scope of the thesis is as follows:

(i) Based on the review made in the current chapter it is observed that consideration of alternative/different features may be useful for limited data scenario. When we go for features having complementary information from the conventional vocal tract features, the voice source features come into the picture. They are found to posses different information from that carried by the vocal tract features. Further, their effectiveness is more for limited data as mentioned in the literature. In this regard, voice source features are considered as the first direction to explore from the view of thesis objective. Three different source features MPDSS, RMFCC and DCTILPR are considered and then investigated for sufficient train and limited test data based SV. These are expected to contribute towards improving the SV performance when used together. The work with respect to the source features is included in Chapter 3.

(ii) The acoustic-phonetic information is found to be a prospective direction for improving the performance of SV in limited data scenario as reviewed. This can be done either by introduction of speaker-specific acoustic-phonetic knowledge or compensating the phonetic information across speakers. It has been found that putting some degree of constraint on the lexical content can also be useful. In this regard, a text-constrained model based framework is proposed, which exploits explicit match of the content of train and test sessions by putting a constraint to use speaker-specific text. Additionally, this framework is extended to fit into a sufficient train and limited test data based SV. The work along this direction later focuses on utilizing the phonetic information in an implicit manner. The vocal tract constriction (VTC) feature that captures

the level of constriction of the vocal tract while producing different sound units is used as an attribute to capture speaker-specific phonetic information. It is hypothesized that the VTC feature in addition to the conventional vocal tract features can be useful for SV with limited test data. These studies with respect to the text-constrained models and VTC information for speaker modeling are discussed in Chapter 4.

(iii) The third direction refers to suitable pattern recognition approaches and normalization techniques in the back-end of SV systems. It has been observed that the performance of SV systems degrades in limited data condition as there are different variabilities involved in such a scenario. In such conditions, the i-vector representations of the utterances from different speakers become less discriminative. The literature discussed in this current chapter highlights various ways to compensate the same in terms of using robust pattern recognition approaches and normalization techniques. In this regard, a compensation technique that can handle the non-linearities in the data can be useful. Kernel discriminant analysis (KDA) uses non-linear mapping for transforming data points into a higher dimensional space, where the classes become more separable. Hence, it is expected to work well for limited test data based SV and studies related to the same are discussed in Chapter 5. Further, this KDA based work along with the other two directions considered in Chapter 3 and Chapter 4 are likely to provide an improved SV system when put into a common platform.

(iv) SV with limited test data based framework is explored with the motivation of having a field deployable system. There comes a lot of issues while going for a system in real-world application scenario as reviewed. In this direction, some issues are considered for investigation. These issues include mismatch in speech tempo between train and test sessions, session variability and template aging. The mentioned issues are expected to have a definite impact on SV performance for limited data condition. Therefore, they are explored in such a scenario for better understanding to have some possible remedies. The studies with respect to these issues are mentioned in Chapter 6.

# 3

# Exploring Different Attributes of Source Information

Contents

## Overview

*This chapter initially concentrates on exploring different scenarios for speaker verification (SV) using limited data. It then projects sufficient train with limited test data as a favorable framework for practical systems. An investigation is made towards possible reasons of having poor performance in such a scenario. Few directions are highlighted to handle them based on the literature review. The work then focuses on exploring different attributes of excitation source information as the first step towards improving speaker characterization. Three source features mel power difference of spectrum in subbands (MPDSS), residual mel frequency cepstral coefficient (RMFCC) and discrete cosine transform of integrated linear prediction residual (DCTILPR) are investigated for limited test data involving short utterances based SV framework. They are found to have different attributes of excitation source information. These source features when fused, provides SV performance comparable to that obtained with the conventional vocal tract based features. Further, their combination with the vocal tract features contributes to achieve significant improvement showing importance with reference to SV under limited test data condition.*

## 3.1 Introduction

SV with limited data is projected as a favorable framework from the view of having practical application oriented systems as discussed in previous chapters. The current advancements in the domain of SV showcase i-vector based modeling as the state-of-the-art technique for performing studies. However, as discussed the performance of such systems is found to drop with reduction of speech data. This chapter mainly focuses on improving the efficacy of SV with limited data by exploring alternative information for speaker characterization in terms of voice source features. Before working in this direction, different possible scenarios for SV systems based on the amount of train/test data are investigated. This is required for defining a limited data condition for SV from the perspective of field deployable systems. Then the issues that are present for implementing SV with limited data conditions are investigated and the possible reasons of failure are discussed. With reference to these, a few directions are identified based on the reviews made in the previous chapter. The first direction deals with exploration of alternative/complementary features capturing speaker characteristics. The literature has shown the importance of the voice source features for SV. They contain complementary information than that of the vocal tract features, which is having significance with respect to the

current work [11, 13, 63, 69]. Thus, the voice source features are considered as the first step towards addressing the objective of the thesis. The details of the work related to the voice source features are detailed in this chapter later.

The chapter is organized in the following order. Section 3.2 details about the i-vector based speaker modeling that has emerged as the state-of-the-art method in the current decade. In Section 3.3, the different scenarios possible for SV with reference to limited train/test condition are explored. Section 3.4 investigates regarding the different obstacles present in case of limited data based scenario and possible suggestions to handle them. Section 3.5 explains the importance of the source information and then explores different attributes of excitation source information followed by supporting experimental studies. The chapter is finally concluded in Section 3.6.

## 3.2 i-vector based speaker modeling

The i-vector [32] system has demonstrated the state-of-the-art approach for speaker recognition. Its compact representation, computational efficiency and easy channel/session compensation make it a benchmark approach for the SV task. The significant improvement in performance, achieved through i-vector based system over other conventional SV systems shows the potential for using it for SV under limited data condition. In this section, an overview of i-vector based speaker modeling is presented, which is used in different studies throughout the thesis.

### 3.2.1 Front-end processing

The utterances from train and test set are processed with short term Hamming windowed frame of 20 ms with a frame shift of 10 ms and 39-dimensional (13-base + 13-$\Delta$ + 13-$\Delta\Delta$) mel frequency cepstral coefficient (MFCC) features are extracted for each frame. Energy based voice activity detection (VAD) is performed over the speech signal to identify the speech regions. The features belonging to those regions are retained for further processing. They are then normalized in the cepstral domain by cepstral mean normalization (CMN) followed by cepstral variance normalization (CVN) technique [25].

### 3.2.2 Modeling and decision

In this kind of speaker modeling approach, the supervector of mean vectors of components of Gaussian mixture model (GMM) of each utterance is represented by a low dimensional vector, which is referred to as i-vector [28, 32]. It is done using a transformation matrix (T-matrix), which is trained

using a set of development data. The development data set contains all the variabilities such as channel, session etc. Given the T-matrix $\boldsymbol{T}$ and universal background model (UBM) mean supervector $\boldsymbol{m}$, for an utterance that has GMM mean supervector $\boldsymbol{M}_u$, it can be used to obtain the i-vector $\boldsymbol{w}$ as follows:

$$\boldsymbol{M}_u = \boldsymbol{m} + \boldsymbol{T}\boldsymbol{w} \tag{3.1}$$

Let us consider, a UBM having a weighted sum of $C$ component Gaussian densities as, $\boldsymbol{U} = \{\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c, \eta_c\}$, $c = 1, 2, ..., C$, where $\eta_c$, $\boldsymbol{\mu}_c$ and $\boldsymbol{\Sigma}_c$ are the weight, mean vector and covariance matrix associated with mixture $c$, respectively. Then for a sequence of $L$ speech feature vectors $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_L\}$ of dimension $F$, the $0^{\text{th}}$ order ($N_c$) and the centralized $1^{\text{st}}$ order ($\boldsymbol{F}_c$) Baum-Welch statistics of the speech frames on the $c^{\text{th}}$ component of the UBM are given by,

$$N_c = \sum_{t=1}^{L} P(c|\boldsymbol{x}_t, \boldsymbol{U}) \tag{3.2}$$

$$\boldsymbol{F}_c = \sum_{t=1}^{L} P(c|\boldsymbol{x}_t, \boldsymbol{U})(\boldsymbol{x}_t - \boldsymbol{\mu}_c) \tag{3.3}$$

where, $c = 1, 2, \ldots, C$ 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 vector of UBM component $c$.

The total variability matrix $\boldsymbol{T}$ is learned from Baum-Welch statistics of the large amount of development data, computed using the UBM to capture different variabilities. For a given $\boldsymbol{T}$, the i-vector $\hat{\boldsymbol{w}}$ for an utterance $u$ is estimated as,

$$\hat{\boldsymbol{w}} = (\boldsymbol{I} + \boldsymbol{T}'\boldsymbol{\Sigma}^{-1}\boldsymbol{N}(u)\boldsymbol{T})^{-1}\boldsymbol{T}'\boldsymbol{\Sigma}^{-1}\boldsymbol{F}(u) \tag{3.4}$$

where, $\boldsymbol{N}(u)$ and $\boldsymbol{\Sigma}$ are diagonal matrix of dimension $CF \times CF$, whose diagonal blocks are $\boldsymbol{N}_c\boldsymbol{I}$ and $\boldsymbol{\Sigma}_c$, respectively. $\boldsymbol{F}(u)$ is a supervector of dimension $CF \times 1$ generated by concatenating all $1^{\text{st}}$ order Baum-Welch statistics ($\boldsymbol{F}_c$) for a given utterance $u$.

Linear discriminant analysis (LDA) and within class covariance normalization (WCCN) are applied on i-vectors for channel/session compensation [33, 34]. The LDA projects the feature vectors to a set of orthogonal axes, where the intra-class variance caused by the channel is minimized and inter-class variance is maximized. The projection matrix is composed of the eigen vectors corresponding to the top eigen values of the eigen analysis equation given by,

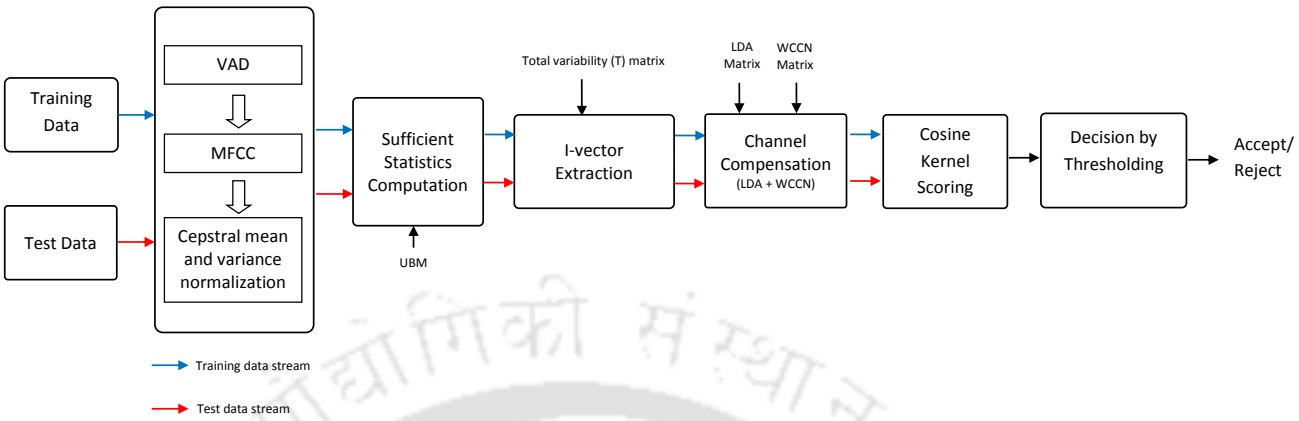


**Figure 3.1:** Block diagram showing different steps involved in the development the i-vector based speaker verification system.

$$(\boldsymbol{W}^{-1}\boldsymbol{B})\boldsymbol{v} = \lambda \boldsymbol{v} \tag{3.5}$$

where, $\boldsymbol{W}$ is the within-class covariance matrix, $\boldsymbol{B}$ is the between-class covariance matrix, $\boldsymbol{v}$ is an arbitrary vector and $\lambda$ is the diagonal matrix of the eigen values.

The WCCN defines a set of upper bounds on the classification error metric to lower the error rate. A transformation matrix is used, by which the feature vectors are transformed to minimize the upper bounds on the classification error metric, which in turn minimizes the classification error. The transformation matrix $\boldsymbol{R}$ is obtained by Cholesky decomposition of the inverse of the within-class covariance matrix $\boldsymbol{W}$ as, $\boldsymbol{W}^{-1} = \boldsymbol{R}\boldsymbol{R}'$. As suggested in [32], when LDA is followed by WCCN better results are obtained, therefore $\boldsymbol{W}$ is calculated in the projected space of LDA.

Figure 3.1 shows the block diagram of i-vector based speaker modeling framework. The train and the test data undergo the same set of procedures as mentioned above to generate train and test i-vectors, respectively. For a pair of train and test i-vectors given by $\hat{\boldsymbol{w}}_{trn}$ and $\hat{\boldsymbol{w}}_{tst}$, the verification of a claim is performed by computing the cosine kernel score between these two i-vectors as follows,

$$\frac{< \hat{\boldsymbol{w}}_{trn}, \hat{\boldsymbol{w}}_{tst} >}{||\hat{\boldsymbol{w}}_{trn}||\ ||\hat{\boldsymbol{w}}_{tst}||} \lessgtr \quad \theta \ \text{(Threshold)} \tag{3.6}$$

## 3.3 Different scenarios for speaker verification with limited data

The SV with limited data can be explored in different ways. The utterances can be of short duration either in train or test or both in train and test sessions. Depending on this, the possible scenarios for SV in a practical setting can be as follows:

- Sufficient train with sufficient test
- Sufficient train with limited test
- Limited train with limited test

The above mentioned three conditions are explored in the i-vector based SV framework to observe the trend for each of them. From the view of field deployable systems, limited amount of speech data is expected. The studies for exploring the limited data conditions are conducted over NIST SRE 2003 database [112]. Although this database is comparatively an older version of SRE database, it is chosen because the main motivation of the considered study is to explore the limited data SV. The recent SRE databases mainly focus on judging the efficacy of SV systems in different adverse conditions, such in presence of noise, language mismatch, sensor/channel mismatch, etc. Thus this does not come within the scope of this work. The NIST SRE 2003 database consists of 356 speakers data containing a population of 212 female and 144 male speakers. There are 2559 test utterances from these speakers in the database that are used for verification against the set of 356 speaker models. The typical duration for enrollment session of the speakers is about 2-3 minutes and the test sessions are of 15-45 s duration. The limited data condition for this work has been fixed as utterances of short duration that is less than or equal to 10 s. This may be seen as truncated utterances of 10 s, 5 s, 3 s and 2 s, respectively. Switchboard Corpus-2 database is used as the development database for this study.

### 3.3.1 Baseline experimental setup

In this work, NIST SRE 2003 database is used for evaluating the performance of SV system. Further, Switchboard Corpus-2 is used as a development database for learning the background models involved. The features extracted from the development data are taken for building a 1024 component UBM. The train and the test features are then processed to extract the sufficient statistics (zeroth and first order statistics) using the UBM parameters. The sufficient statistics of the development data are also extracted using which a T-matrix of 400 columns is trained. This is followed by estimation of respective i-vectors with respect to the extracted sufficient statistics over i-vector based modeling approach. The LDA (150-dimensional) and WCCN (full rank) matrices are computed using the development data i-vectors for the compensation of channel/session information. The trials are then made as per the evaluation plan of NIST SRE 2003 database. The performance of the same is reported in Table 3.1 in terms of equal error rate (EER) and detection cost function (DCF) [112].

**Table 3.1:** Performance of the baseline framework for the sufficient train and sufficient test condition.

| EER (%) | DCF |
|---|---|
| 2.48 | 0.0474 |

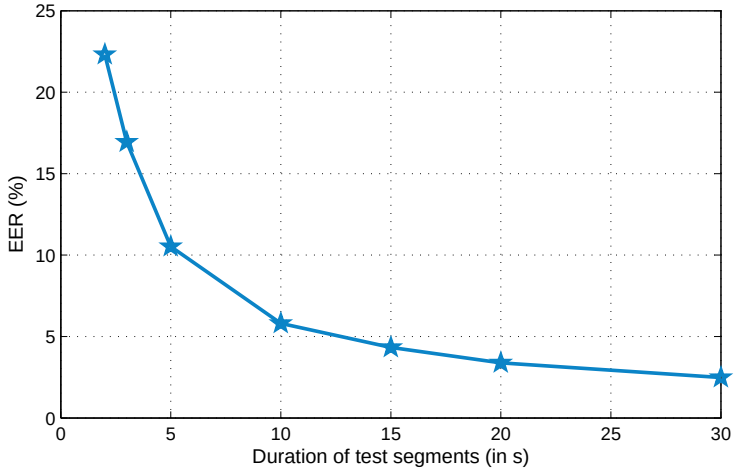


**Figure 3.2:** EER trend with respect to the duration of test segments.

**Table 3.2:** Performance of the baseline framework for sufficient train and limited test condition.

| Test Duration | EER (%) | DCF |
|---|---|---|
| 10 s | 5.81 | 0.1090 |
| 5 s | 10.52 | 0.1977 |
| 3 s | 16.94 | 0.3100 |
| 2 s | 22.31 | 0.4128 |

### 3.3.2 Studies on limited data SV

For investigating the limited data SV, short speech segments are considered for the study. The test utterances are truncated as limited test data is expected from the view of a field deployable system. Then the performance is evaluated for different durations of test segments. The trend of performance in terms of EER for different durations of test segments is shown in Figure 3.2. It depicts that the drop in performance is more prominent when short segments of speech less than 10 s are used for testing. Thus the cases of short test segments below 10 s gain attention, which are also favorable from user comfort and decision on the fly. The EER and DCF values for different duration short test segments in the range 10 s to 2 s are tabulated in Table 3.2. It is observed that the EER increases significantly with a decrease in test duration, which becomes a concern for the realizable systems in

**Table 3.3:** Performance of the baseline framework for limited train and limited test condition.

| Train-Test Duration | EER (%) | DCF |
|---|---|---|
| 10 s | 12.24 | 0.2300 |
| 5 s | 21.68 | 0.4079 |
| 3 s | 29.58 | 0.5591 |
| 2 s | 36.27 | 0.6827 |

field settings. Additionally, a framework of the limited train and test data is also explored for the considered duration cases and their performance is reported in Table 3.3. It indicates a large gap in performance from the former baseline studies as the models are trained poorly due to limited train condition. This indicates that the scenario of the sufficient train with limited test data is a potential framework for practical systems. However, substantial work is needed on different modules of an SV system to overcome the issues that come across while dealing with the limited test data scenario.

## 3.4 Issues in dealing with limited data

The previous section brought out the shortcomings of the limited test data based SV for application services as the performance drops significantly with reduction in the amount of data. In this section, the issues related to the limited data are analyzed in detail at different levels and possible directions for handling them are put forward.

The main issues in handling limited data for SV are as follows:

- Low coverage of acoustic space for a speaker, that affects extraction of speaker characteristics for modeling.
- As the available data is very small, the adaptation process is not much effective and results in poor speaker modeling.

To analyze these issues we plan to have investigations at different levels. The first exploration is based on the amount of speech data available for test utterances after performing VAD. Figure 3.3 represents the histogram for the number of speech frames after VAD for the test set of NIST SRE 2003 database. It is seen that around 15-20 s of speech is available from most of the test segments for making a claim. Similarly, Figure 3.4 shows the histograms for different short speech segments of durations ranging from 2 s to 10 s. It can be seen that for 10 s of truncated test utterances only about 5 s of speech is available for most of the trials after VAD, which is about 1 s speech for 2 s of

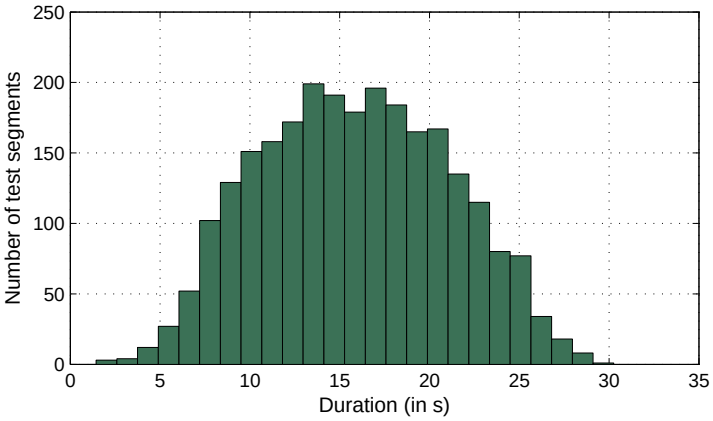


**Figure 3.3:** Distribution of speech data available after voice activity detection for test segments.

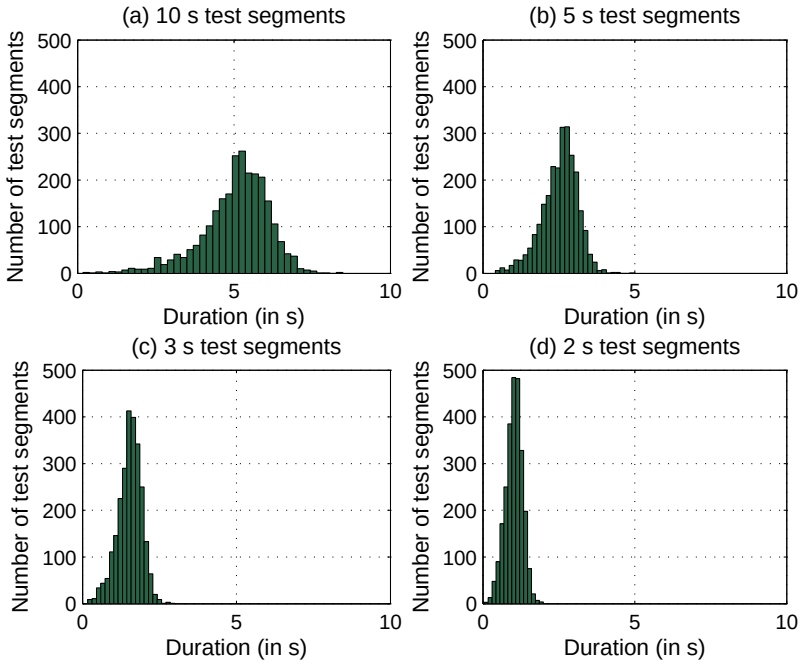


**Figure 3.4:** Distributions of speech data available after voice activity detection for test segments of different durations.

test segments. Thus the amount of speech data available is even very less compared to the case of sufficient test data. This in turn affects the performance due to poor characterization of the speakers.

To visualize the effect of limited data, the i-vectors of sufficient train utterances and sufficient test utterances from three different speakers in NIST SRE 2003 database are considered. The corresponding i-vectors of truncated utterances of 2 s duration are also considered. Figure 3.5 shows that the separation among the three speakers is more distinct, when the i-vectors are estimated with sufficient data for both train and test. On the contrary, they are quite overlapping for both train and test data

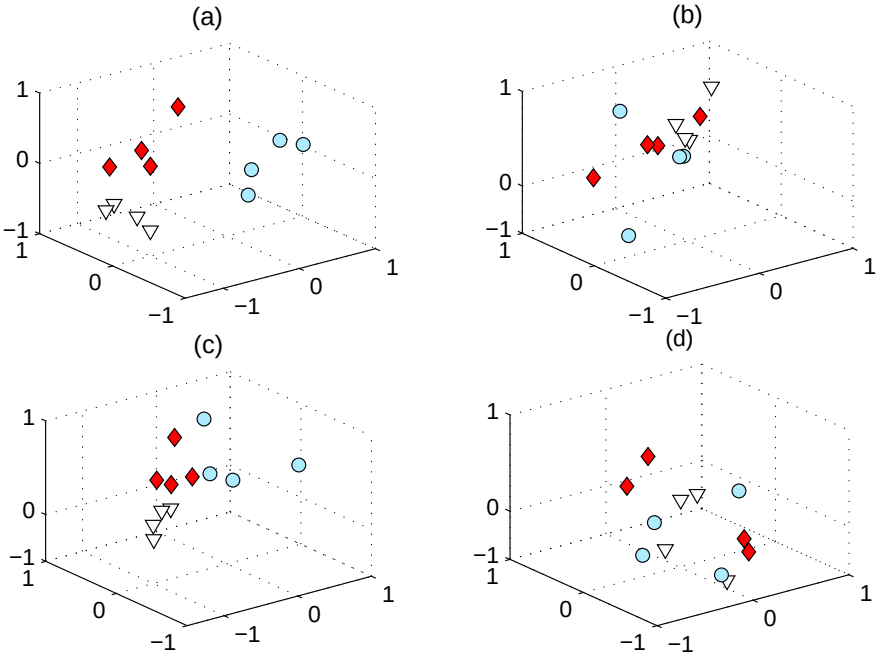


**Figure 3.5:** 3-D plots of i-vectors considering the top three dimensions obtained by performing PCA for three different speakers for (a) sufficient train data (b) limited train data of 2 s (c) sufficient test data (d) limited test data of 2 s. The three different shapes denote three different speakers.

of 2 s case in the 3-dimensional (3-D) space. They are plotted by performing principal component analysis (PCA) on the i-vectors to consider the top three dimensions. Thus when there is sufficient train and test data, it provides a better performance than using limited data. This highlights the poor estimation of i-vectors for the short segments of speech causing performance to drop drastically. Further, the observations are extended to the score level to view the trend of the genuine and impostor scores as the limited test data is used. Figure 3.6 illustrates the separation of genuine and impostor trial scores for the three different conditions. These conditions are sufficient train with sufficient test, sufficient train with limited test and limited train with limited test. The figure shows an interesting trend that the genuine scores are much affected when the truncated speech segments are used. On the other hand, the impostor scores are less affected. This is due to the reason that the i-vectors extracted from limited test data generate less similarity score than the i-vectors extracted from sufficient test data when tested against sufficient train data based i-vectors. Additionally, the similarity score is even lesser while using limited train data based i-vectors for genuine trials. On the contrary, the impostor trials are not much affected as their possibility of producing a higher similarity score further reduces due to consideration of limited data. The goal of this work will be to improve the performance by

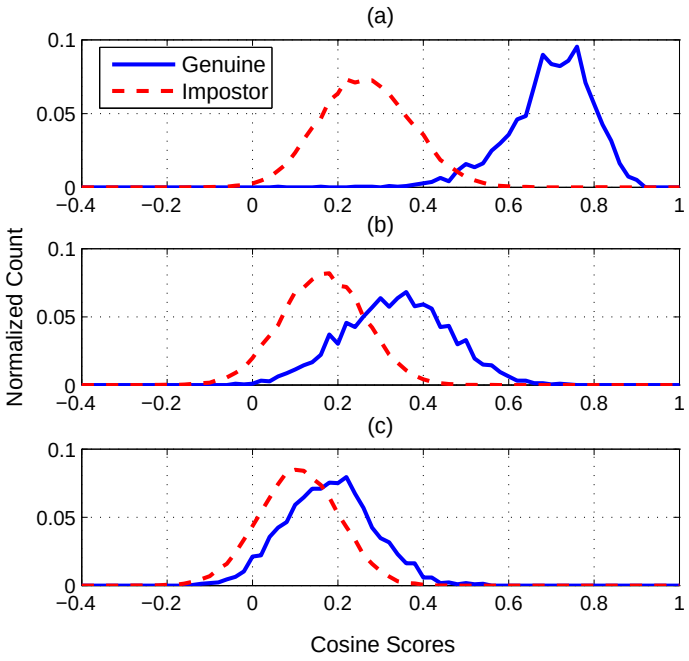


**Figure 3.6:** Distributions of genuine and impostor scores for (a) sufficient train with sufficient test (b) sufficient train with limited test data of 2 s (c) limited train data of 2 s with limited test data of 2 s.

enhancing the separation between genuine and impostor scores for sufficient train with limited test data condition. In this direction, different attempts may be made for overcoming the shortcomings for handling limited data. Some of the possible directions are as follows:

- Use of alternative/complementary features:

  Conventionally MFCC based vocal tract features are used for speaker modeling. However, while dealing with limited data the feature vectors belonging to a particular utterance are small in number. Thus it would be preferable if alternative features based on excitation source information or any other feature having additional information is extracted to capture the speaker characteristics. These alternative/complementary features may be used in fusion to the conventional vocal tract features for improved speaker characterization.

- Utilization of acoustic-phonetic information:

  The speaker-specific acoustic-phonetic information can be useful for providing additional information about the speakers.

- Use of proper channel/session compensation techniques along with pattern recognition approaches that have more efficacy in handling limited data may be helpful.

## 3.5 Source features for speaker verification

The significant performance of the i-vector based system for the sufficient train and sufficient test data shows the potential for using it in SV under limited data condition [32]. In [54], i-vector based SV system for short utterances is analyzed for the train as well as test segments of different durations. From a practical system point of view, we consider the framework of sufficient training data and limited test data to be the favorable framework as discussed earlier in this chapter. The analysis given in [54] for very small amount of test data ($\leq$10 s) shows that the performance drops significantly even though sufficient speech data is used during training. The voice source features represent speaker characteristics in terms of modeling the glottal excitation signal generated from the voiced sound units. This information may be in terms of glottal pulse shape, fundamental frequency and many such aspects of excitation source. Since the glottis and associated muscle structures are unique for each individual, the information represented by the source features is expected to be specific for each speaker and can be utilized for SV. The literature shows that, though the voice source features are not as discriminative as the vocal tract features, the fusion of the two can improve the accuracy [13,63]. Further, the studies of [11, 69] suggest that the amount of train/test data for the voice source features can be less than that for the vocal tract features. This is because the voice source features do not depend much on the phonetic content. On the other hand, the robustness of vocal tract features depends on the amount of phonetic content that they capture. This motivates us to use the voice source features along with the conventional vocal tract features for limited test data based SV framework.

The linear prediction (LP) residual of speech that contains the excitation source information does not contain second order relations as they are already extracted by LP analysis [11, 113]. Therefore, when compact representation of a source feature is obtained by some signal processing method, it may not capture all the aspects of source. Moreover, the noise like structure of the LP residual itself creates difficulty in compact representation of the source information. This signifies the need for an alternative approach for source modeling. One such direction is extracting different attributes of source information and using them to build a better speaker model.

This work focuses on considering different types of source features along with the conventional vocal tract features for improving the SV system performance for limited data test conditions. Three types of source features considered are MPDSS, RMFCC and DCTILPR. The different attributes of excitation source are explored by analyzing these three source features. Further, their effectiveness

in representing the source characteristics when used in combination is investigated. The RMFCC feature is obtained by frame based processing of LP residual and provides a compact representation of source information using cepstral analysis. Alternatively, DCTILPR feature is obtained by pitch synchronous analysis, which provides a compact representation of source information using DCT. Even though the LP residual is common for extraction of both of these features, the signal processing approaches employed in them cases are different. Accordingly, the source information represented by each of the source features may also be different. The source features explored in this work have been previously used for sufficient data condition and found to be carrying speaker-specific information. However, the novelty of this work lies in exploring their importance for limited test data condition and extracting different attributes of source information carried by each of them.

### 3.5.1 MPDSS features

Speaker-specific vocal tract information is modeled from the spectral modulation of the speech signal. On the other hand, the speaker-specific excitation source information can be modeled from the LP residual spectrum [11]. Here, we elaborate the modeling of speaker-specific excitation information from the LP residual spectral harmonics. The dynamic range of a spectrum is defined as the difference between peak and dip within a frequency range that represents the periodicity nature [114]. Depending on the type of voice (say, hard or soft), the spectral flatness of the LP spectrum also changes. If the voice is hard, then it results in rapid and complete closure of the vocal folds. Accordingly, the flow is discontinuous and excitation is more impulse-like. This results in high spectral flatness, equivalently short dynamic range or less periodic nature. Similarly, larger the periodicity, larger will be the difference between peaks and dips of a spectrum [114]. We can observe it better from the power spectrum of a signal. The LP residual signals $r(n)$ and their corresponding power spectra $P(k)$ of utterances from two speakers are shown in Figure 3.7 (a)-(b) and Figure 3.7 (c)-(d), respectively.

The LP residual signal of Speaker-2 shows much stronger periodicity than that of Speaker-1, which has relatively larger dynamic range. The dynamic range of LP residual power spectrum around 500 Hz is 20 dB for Speaker-2 and 5 dB for Speaker-1. However, the overall dynamic range is 30 dB for both the speakers. The excitation periodicity nature is unique for a speaker. So, we explored the spectral harmonics of LP residual for modeling speaker-specific excitation information. Earlier attempt [114] used power differences of subband spectra (PDSS). These PDSS values are computed from the spectral flatness measure of the subbands [115].

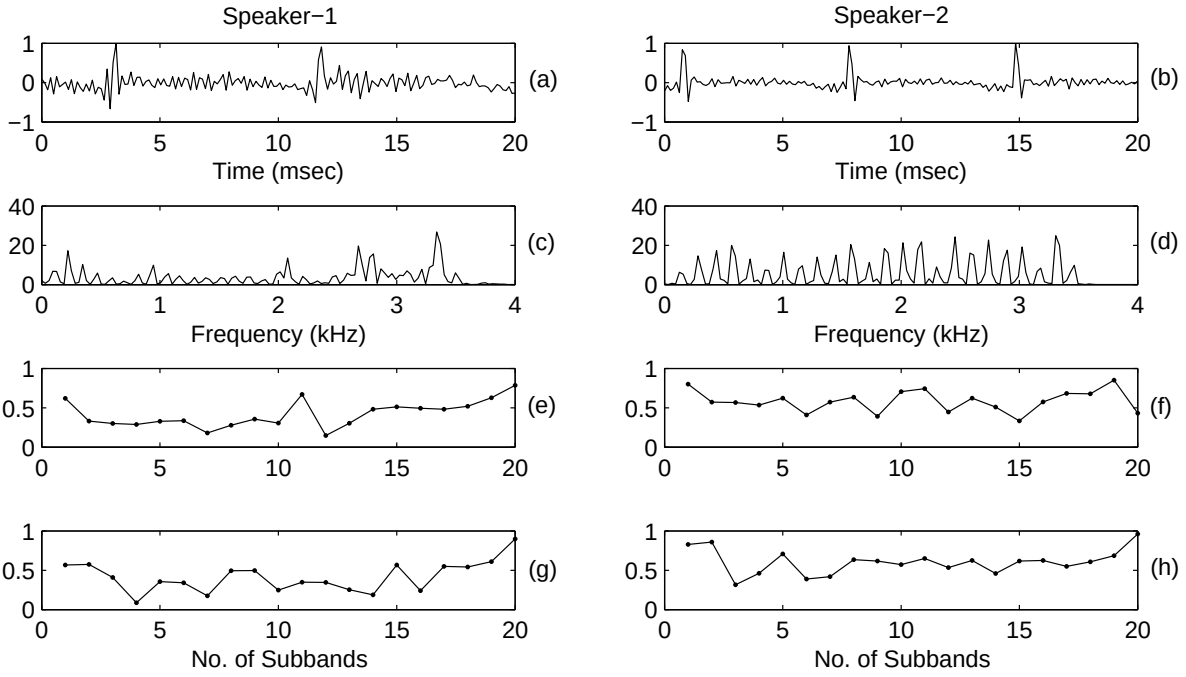


**Figure 3.7:** Speaker-specific excitation information from LP residual spectral harmonics for two speakers. (a)-(b) LP residual signals, (c)-(d) LP residual power spectra, (e)-(f) PDSS features, (g)-(h) MPDSS features.

The PDSS can be interpreted as the subband version of spectral flatness measure. The spectral flatness of a spectrum is defined as the ratio of the geometric mean (GM) to the arithmetic mean (AM) of spectral samples. Since $0 \leq \text{GM} \leq \text{AM}$, the spectral flatness values vary from 0 to 1 and hence the PDSS values also vary from 0 to 1. If the power spectrum has short dynamic range, for example, nearly flat, then $\text{GM} \simeq \text{AM}$ and PDSS is less than 1. Alternatively, if PDSS is low, the spectrum is less periodic. When the spectrum has peaks and dips, for example, the dynamic range is high, then GM is less than AM and PDSS value is close to 1. In this case, the spectrum is more periodic, which makes PDSS measure to provide information about the periodicity nature of a spectrum.

As the PDSS values are computed for subbands, they are expected to be more effective than a global measure of spectral flatness in capturing the speaker-specific characteristics. However, these PDSS values are computed from linearly spaced overlapping filters. We may be benefited by using non-linearly spaced subbands. The nonlinearity nature of the human auditory perception system (like mel bands) is expected to be effective in deciding the subband spacings. The other motivation is from the property of the mel filterbank that provides fewer spectral samples to lower bands and more to higher bands (beyond 1 kHz). The dominant speaker information in the source is manifested in

the higher frequency range. Since PDSS is a statistical measure, with an increase in the number of samples, PDSS may be more accurately measured in the higher frequency range. The MPDSS ($M_p(m)$) feature is computed as follows:

$$M_p(m) = 1 - \frac{\left[\prod_{k=l_m}^{h_m} P(k)\right]^{\frac{1}{N_m}}}{\frac{1}{N_m}\sum_{k=l_m}^{h_m} P(k)} \tag{3.7}$$

where, $N_m = h_m - l_m + 1$ is the total number of samples; $l_m$ and $h_m$ denotes the first and last sample of the subband in the $m^{\text{th}}$ filter. The $P(k)$ corresponds to the power of the $k^{\text{th}}$ sample of the subband.

In this case, 20 mel filterbanks are used that forms a 20-dimensional MPDSS feature vector for each frame. The use of twenty overlapping filters is followed from [114]. The PDSS and MPDSS features for utterances from Speaker-1 and Speaker-2 are shown in Figure 3.7 (e)-(f) and Figure 3.7 (g)-(h), respectively. The PDSS values for Speaker-1 are smaller than that for Speaker-2. This depicts that Speaker-2 shows much stronger periodicity than Speaker-1. Additionally, the MPDSS values of Speaker-1 are less than Speaker-2. This shows that the PDSS property holds well for MPDSS also, even if it is computed from mel warped spectrum. Further, it is observed that MPDSS values can more accurately distinguish the periodicity nature of the speakers. For example, in case of Speaker-1, PDSS values from subbands 3, 4 and 5 indicate that they are almost same. But this is not the case as observed from the trend of MPDSS values. The MPDSS value of subband 4 is comparatively less. This indicates that perceptually, the periodicity of the Speaker-1 in the fourth subband is weak. Similar observation can also be made for Speaker-2 in subbands 2, 3 and 4 showing better capture of periodicity information by MPDSS. Thus the MPDSS feature is an attempt to capture the periodicity aspect of excitation source information by using mel filterbanks over subbands of power differences in the spectrum.

### 3.5.2 RMFCC features

Cepstral analysis of LP residual is found to be one of the suitable approaches due to its simplicity [116]. The approach involving cepstral analysis has scope for improvement by use of spectral subband energies. The spectrum of LP residual is flat in nature. Therefore, if the spectral energies are accumulated over the subbands, the benefit of using them as features can be achieved. With this motivation, the authors of [66] proposed the source feature RMFCC, which involves processing the

LP residual in the cepstral domain unlike the former feature MPDSS. The log magnitude spectrum of LP residual is passed through a non-uniform filterbank with triangular windows placed on the mel frequency scale. Then inverse discrete Fourier transform (IDFT) is computed over it to obtain RMFCC feature [66]. Let, $r(n)$ be the LP residual of a speech segment and $R(w)$ its spectrum, the log magnitude of which is passed via mel filterbanks $M_l$ for non-linear transformation. Then RMFCC feature ($\boldsymbol{R_m}(k)$) is computed in the following way,

$$\boldsymbol{R_m}(k) = \text{IDFT}[M_l(\log|R(w)|)] \tag{3.8}$$

The first 13 dimensions along with 13-$\Delta$ and 13-$\Delta\Delta$ are considered to form a 39-dimensional RMFCC feature vector for short term processed speech signal with frames of 20 ms with a shift of 10 ms. Accordingly, RMFCC represents segmental level smoothed spectrum information due to mel filterbanks. This in turn may correspond to the average glottal signal information.

### 3.5.3 DCTILPR features

The DCTILPR source features are obtained by using the integrated linear prediction residual (ILPR) that closely resembles the glottal flow derivative signal [67]. The ILPR is used as a voice source estimate and its pitch synchronous discrete cosine transform (DCT) coefficients are used to form the feature vector [117]. The DCTILPR has been shown to perform on par with existing voice source-based speaker-specific features in a speaker identification task [67].

Figure 3.8 shows the block diagram to extract the DCTILPR features. The energy-based VAD is applied on the speech signal to get frames with significant voice activity. On these frames, an epoch extraction algorithm [117] is applied. Then using these epochs, a voiced/unvoiced (V/UV) decision based on maximum normalized cross-correlation is performed as in [118]. Only the voiced regions are retained for further processing and the ILPR is extracted on the voiced regions as in [117]. Considering the epochs in the voiced regions as glottal closure instants (GCIs) and the interval between two successive GCIs as a pitch period, pitch synchronous DCT-II having compaction property is obtained to obtain the DCTILPR features.

Let, $i_r(n)$ be the ILPR corresponding to the LP residual $r(n)$ extracted between epoch locations $j$ and $(j+1)$ of a speech segment. The respective DCTILPR feature ($\boldsymbol{D_r}(k)$) taken in pitch synchronous manner is given by,

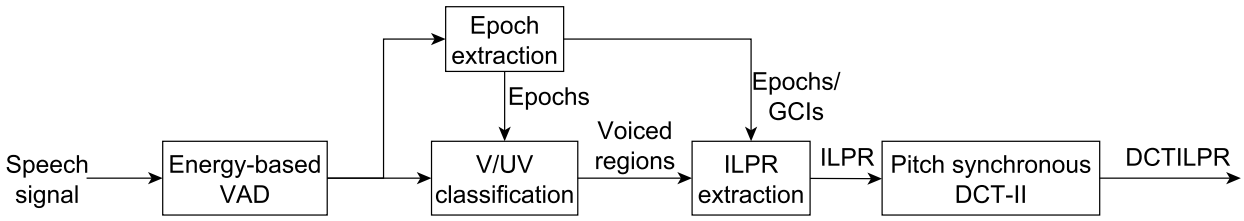


**Figure 3.8:** Block diagram of the system for the extraction of DCTILPR features.

$$\boldsymbol{D_r}(k) = \sum_{n=0}^{N-1} i_r(n) \cos\left[\frac{\pi}{N}(n+\frac{1}{2})k\right] \tag{3.9}$$

where, $N$ is the number of samples between the epoch locations $j$ and $(j+1)$ and $k = 0, 1, 2, \ldots, N-1$. As shown in [67], the first 24 DCT coefficients are sufficient to capture the speaker information contained in the voice source and are taken as the feature vector. As the source feature DCTILPR captures the glottal signal shape information, pitch synchronous analysis is made for precisely capturing this aspect of source characteristics.

### 3.5.4 Different attributes of source

Figure 3.9 shows the nature of LP residual, ILPR and the three source features for utterances of the vowel /a/ from the word "dark" for two different speakers in TIMIT database [111]. The features MPDSS, RMFCC and DCTILPR involve spectral, cepstral and temporal domain processing of LP residual, respectively. The MPDSS and RMFCC features involve segmental processing on the speech signal. Thus they model the excitation source information averaged over 2-3 pitch periods. On the contrary, DCTILPR feature is extracted by pitch synchronous analysis. Hence it models the source information within a pitch period representing the shape of glottal signal. Due to different domains of processing, different equations for extraction and segmental vs. pitch-synchronous ways of extracting information, each of these features is hypothesized to capture different attributes of excitation source.

The DCTILPR feature captures the glottal shape information of a speaker in a pitch synchronous manner [67]. However, this does not capture the periodicity information of the signal denoting how much periodic it is. The MPDSS feature is a variant of spectral flatness measure. It captures the periodicity information as the peak to dip ratio of the spectrum of a signal measures the periodicity [65]. Thus the periodicity attribute captured by the MPDSS is an additive information for representing

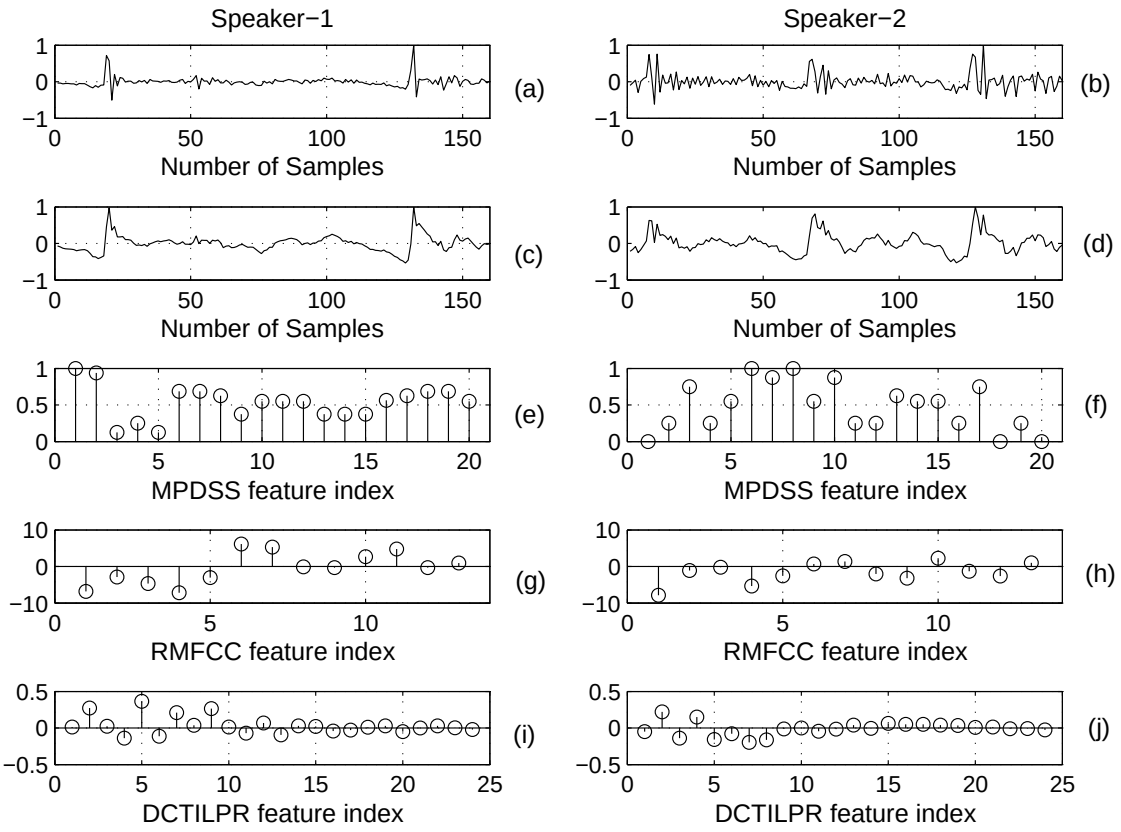


**Figure 3.9:** Three types of source features for the utterances of the vowel /a/ in the word "dark" for two speakers in TIMIT database. (a)-(b) LP residual signals, (c)-(d) ILPR signals, (e)-(f) MPDSS features, (g)-(h) RMFCC features, (i)-(j) DCTILPR features.

the source characteristics. Further, the RMFCC feature is extracted from the LP residual by its cepstral analysis. Due to the noise like structure of LP residual, the information captured without any processing is less distinctive across the speakers. The RMFCC feature computed from the subband energies, provides a segmental level smoothed spectrum information by capturing the strength of excitation [66]. Therefore, a combination of the different attributes of these source features is expected to enhance SV performance.

### 3.5.5 Experimental results and analysis

The three source features MPDSS, RMFCC and DCTILPR are extracted for the NIST SRE 2003 database and then three parallel systems are built over the i-vector based speaker modeling framework. Table 3.4 shows the performance of the baseline system using MFCC features along with the three

**Table 3.4:** Performance of different features for sufficient train and limited test data condition.

| Test Duration | MFCC | | MPDSS | | RMFCC | | DCTILPR | |
|---|---|---|---|---|---|---|---|---|
| | EER (%) | DCF | EER (%) | DCF | EER (%) | DCF | EER (%) | DCF |
| 10 s | 5.81 | 0.1090 | 17.43 | 0.3269 | 12.96 | 0.2466 | 13.91 | 0.2497 |
| 5 s | 10.52 | 0.1977 | 22.58 | 0.4250 | 18.88 | 0.3462 | 18.65 | 0.3460 |
| 3 s | 16.94 | 0.3100 | 27.60 | 0.5202 | 23.62 | 0.4362 | 22.13 | 0.4077 |
| 2 s | 22.31 | 0.4128 | 31.44 | 0.5958 | 27.55 | 0.5203 | 27.78 | 0.5198 |

systems using the source features for limited test data conditions ($\leq$10 s) in terms of EER and DCF. As reported in the literature, the performance of an individual source feature is significantly lower than that of MFCC. However, as the amount of test data decreases, the performance degradation in case of source features is relatively less than that for MFCC. This is because the voice source features are less dependent on the phonetic content. On the contrary, the system features model the speaker characteristics based on more coverage of acoustic space for a speaker, capturing the phonetic variation. Moreover, the MFCC features model the vocal tract characteristics to a large extent, whereas each of the source features model one aspect of the source information. Thus, under limited test data condition, even though a single source feature does not perform better than MFCC, the fusion of multiple source features may be useful. This in turn may provide effective characterization of source information, which can help in improving SV performance.

As demonstrated in earlier studies, the source features perform well in combination with MFCC features, especially for limited test data condition [11, 13]. The study is extended for the RMFCC feature to view the trend. Each of the source features MPDSS, RMFCC and DCTILPR are fused at the score level by the following way,

$$S_c = \alpha S_s + (1 - \alpha) S_m \tag{3.10}$$

where, $S_s$, $S_m$ and $S_c$ denote the scores obtained using particular source feature, MFCC feature and the combination of the two, respectively. The learning of the weights is made on the development set. It is also to be noted that the weights are learned for different durations of test utterance cases. The scores obtained from two features are fused with different weights varying between 0 to 1 in steps of 0.05. Then the optimal weight value $\alpha$ is considered for which the performance in terms of EER is found to be the least. Thus based on the different durations of the test utterances considered for the study, respective weights are used.

**Table 3.5:** Performance under fusion of different source features with MFCC and their comparison to baseline performance suing MFCC showing improvements in each of the combinations for limited test data.

| Test Duration | MFCC | | MPDSS+MFCC | | RMFCC+MFCC | | DCTILPR+MFCC | |
|---|---|---|---|---|---|---|---|---|
| | EER (%) | DCF | EER (%) | DCF | EER (%) | DCF | EER (%) | DCF |
| 10 s | 5.81 | 0.1090 | 5.56 | 0.1048 | 5.78 | 0.1087 | 5.33 | 0.0971 |
| 5 s | 10.52 | 0.1977 | 10.12 | 0.1850 | 9.67 | 0.1829 | 8.45 | 0.1567 |
| 3 s | 16.94 | 0.3100 | 15.04 | 0.2811 | 14.96 | 0.2790 | 12.46 | 0.2325 |
| 2 s | 22.31 | 0.4128 | 19.96 | 0.3720 | 20.19 | 0.3767 | 17.71 | 0.3351 |

**Table 3.6:** Canonical correlation analysis (CCA) measure to highlight the nature of complementary characteristics between different types of features.

| Feature Pairs | Correlation |
|---|---|
| MPDSS vs. MFCC | 0.89 |
| RMFCC vs. MFCC | 0.91 |
| DCTILPR vs. MFCC | 0.79 |
| MPDSS vs. RMFCC | 0.91 |
| MPDSS vs. DCTILPR | 0.81 |
| RMFCC vs. DCTILPR | 0.82 |

**Table 3.7:** Performance under fusion of two source feature pairs and combined fusion of three source features showing different attributes of excitation source information. The boldface numbers showing improved results compared to the baseline.

| Test Duration | DCTILPR+MPDSS | | DCTILPR+RMFCC | | MPDSS+RMFCC | | Source Fusion | |
|---|---|---|---|---|---|---|---|---|
| | EER (%) | DCF | EER (%) | DCF | EER (%) | DCF | EER (%) | DCF |
| 10 s | 10.93 | 0.2052 | 12.33 | 0.2285 | 14.96 | 0.2815 | 10.57 | 0.1964 |
| 5 s | 14.99 | 0.2785 | 13.37 | 0.2492 | 15.58 | 0.2955 | 11.97 | 0.2252 |
| 3 s | 18.29 | 0.3413 | **16.67** | **0.3109** | 20.28 | 0.3802 | **15.85** | **0.2854** |
| 2 s | 23.85 | 0.4456 | **21.59** | **0.4065** | 24.16 | 0.4578 | **20.19** | **0.3759** |

Table 3.5 shows the performance for the fusion of each of the three source features with the MFCC feature. The performance improvement for each case is more apparent as the duration of the test speech segment is reduced. To study the different attributes in terms of correlation measure for each of the source features, canonical correlation analysis (CCA) is performed among the source features and MFCCs. Table 3.6 depicts that there is some complementary nature of information carried by each feature in combination with another (correlation value being lesser than 1). It is observed that the DCTILPR feature shows more complementary information to the other two source features as well as MFCC. The three source features are combined at the score level similar to the case as given by (3.10) considering scores obtained from two features at a time. Table 3.7 shows the results of fusion for different source features that provides improvement when combined with one another depicting

**Table 3.8:** Performance under fusion of two source features with MFCC and their comparison to (Source Fusion+MFCC) indicating better results for fusion of three source attributes when combined to MFCC.

| Test Duration | (DCTILPR+MPDSS) +MFCC | | (DCTILPR+RMFCC) +MFCC | | (MPDSS+RMFCC) +MFCC | | Source Fusion +MFCC | |
|---|---|---|---|---|---|---|---|---|
| | EER (%) | DCF | EER (%) | DCF | EER (%) | DCF | EER (%) | DCF |
| 10 s | 5.24 | 0.0975 | 5.19 | 0.0979 | 5.42 | 0.1008 | **5.10** | **0.0965** |
| 5 s | 8.40 | 0.1531 | 8.36 | 0.1553 | 9.58 | 0.1776 | **8.18** | **0.1524** |
| 3 s | 12.15 | 0.2242 | 11.97 | 0.2215 | 13.96 | 0.2578 | **11.47** | **0.2148** |
| 2 s | 17.12 | 0.3216 | 16.98 | 0.3262 | 18.34 | 0.3466 | **16.08** | **0.3025** |

different nature of source information carried by each of them. The trend shown by CCA is also reflected while performing fusion of different features as can be observed from Table 3.5 and Table 3.7. This is because more the amount of complementary information for each feature pair, it helps for achieving better performance on fusion. Thus each of the source features showed improvement on fusion as a pair of two features, depicting different aspects of source. Then the fusion of all the three source features is carried out by averaging the scores and the performance is reported in the last column of Table 3.7. Due to the combination of the three source features carrying different attributes of source, the performance obtained for 2 s and 3 s cases outperform the MFCC features for respective cases by a larger margin. Further, the fusion of three features results in improved performance than that observed from fusion of only DCTILPR and RMFCC features. This infers that, even though any one source feature does not provide good performance, they combine altogether well to outperform MFCC. Thus source features may be capturing different attributes of excitation source and hence significant improvement is obtained when they are combined.

Finally, the combination of two source features at a time with MFCC is carried out, followed by the fusion of all the three source features with MFCC. Table 3.8 shows the results belonging to the same. In all the cases, significant improvements are observed, which is more prominent when all the three source features carrying different attributes of source are fused with MFCC. It indicates the importance of each source feature for excitation source characterization that helps in improving SV performance. Thus all the source features are found to be necessary for improved speaker characterization. Further, their necessity increases with reduction of test data. Figure 3.10 shows the histograms of the scores obtained from the genuine and impostor trials of NIST SRE 2003 database, for 2 s test data case for different features and their combinations. It indicates more separability of genuine and impostor scores for the fusion of the three source features than the baseline MFCC feature based system.

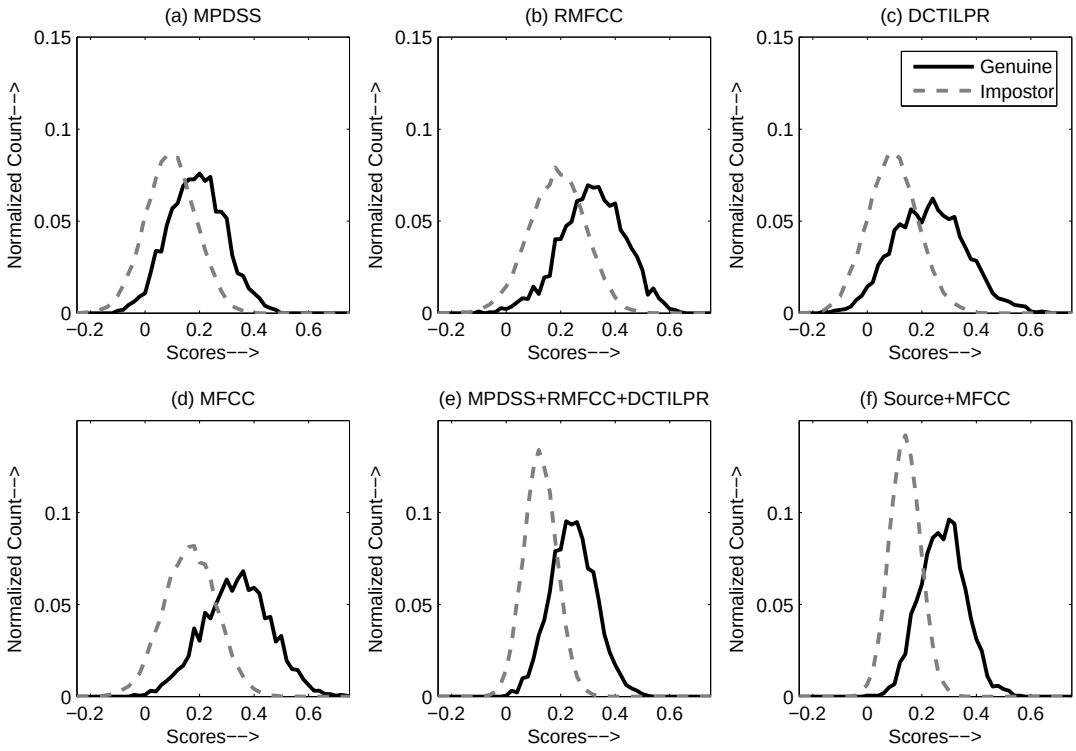


**Figure 3.10:** Histograms of scores for different features and their combinations for 2 s test data.

**Table 3.9:** Area of overlap of genuine and impostor score histograms indicating better separability for three source features fusion and their fusion with MFCC features.

| Feature | Overlap (%) |
|---|---|
| MPDSS | 62.17 |
| RMFCC | 55.20 |
| DCTILPR | 52.94 |
| MFCC | 42.90 |
| Source Fusion | **39.09** |
| Source+MFCC | **30.85** |

To quantify the same the area of overlap (in %) is computed for the genuine and impostor score histograms and is shown in Table 3.9. This depicts that the separability between the histograms enhances on fusion of MFCC with three source features. Figure 3.11 illustrates the detection error tradeoff (DET) curve trends for different features and their combinations for the case of 2 s test data [119]. The combination of three source features gives better performance than the stand-alone vocal tract features. Additionally, the fusion of three source features with the vocal tract information enhances the baseline performance based on MFCC features by a large margin.

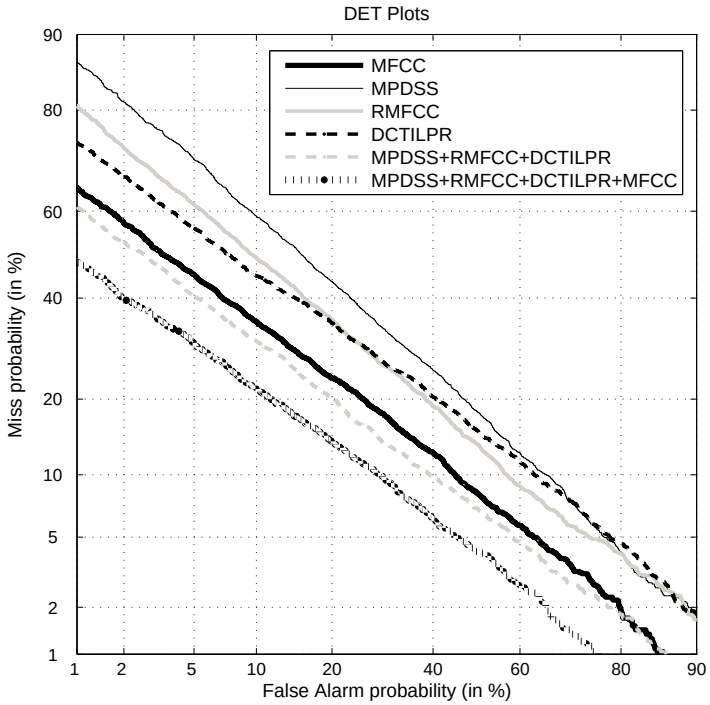


**Figure 3.11:** DET plots obtained using different features and their combinations for 2 s test data.

## 3.6 Summary

This chapter puts the first step into the thesis objective by bringing out a framework of SV with sufficient train and limited test data favorable from the view of field deployable systems. It highlights the obstacles that cause poor performance in such a scenario and discusses their possible solutions. The work then explores different attributes of excitation source information for speaker modeling. The three source features, namely, MPDSS, RMFCC and DCTILPR are found to capture different aspects of source information, which are the periodicity, smoothed spectrum information and shape of glottal signal, respectively. These different attributes are visible on the fusion of multiple features, which becomes more evident with reduction in duration of test segments. The combination of these three features outperforms the MFCC features for very less duration test data cases. Further, the three source features in combination with MFCC features provide better performance than their individual combination due to presence of different attributes of source characteristics. This highlights their importance for SV under limited test data scenario.

# 4

# Text-constrained Speaker Models and Vocal Tract Constriction Information

Contents

### Overview

*This chapter focuses on exploring phonetic information for speaker modeling that has shown potential under limited test data based speaker verification (SV). In this regard, a text-constrained model based framework is proposed that explicitly utilizes lexical match between train and test sessions by putting a constraint on the text to be spoken. This framework is found to work better than the conventional SV systems using limited data involving short utterances. Further, the different attributes of source features explored in the previous chapter are investigated for text-constrained models. The chapter later explores implicit utilization of speaker-specific phonetic content by use of vocal tract constriction (VTC) evidence. The fusion of VTC feature with conventional mel frequency cepstral coefficient (MFCC) features is found to work well signifying the importance of the speaker-specific phonetic information for SV with limited test data.*

## 4.1 Introduction

The i-vector based modeling is considered as the benchmark approach for text-independent SV [32]. However, as discussed in the previous chapter, it fails under short utterance based scenario for SV due to poor capture of speaker characteristics. To improve the performance for SV with limited test data condition, different attributes of source information are explored in the previous chapter. In this chapter, we have explored the speaker-specific phonetic information, which can provide additional information for speaker characterization. The results in [94] show that although the i-vector based modeling performs poorly under limited data, the use of this approach for the text-dependent SV, that constrains the phonetic content may help it. Further, the authors showed that the performance can be enhanced by estimation of within class covariance normalization (WCCN) and eigen factor radial (EFR) using the development data having the same phonetic content. However, the work is conducted on text-dependent SV database and the efficacy of i-vector based modeling for phonetically constrained test data is studied. In [95], the authors showed that developing the phonetically constrained probabilistic linear discriminant analysis (PLDA) model helped in improvement of text-dependent SV performance. In one of our works, Gaussian posteriorgram based transformed features when computed from speaker-specific and sentence-specific Gaussian mixture model (GMM) performed well in case of text-dependent SV [120]. Motivated by these works, we tried to adopt the phonetic constraint based exploration in the context of text-independent SV.

In the current work, initially a text-constrained model based framework is proposed. This is based on explicit utilization of the lexical content between train and test sessions. The significance of this work is demonstrated on data collected over a practical application oriented system under limited train and limited test condition. Additionally, the importance of the text-constrained model based framework has been explored for the recently available RedDots database for sufficient train and limited test data scenario. The Part IV of RedDots database contains two enrollment conditions, which are text-dependent and text-independent. The text-constrained model based framework focuses to reduce the performance difference between text-dependent and text-independent modalities. Further, the source features explored in the previous chapter are investigated in the context of text-constrained models to observe their effectiveness. The chapter later focuses to utilize the speaker-specific phonetic information in an implicit manner by use of the VTC feature. It captures the level of constriction in the vocal tract while producing different sound units. This information is specific for every individual due to the structure of the vocal tract. The fusion of VTC feature is carried out with the conventional MFCC based vocal tract features to have additional information from the speaker-specific phonetic information perspective.

The chapter is organized as follows. Section 4.2 details the proposed framework of the text-constrained model that utilizes the phonetic content in an explicit manner. In Section 4.3, the text-constrained model based framework is explored for sufficient train with limited test condition and then the different aspects of source information are investigated for such a framework. Section 4.4 describes the VTC feature used for capturing the speaker-specific phonetic information in an implicit manner and the related experimental studies for limited test data scenario. Finally, the summary of the chapter is mentioned in Section 4.5.

## 4.2 Exploring text-constrained models for speaker verification

The robustness of text-independent SV is achieved with research advances in many directions, starting from different features to classifiers suitable for it. This section highlights the genesis of text-constrained models for text-independent SV. In a text-constrained model based framework, the speaker models are created with user chosen text of short duration of around 10 s. Then the same phonetically constrained texts are spoken by the users during testing sessions. The performance of text-constrained models is compared to that of text-independent SV without any phonetic constraint.

The proposed framework of text-constrained model is found to work significantly better than the conventional text-independent SV system. Another study is conducted, where the speaker models are created using data that contains constrained text as a subset of a larger train session around 1 minute. This study highlights the importance of the phonetic content match between train and test sessions. The studies are reported over a 100 speaker database. This database is collected over an online telephone based framework using a voice-server having interactive voice response system (IVRS) callflow in view of field deployable systems [50]. The contribution of this work mainly lies in adopting the phonetically constrained models for text-independent SV under limited data condition and highlighting its importance experimentally.

### 4.2.1 Motivation for text-constrained models

The motivation for the proposed text-constrained model is based on the phonetic match that is retained for a speaker across different examples of the same phonetic content. To view the feature distribution based on this direction, two examples each of three different sentences (sen I, sen II and sen III) having distinct phonetic contents from two speakers are considered. A 39-dimensional MFCC feature vector is extracted from every frame for each of the examples. Then a 3-dimensional (3-D) representation is obtained considering the first 3 dimensions for 50 feature vectors for each examples and is shown in Figure 4.1. It can be seen from Figure 4.1 (a)-(b) and Figure 4.1 (c)-(d) that for the same speaker, the feature vectors of a particular sentence category having distinct phonetic content occupy similar regions in the acoustic space scatter plot. On the contrary, that is somewhat different for different speakers. However, as the sentences spoken in the two examples are same, there is some similarity in the scatter plots of features of Figures 4.1 (a)-(d).

Based on this analysis, we came up with the proposal for speaker-specific text-constrained speaker models, as they are expected to provide a better match for similar speakers and vice versa. In this regard, two examples of two different sentences from two speakers are considered. The examples are taken from the set of 100 speakers data that is considered in this work, which is collected over online IVRS callflow based telephone network. They are also viewed in the 3-D space to verify our hypothesis. Figure 4.2 clearly conveys that upon putting a constraint on speaker-specific phonetic content condition, the speakers are more distinguishable. Additionally, the similarity of different examples of the speakers is also better visible. This made us put forward the proposal of the text-constrained model based framework for text-independent SV under limited data condition.

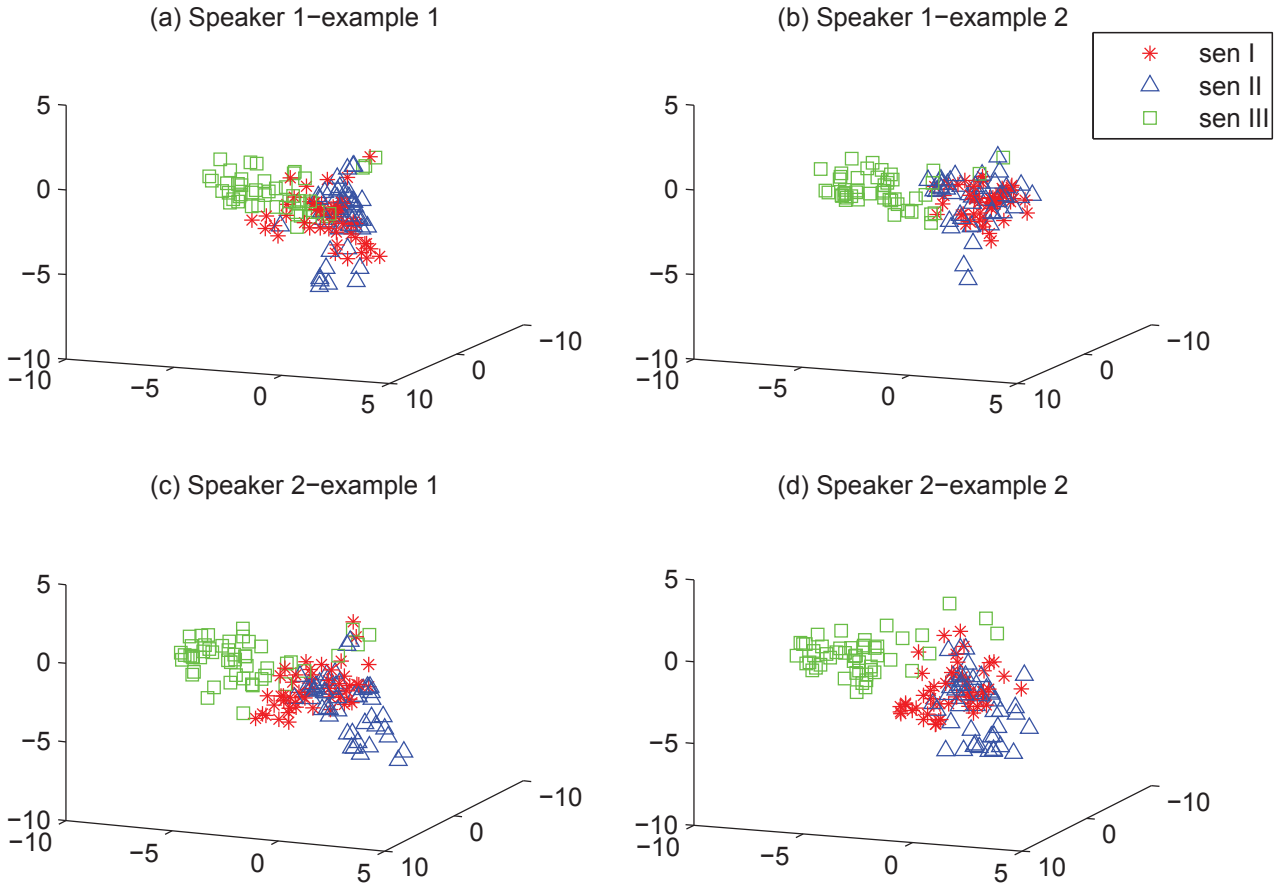


**Figure 4.1:** Features for three different sentences of two examples from two speakers.

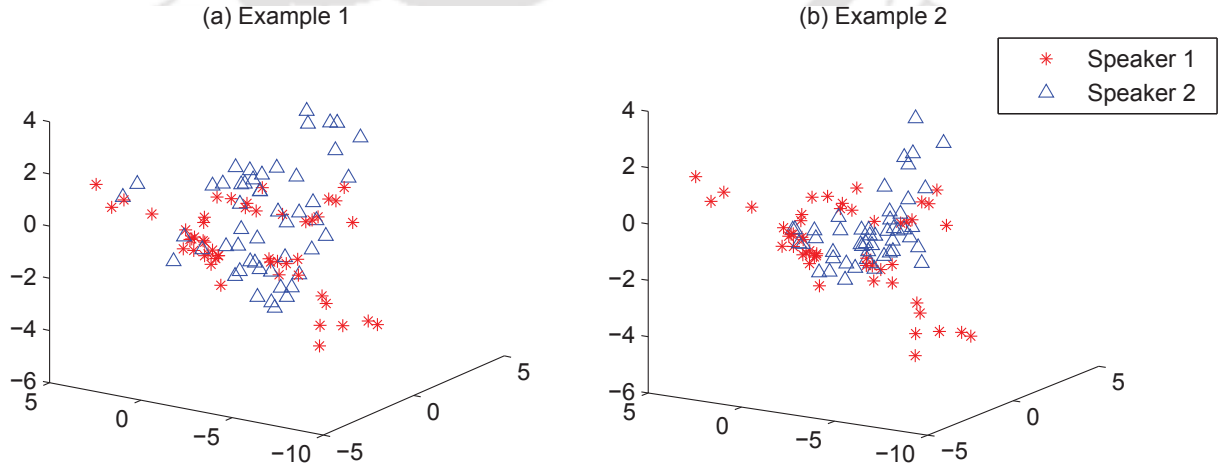


**Figure 4.2:** Features for speaker-specific constrained texts of two examples from two speakers.

### 4.2.2 Proposed framework using text-constrained models

The proposed framework of the text-constrained model under limited data, which is inspired from the phonetic content match is explained in this subsection. It is used as a case of considering constrained data in i-vector based speaker modeling. The studies performed for the same are particularly designed with respect to the practical field deployable systems.

#### 4.2.2.1 Database, preprocessing and feature extraction

As the proposed work is directed towards addressing issues related to deployable systems, the database considered in the current study is taken under such a scenario. A database of 100 students is collected, where the students enroll to the system by an online IVRS callflow over the telephone channel [48]. During training session, the students are asked to speak their roll number and name, three text-constrained sentences and 3 minutes of read text speech. During the testing phase, to reduce the amount of time involved, they are asked only for their roll number and name, then one out of the randomly generated text-constrained sentences. The database contains 10 trials from each speaker taken for this study. The studies of the previous chapter are carried out on NIST SRE 2003 database with limited test data. However, the current study for the proposed framework is carried out on the database explained here. This is because it is structured to fit into this study based on the proposed framework for text-constrained speaker models.

The train and the test speech are considered as frames of 20 ms duration at a 10 ms frame shift. A 39-dimensional (13-base + 13-$\Delta$ + 13-$\Delta\Delta$) MFCC feature vector is extracted for every Hamming windowed frame. Energy based voice activity detection (VAD) is performed to retain the feature vectors from the speech regions. The features are then normalized in the feature domain using cepstral mean and variance normalization (CMVN) [25].

#### 4.2.2.2 Structure of proposed text-constrained model framework

The baseline framework for text-independent SV under the sufficient train with limited test data condition is developed using the discussed database over i-vector based speaker modeling. In this proposed text-constrained model based framework, 3 minutes of read speech is considered for training the speaker models. On the other hand, a user chosen phrase and a text-constrained phrase are used for test sessions. A gender independent universal background model (UBM) of 1024 components is built using an equal amount of female and male speech data of around 10 hours from NIST SRE

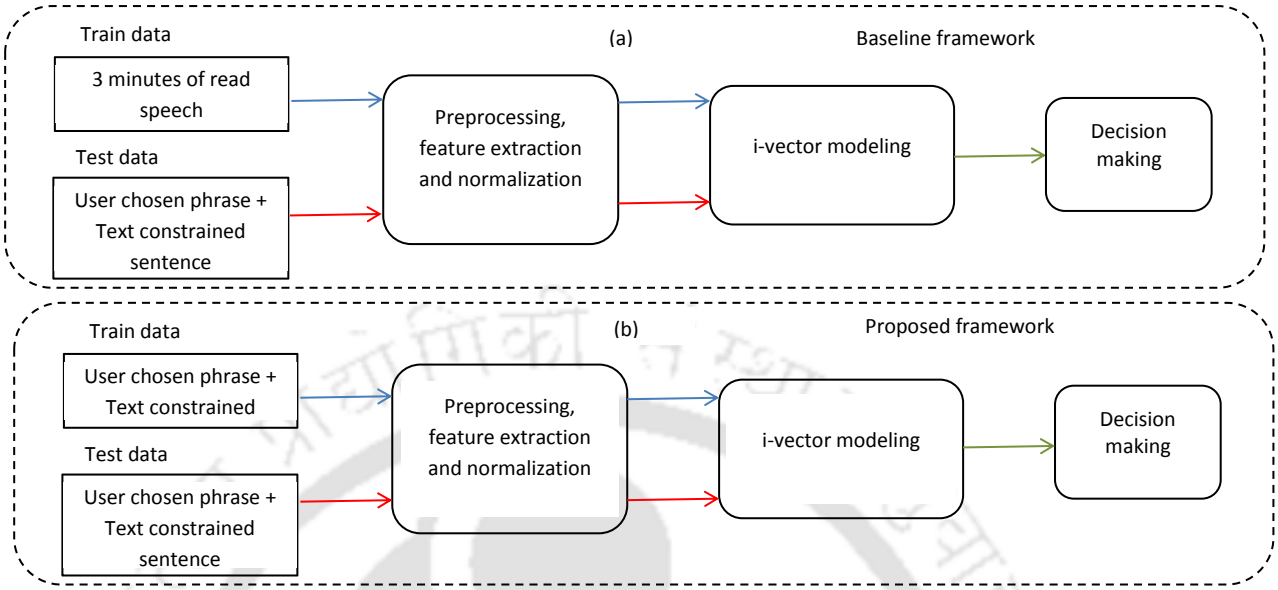


**Figure 4.3:** Block diagram of (a) Baseline framework (b) Proposed framework.

2010 telephone channel data as development database [121]. Then the sufficient statistics of the train and the test data are extracted using the UBM. A total variability matrix (T-matrix) of 400 columns is trained using the sufficient statistics extracted from the development set. Then the train, test and development data i-vectors are extracted using transformation by T-matrix. Linear discriminant analysis (LDA) and WCCN matrices are trained using the development data i-vectors of dimension 250 and full rank, respectively for having channel/session compensation as discussed in Chapter 3. Finally, cosine kernel scoring between the test and the claimed i-vector is performed to generate the similarity score. The block diagram of the baseline framework can be seen from Figure 4.3 (a).

The structure of the proposed framework is different from the baseline framework. The former deals with a user chosen phrase and a text-constrained sentence for training, whereas the latter uses 3 minutes of read speech for training the speaker models. This provides the phonetic content match for the text-constrained model as the same user chosen phrase along with text-constrained sentence is spoken during testing. The rest of the structure and the methodology of the text-constrained model based framework are the same as that of the baseline system as discussed. Figure 4.3 (b) shows the proposed text-constrained model based framework designed for text-independent SV studies. The studies to compare the performance of the two frameworks are presented in the following subsection.

**Table 4.1:** Performance of different SV frameworks.

| SV System Framework Condition | EER (%) |
|---|---|
| (I) 3 minutes training vs. limited test data | 23.00 |
| (II) 1 minute training vs. limited test data | 23.70 |
| (III) text-constrained model vs. limited test data | **11.30** |
| (IV) (text-constrained + 1 minute) model vs. limited test data | **20.60** |

### 4.2.3 Experimental studies and analysis of results

The baseline system for the current study is built by considering the 3 minutes of read text speech in the train session. On the other hand, roll number with name and a text-constrained sentence are used as the limited test data. This SV system is found to give a high equal error rate (EER) of 23.00% as can be seen from Table 4.1. Another study is then made by limiting the amount of training session data to 1 minute for creation of speaker models. This yields a slightly poorer EER of 23.70% when compared to the baseline system. It indicates that when the amount of test data is limited, the amount of train data of 1 minute or 3 minutes does not significantly affect the performance.

The proposed framework based study using text-constrained models is carried out by considering the roll number and name along with the text-constrained sentence for training. The same is repeated by the users during testing sessions, which is specific to each user. The performance of this framework can be seen from Table 4.1. It shows a considerable improvement over the baseline, giving an EER of 11.30%. It is hypothesized that this improvement is due to the phonetic match between the train and the test sessions, while extracting the i-vectors in a speaker-specific text-constrained framework. Additionally, another experimental study is performed to determine the effect of including the data of the same phonetic content of the test session as part of the train session. For this, the speaker models are trained by adding 1 minute of text-independent data along with the speech that is considered for building the text-constrained speaker-specific models. For this condition, the SV system gives an EER of 20.60% that is significantly better than the EER of the baseline case, where 3 minutes of read speech is used. This further confirms the importance of the phonetic match for text-independent SV. Thus it supports our hypothesis of using a speaker-specific text-constrained model based SV framework under limited data condition.

Figure 4.4 shows the detection error tradeoff (DET) plots for tdifferent SV system framework conditions listed in Table 4.1. It is observed that the proposed framework represented by, Condition

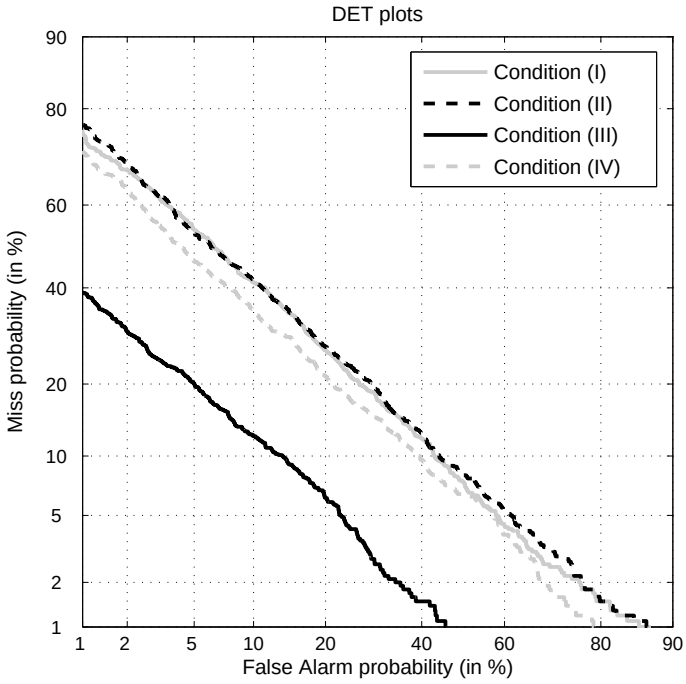


**Figure 4.4:** DET plots for different SV system framework conditions considered.

(III) having the text-constrained speaker-specific model performs better than the remaining systems. This signifies the importance of a phonetic match for text-independent SV from the perspective of practical systems under limited data condition.

## 4.3 Text-constrained models for sufficient train with limited test data speaker verification

The recently made available RedDots database is designed for SV studies with short utterances [53]. Various attempts have been carried out for studying different aspects from the short utterance perspective [122–127]. The authors have used tandem features for text-dependent SV studies using RedDots corpus and proved their importance for short utterances [122]. In [123], an i-vector/hidden Markov model (HMM) based system that performed better than the conventional i-vector/Gaussian mixture model (GMM) based system is proposed. Similarly, the significance of unsupervised HMM-UBM and temporal GMM-UBM over the classical modeling techniques is demonstrated in [124]. The joint speaker and lexical modeling for short-term characterization of speakers on Part IV of RedDots database for two different enrollment conditions is considered in [125]. Studies related to parallel speaker and content modeling on RedDots database are presented in [126].

In this work, the focus is to investigate the importance of constraining the text for text-independent SV systems for sufficient train with limited test data based scenario. The RedDots database has a subset for which the effectiveness of text-dependent and text-prompted enrollment conditions can be explored. The mentioned subset of the database, i.e. Part IV can be used for sufficient train with limited test data based text-independent SV studies. This is because the prompted text based enrollment condition deals with 10 different fixed phrases for training the speaker models. During testing a short utterance of different lexical content is used. Thus it provides a framework of sufficient train with limited test data based SV, which is of our interest. For maintaining uniformity with the thesis framework, the text-prompted is referred to as the text-independent enrollment condition from this point onwards. On the contrary, the text-dependent based enrollment condition is similar to the conventional framework of text-dependent SV. It considers three fixed phrases of the same lexical content for training and a phrase of same lexical content during testing. We intend to explore the significance of the phonetic match in terms of text-constrained models in place of conventional text-independent based enrollment condition. Two different frameworks are proposed for exploring the significance of content match under sufficient train with limited test data condition. The text-constrained model based framework is expected to increase the efficacy of the text-independent enrollment condition and bring it closer to that of the text-dependent SV. The initial SV studies are conducted using MFCC features over i-vector and PLDA based framework [32, 128].

In the previous chapter, the different attributes of source information are combined with the MFCC features for the sufficient train and limited test data based framework. Motivated from their proven significance, they are considered for the current work. The three source features mel power difference of spectrum in subbands (MPDSS), residual mel frequency cepstral coefficient (RMFCC) and discrete cosine transform of integrated linear prediction residual (DCTILPR) are considered in the context of text-dependent and text-independent based enrollment conditions on Part IV of RedDots database. Further, their significance is investigated under the text-constrained models. The source features are expected to work better for the text-independent based enrollment condition than the text-dependent based enrollment condition as observed from the studies in the previous chapter.

### 4.3.1 Investigating text-dependent and text-independent enrollment conditions

As discussed, the Part IV of RedDots database has two different enrollment conditions, which are text-dependent and text-independent. Here, the two enrollment conditions are investigated and

studied in detail. A baseline system using i-vector based modeling at front-end with PLDA at back-end is developed for performing the SV studies.

#### 4.3.1.1 Database, preprocessing and feature extraction

The RedDots database used in this study is collected with the collaborative effort of 21 countries across the world as a part of the RedDots project. This database comprises of speech data from 49 male and 13 female speakers totaling 62 speakers. It contains four different parts Part I to Part IV categorized according to the nature of the short utterances, the details of which may be found from [53]. However, for this work we focus on the Part IV of the database that has two different enrollment conditions. In one of these conditions, three short fixed phrases of the same text content for creating sentence specific speaker models. In the other condition, 10 short fixed phrases with different text content are used to build the speaker models. The first condition corresponds to the text-dependent framework and the second condition corresponds to the text-independent framework. It is to be noted that in text-independent based enrollment condition, the fixed phrases used are different from those that are used for text-dependent based enrollment condition for the same speaker. However, the test trials are same for both the enrollment conditions, where they match in lexical content for the text-dependent case.

The utterances of the database are short term processed with Hamming windowed frame of size 20 ms keeping a shift of 10 ms. A 39-dimensional (13-base + 13-$\Delta$ + 13-$\Delta\Delta$) MFCC feature vector is extracted for each short term processed frame. The features of the speech regions are taken by performing energy based VAD over which CMVN is performed.

#### 4.3.1.2 Baseline experimental setup and studies

The SV studies are performed on RedDots database using i-vector modeling in the front-end and PLDA in the back-end. The RSR2015 database containing the speech data from 300 speakers data for text-dependent SV studies is used as development data for training the UBM, T-matrix and PLDA model [52]. Two gender dependent UBMs of 512 mixture components are trained using the male and the female subsets of RSR2015 database. The sufficient statistics of the RedDots database as well as the development database are extracted for the male and the female subset separately, using respective male and female UBM. Two gender dependent T-matrices of 150 columns are computed using development data statistics of male and female subset, respectively. The i-vectors are then

**Table 4.2:** Number of trials for Part IV of RedDots database.

| RedDots Subset | Target Correct | Impostor Correct | Target Wrong | Impostor Wrong |
|---|---|---|---|---|
| Female | 1,122 | 3,906 | 25,806 | 180,462 |
| Male | 5,696 | 99,264 | 131,002 | 4,999,686 |

**Table 4.3:** Baseline system performance using MFCC features: Text-dependent and text-independent framework based enrollment conditions on Part IV of RedDots database.

| Male Subset | | | | Female Subset | | | |
|---|---|---|---|---|---|---|---|
| Text-dependent | | Text-independent | | Text-dependent | | Text-independent | |
| EER (%) | DCF | EER (%) | DCF | EER (%) | DCF | EER (%) | DCF |
| 9.50 | 0.0433 | 13.55 | 0.0598 | 13.73 | 0.0514 | 19.25 | 0.0679 |

extracted for the mentioned database using the T-matrix . Further, two 100-dimensional PLDA models are learned using the development data i-vectors of the male and the female subsets. The verification of a claim is performed according to the evaluation procedure of RedDots database.

The two different enrollment conditions for Part IV of RedDots database, which are text-dependent and text-independent are evaluated. There are four kinds of trials considered for the evaluation of RedDots database, namely *Target Correct*, *Target Wrong*, *Impostor Correct* and *Impostor Wrong* [53]. Table 4.2 shows the details of the number of trials for each category. Performance for *Target Wrong*, *Impostor Correct* and *Impostor Wrong* are computed against *Target Correct* category. The test condition *Target Wrong* refers to the genuine speakers producing wrong phrases. On the other hand, *Impostor Correct* denotes the impostors producing correct fixed phrases and *Impostor Wrong* represents the impostors producing wrong fixed phrases. The testing conditions based on *Target Wrong* and *Impostor Wrong* are not considered as their scope is limited in a cooperative scenario for practical systems. The trial condition of only *Impostor Correct* comes under the scope of this work and the results of baseline system using MFCC features are shown in Table 4.3 in terms of EER and detection cost function (DCF). It shows that for both male and female subsets of Part IV of RedDots database, the text-dependent based enrollment condition provides better performance than that of the text-independent condition. This indicates the importance of the lexical match between train and test sessions while dealing with limited data. Further, the results show that the performance for the male subset is better than that obtained with the female subset. This may be due to the smaller ratio of true trials to false trials in case of the male subset than that of the female subset as the number of speakers is very small in the latter subset of RedDots database considered for this study.

### 4.3.2 Text-constrained framework for sufficient train and limited test data

To explore the significance of lexical content match in text-independent enrollment condition for sufficient train with limited test data based SV, two frameworks are proposed. These have evolved from the text-dependent and text-independent enrollment conditions.

#### 4.3.2.1 Condition 1

The first framework is based on the creation of speaker models by considering models of both text-dependent and text-independent enrollment conditions. This is performed by averaging the i-vectors computed from the two enrollment conditions. For better understanding let us consider a text-dependent model for sentence 'a' and speaker 'X', which is referred to as $TD_{Xa}$. The corresponding text-independent based model is represented by $TI_X$. Then the result of *Condition 1* based framework model will be the average of $TD_{Xa}$ and $TI_X$ i-vectors to generate a model for each speaker-sentence pair. This setup has been made to check whether merging the presence of same phonetic content based models along with models trained with sufficient data, provides some impact on SV performance.

#### 4.3.2.2 Condition 2

The second framework is obtained by replacing one of the 10 short fixed phrases by a fixed phrase having the lexical content match to that of the test phrase in text-independent enrollment condition. It is done by replacing one of the phrases so that the number of examples taken for the creation of sufficient train based models remain same as that of the baseline setup. For the current study, this is made by choosing one out of the three fixed phrase examples of the text-dependent based enrollment condition. It is also to be noted that the replacement of fixed phrases here indicates the replacement of respective i-vector representation of the fixed phrases. This is made so that the effectiveness in presence of same text content in the text-independent enrollment condition can be judged. The *Condition 2* refers to the text-constrained model based framework having sufficient train with limited test data.

Figure 4.5 shows the block diagrammatic representation of different frameworks that are discussed. The text-dependent enrollment condition for Part IV of RedDots database is shown in Figure 4.5 (a), where three examples of a fixed phrase (say some Phrase 'a') are used for training the speaker model and the same lexical content based phrase is used during testing. In Figure 4.5 (b), text-independent enrollment condition is illustrated, where 10 different fixed phrases (say Phrases 'b', 'c', ..., 'k') are used for training the model. Further, testing is made with a fixed phrase which is not a part of the set

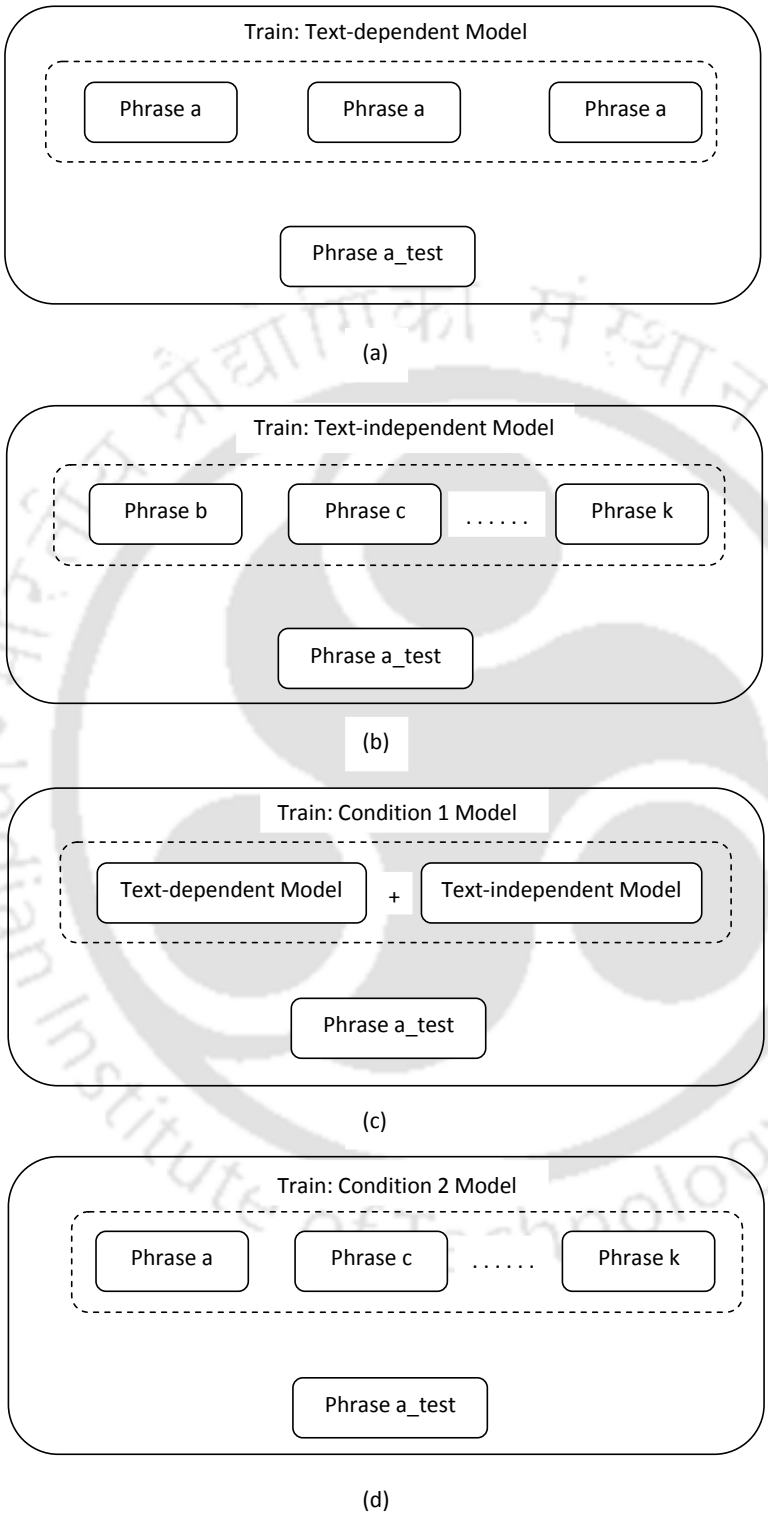


**Figure 4.5:** Block diagram representation of different enrollment conditions at the i-vector level. (a) Text-dependent (b) Text-independent (c) Condition 1 (d) Condition 2.

of phrases used for training the speaker models (say Phrase 'a' in this case). Thus for text-dependent based enrollment condition, the speakers of the RedDots database have a fixed phrase specific speaker model. On the contrary, the latter deals with a single speaker model trained with 10 different short phrases. The *Condition 1* deals with the creation of speaker models by averaging the i-vector models of text-dependent and text-independent based enrollment conditions as can be seen from Figure 4.5 (c). Finally, Figure 4.5 (d) shows the text-constrained model based framework, which is termed as *Condition 2*. In the text-constrained framework, as observed from the figure, one out of the 10 phrases taken for the text-independent based enrollment condition is replaced by a fixed phrase of matching lexical content to that of the test speech (say Phrase 'b' is replaced with Phrase 'a'). This has been done to maintain the match in phrase between train and test sessions. As discussed earlier, the text-constrained model is expected to perform better due to having some lexical content match compared to the text-independent based enrollment condition.

#### 4.3.2.3 Experimental results and observations

The *Condition 1* and *Condition 2* based setups are evaluated on the Part IV of RedDots database and the results are reported in Table 4.4. It shows that due to the addition of a sentence specific model information to sufficient train scenario in *Condition 1*, it helps in speaker modeling. Hence this framework yields better results than the text-dependent and text-independent based frameworks, when compared to the baseline framework as given in Table 4.3. In case of *Condition 2*, one of the examples of fixed phrases for sufficient train based speaker model is replaced with a fixed phrase having a matching lexical content to the test session. Thus it provides a better match and the same is reflected in terms of performance improvement. Figure 4.6 shows the DET plots for text-dependent and text-independent based enrollment setups and their comparison to *Condition 1* and *Condition 2* based frameworks for the Part IV of RedDots database. It indicates the significance of both the frameworks generating improved results than the baseline text-independent based enrollment condition. However, *Condition 1* based setup requires both text-dependent and text-independent based models. In this regard, the text-constrained model based framework represented by *Condition 2* is found to be favorable as it is similar to the implementation of text-independent SV. It is also to be noted that the text-constrained model helps to minimize the performance difference between text-dependent and text-independent enrollment conditions. Thereby showing its significance, which is of our interest in this current work for SV using sufficient train and limited test data based scenario.

**Table 4.4:** Performance for text-constrained based setups on Part IV of RedDots database.

| Male Subset | | | | Female Subset | | | |
|---|---|---|---|---|---|---|---|
| Condition 1 | | Condition 2 | | Condition 1 | | Condition 2 | |
| EER (%) | DCF | EER (%) | DCF | EER (%) | DCF | EER (%) | DCF |
| 8.13 | 0.0372 | 11.20 | 0.0505 | 12.30 | 0.0489 | 17.38 | 0.0646 |

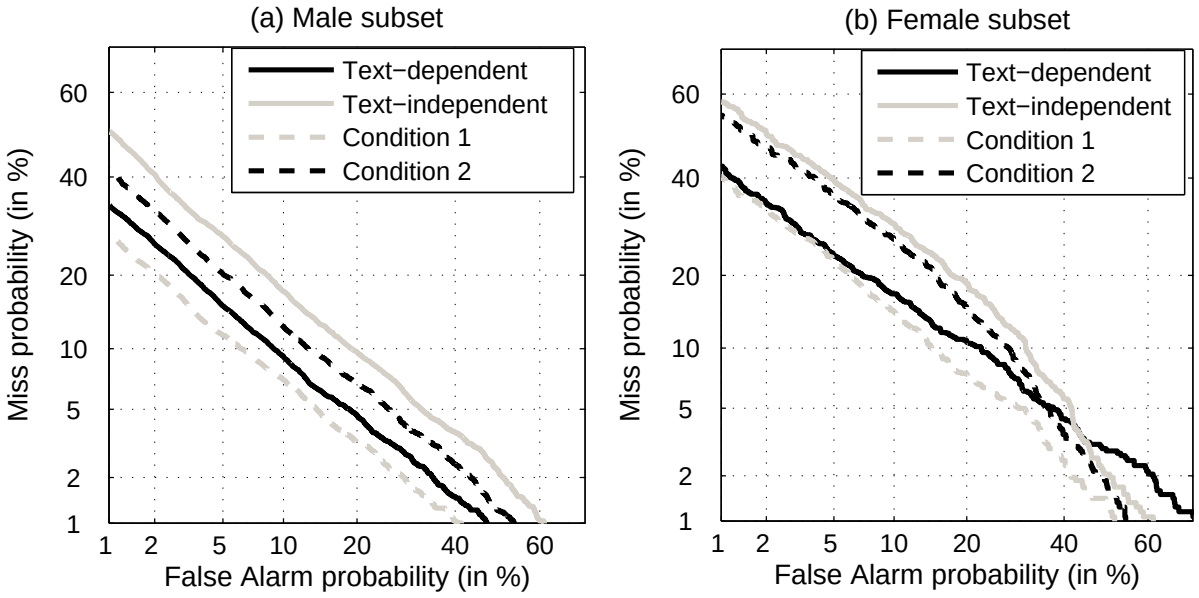


**Figure 4.6:** DET plots for text-dependent and text-independent based enrollment and their comparison to *Condition 1* and *Condition 2* based setups for (a) male and (b) female subsets of Part IV of RedDots database.

The text-constrained model based framework discussed here has significance in real-world scenario for a text-independent SV task under sufficient train with limited test data condition. This can be visualized by a system having enrollment condition that deals with collecting a set of fixed phrases of 10 or more in number from the users for training the speaker models. On the other hand, during testing only a random phrase out of the set of phrases taken for enrollment is expected. This will provide a match in the lexical content of the test phrase to one of the phrases from the train session. It will then provide a better way of testing than with the case of generic text-independent SV having sufficient train with limited test data condition. Thus a text-constrained model based framework may provide some degree of improvement over the conventional text-independent SV. The remaining part of this current work focuses on ways to improve the performance of this text-constrained model to reach closer to that obtained with text-dependent based SV system.

**Table 4.5:** Performance on Part IV of RedDots database for the three source features in different setups.

| Feature Used | RedDots Subset | Text-dependent | | Text-independent | | Condition 1 | | Condition 2 | |
|---|---|---|---|---|---|---|---|---|---|
| | | EER (%) | DCF | EER (%) | DCF | EER (%) | DCF | EER (%) | DCF |
| MPDSS | Male | 25.63 | 0.0904 | 23.56 | 0.0939 | 23.26 | 0.0905 | 23.05 | 0.0920 |
| | Female | 24.51 | 0.1000 | 20.86 | 0.0978 | 19.43 | 0.0992 | 21.21 | 0.0991 |
| RMFCC | Male | 26.58 | 0.0913 | 24.93 | 0.0923 | 23.44 | 0.0898 | 23.86 | 0.0915 |
| | Female | 32.71 | 0.0924 | 29.77 | 0.0940 | 28.25 | 0.0929 | 30.39 | 0.0940 |
| DCTILPR | Male | 31.95 | 0.0996 | 30.16 | 0.0999 | 29.20 | 0.0993 | 30.27 | 0.0999 |
| | Female | 37.17 | 0.0995 | 34.40 | 0.0995 | 34.49 | 0.0996 | 34.76 | 0.0995 |

### 4.3.3 Source features for text-constrained models

The source features MPDSS, RMFCC and DCTILPR explored in Chapter 1 are found to contain different attributes of excitation source information. MPDSS captures periodicity, while RMFCC captures the smoothed spectrum information. On the other hand, the source feature DCTILPR captures the shape of glottal signal. These three different attributes on fusion provided an improvement due to complementary nature of information carried by each of them. With their proven significance under SV limited test data as discussed in the previous chapter, they are considered for the studies related to text-constrained models.

The excitation source features MPDSS, RMFCC and DCTILPR are evaluated on the Part IV of RedDots database for text-dependent and text-independent based enrollment conditions. The results for the same can be observed from Table 4.5. They indicate that the source features behave better for the text-independent than the text-dependent based enrollment condition in almost all the cases. It thus signifies their less dependency on the amount on phonetic match. Further, the performance of each source feature is poorer compared to that obtained with MFCC features. The proposed frameworks *Condition 1* and *Condition 2* that are explained in Section 4.3.2 are also explored using the source features as can be seen from Table 4.5. In these experimental setups, the source features are not able to contribute as much. The performance is more or less similar to that obtained with text-independent based SV framework. However, the literature mentions regarding the complementary nature of information carried by the source features to that carried by the vocal tract features. Hence their fusion is expected to contribute towards improving the performance.

The fusion of the three source features MPDSS, RMFCC and DCTILPR having different aspects of excitation source information is carried out at the score level similar to that made in Chapter 3. The results of fusion for each of the source features with MFCC and that with one another is shown in the

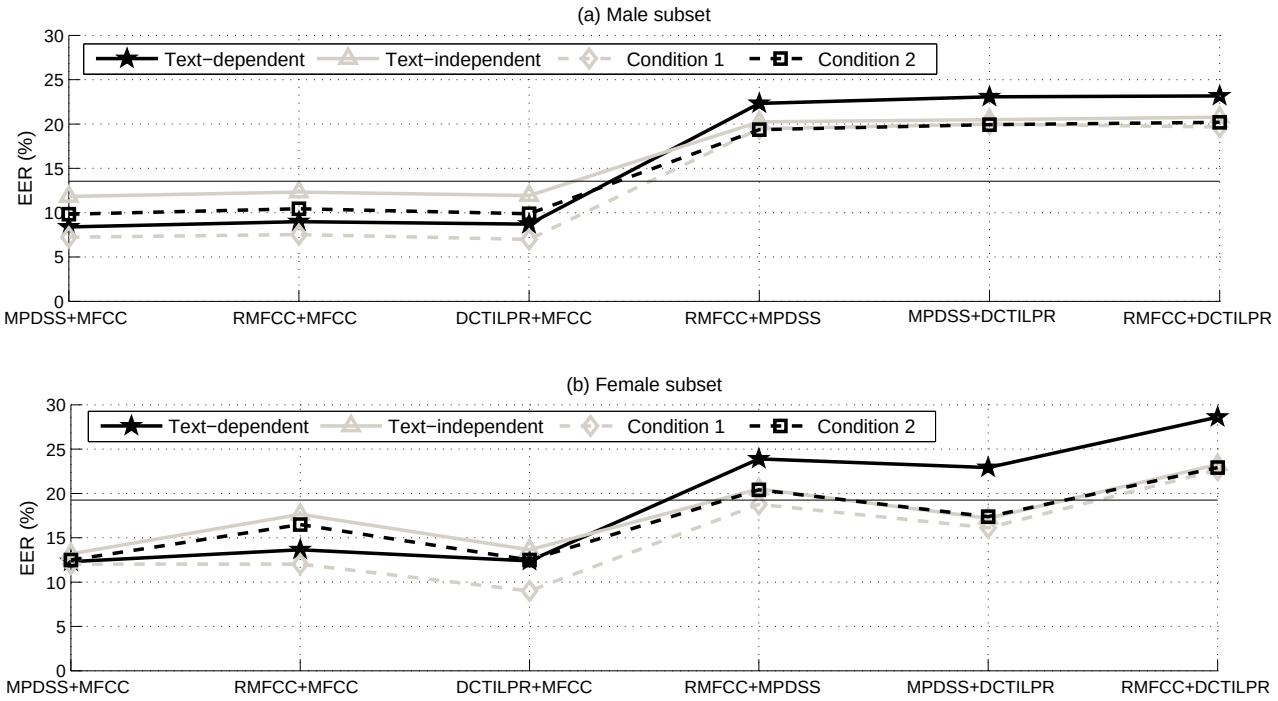


**Figure 4.7:** Performance trends for combination of two feature pairs on Part IV of RedDots database for different setups. The horizontal thin black line indicates the baseline result with MFCC features for the text-independent based enrollment condition for comparison.

Figure 4.7. The observations from the figure bring light towards the significance of source information for each case. All the three source features in combination to MFCC features produce significant improvement for the different enrollment conditions considered in this work. The improvements are more evident for MPDSS and DCTILPR features in most of the cases. Additionally, the nature of complementariness is found in all the three aspects of source information. This is because they showed better results on fusion with one another as source feature pairs than that with an individual source feature. Then a combination of three features is also carried out in order to investigate their effectiveness for different enrollment conditions. Figure 4.8 shows the performance trends under the fusion of three features at a time. It indicates that the combination of MPDSS, DCTILPR and MFCC provides the best among them for most of the cases. Further, the three source features MPDSS, RMFCC and DCTILPR are fused with MFCC as they have already shown their significance for the conventional text-independent SV with limited test data based framework.

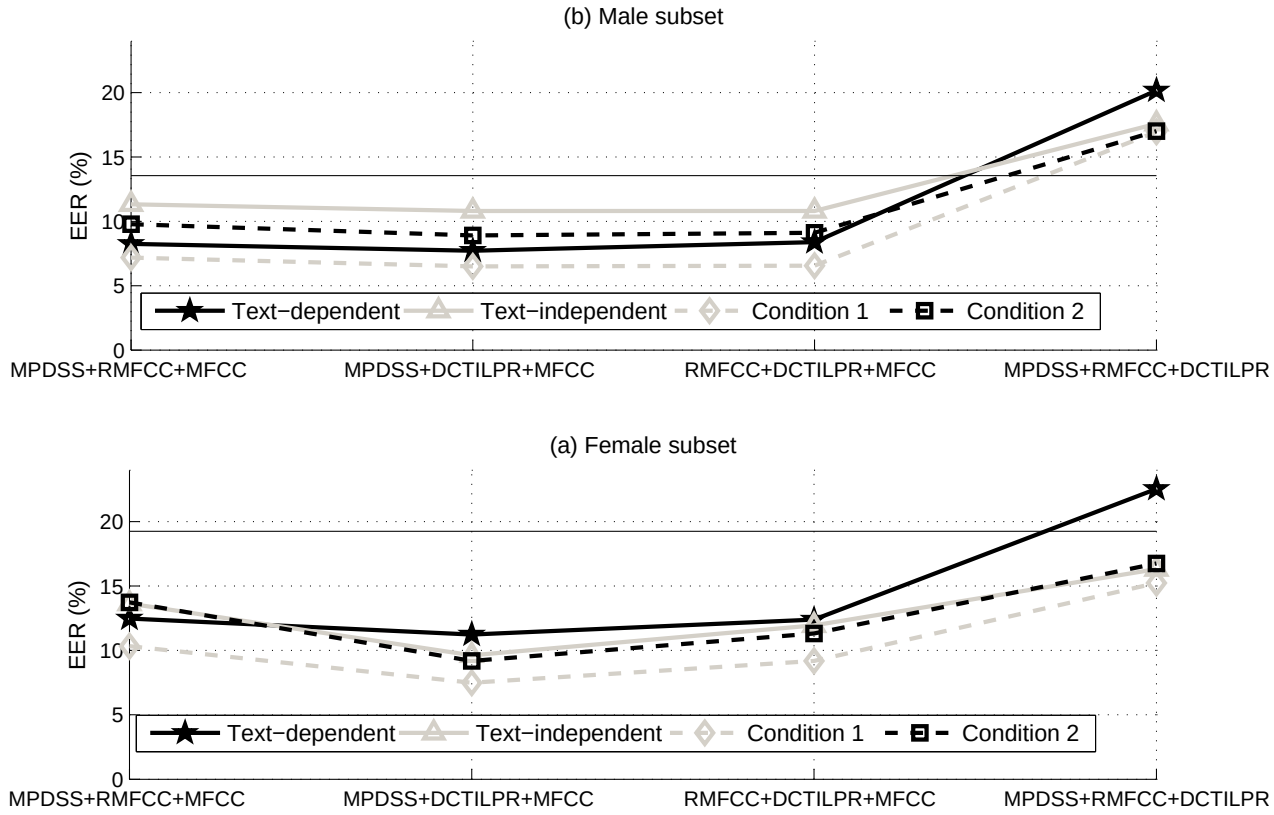


**Figure 4.8:** Performance trends for combination of three feature pairs on Part IV of RedDots database for different setups. The horizontal thin black line indicates the baseline result with MFCC features for the text-independent based enrollment condition for comparison.

The final combined system performance in consideration of all the features is shown in the Table 4.6. It is observed that this can further bring down the EER and DCF measures, which minimizes the performance difference between text-dependent and text-independent based enrollment condition. Additionally, it is observed that the text-constrained model based condition, i.e. *Condition 2* for sufficient train with limited test data based text-independent SV framework is benefited significantly in combination of the source features. This has resulted due to the complementary nature of information carried by each of them that helps to achieve closer performance to that obtained from text-dependent based enrollment condition. Although the source features are less dependent on the phonetic content, there is some amount of dependency that helps in the current study. This seems to produce better results for *Condition 2* than that for the text-independent based enrollment framework when fused

**Table 4.6:** Performance on Part IV of RedDots database using fusion of MFCC with three source features MPDSS, RMFCC and DCTILPR.

| RedDots Subset | Text-dependent | | Text-independent | | Condition 1 | | Condition 2 | |
|---|---|---|---|---|---|---|---|---|
| | EER (%) | DCF | EER (%) | DCF | EER (%) | DCF | EER (%) | DCF |
| Male | 7.90 | 0.0358 | 10.11 | 0.0445 | 6.37 | 0.0284 | 8.66 | 0.0383 |
| Female | 11.32 | 0.0452 | 9.45 | 0.0490 | 7.49 | 0.0381 | 9.54 | 0.0487 |

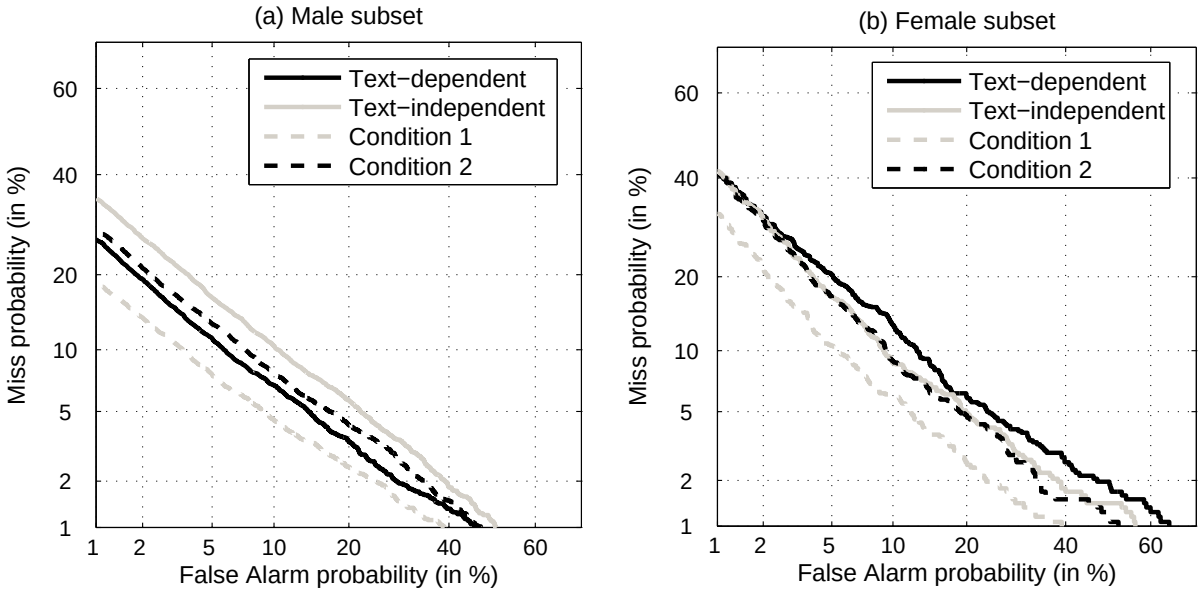


**Figure 4.9:** DET plots for different setups obtained using fusion of MFCC with the three source features MPDSS, RMFCC and DCTILPR on (a) male and (b) female subsets of Part IV of RedDots database.

with the vocal tract features. It thus signifies the usefulness of the source features. Further, they are more evident in case of text-constrained models than the conventional text-independent framework under the sufficient train with limited test data based SV that has scope for practical systems.

Figure 4.9 shows the DET curve trends under the combination of all the features depicting alternative speaker information for different setups when compared to the earlier results obtained with the baseline setups. It clearly conveys the importance of the current exploration in terms of text-constrained models with source information given by *Condition 2* based setup. This framework is able to minimize the performance difference between text-dependent and text-independent enrollment condition. Thus the studies showcase the effectiveness of a text-constrained model with consideration of source information for sufficient train and limited test data based practical SV systems.

## 4.4 Implicit utilization of phonetic information

The text-constrained model explored in this chapter utilizes the lexical content between train and test session in an explicit manner. It puts some constraint on the text to be spoken during train and test sessions. In this section, the VTC feature is used for the utilization of phonetic content for speaker-specific information in an implicit manner. The VTC evidence is investigated along with the studies conducted for SV using sufficient train and limited test data based scenario.

### 4.4.1 VTC feature

The VTC feature is an attempt to capture the level of constriction in the vocal tract while producing different sound units [129]. The level of constriction is different for various categories of sound units starting from the low vowels to the voice bars. It is found that the low vowels show the least, whereas the voice bars produce the highest level of constriction in the vocal tract. The VTC evidence is computed as the dot product of speech signal and the differenced zero frequency filtered signal (ZFFS) in an epoch synchronous manner [129, 130]. Let, $x(n)$ and $z(n)$ be the speech and the ZFFS between two successive epochs, then the corresponding VTC feature ($V_{tc}$) for that epoch interval is given by,

$$V_{tc} = \frac{< x(n),\ z(n) >}{\|x(n)\|\ \|z(n)\|} \tag{4.1}$$

The VTC evidence is found to be useful for the phoneme recognition task as mentioned in [129]. The authors showed that the VTC evidence contains complementary information to that carried by MFCC and helps in fusion. However, the distributions of the VTC evidence shown for different categories of sound units in [129] overlapped by a visible margin. This may have resulted due to no consideration of normalization of speaker-specific information. Generally, while working with tasks related to speech recognition, normalization techniques like vocal tract length normalization are used. These techniques are used for nullifying the speaker-specific information. Conversely, no such normalization techniques are applied on top of the VTC. Every speaker has unique vocal tract shape, size, etc. This may result in level of constriction to be specific for different phonemes in a particular language. Therefore, the VTC feature is expected to carry speaker-specific phonetic information. Figure 4.10 shows the nature of VTC evidence plotted for two different speakers from TIMIT database for a portion of same lexical content based sentence SA1 [111]. This shows that although there is a similar trend, it has some speaker-specific nature that is visible during transitions from one sound unit to the other. Therefore,

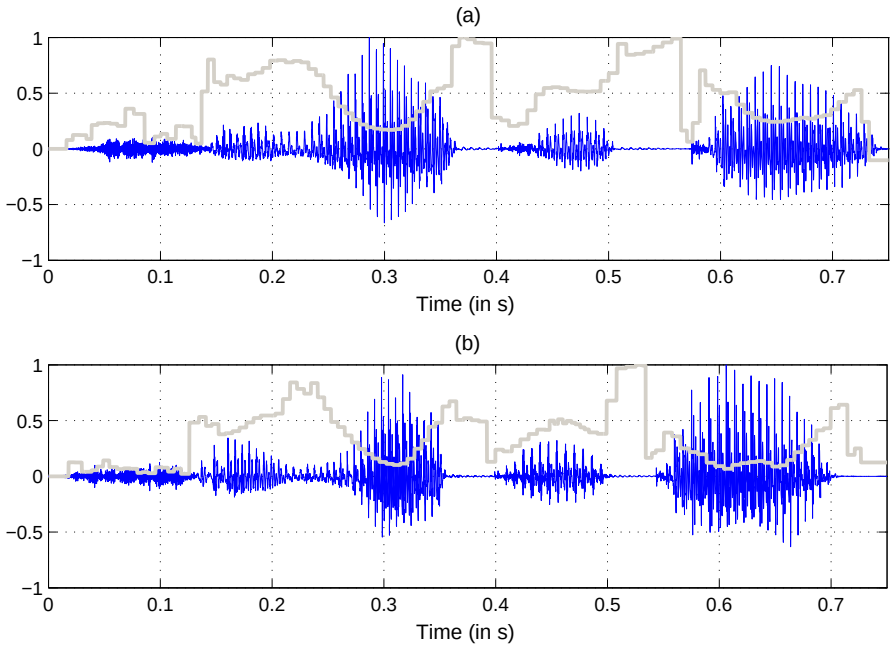


**Figure 4.10:** Vocal tract constriction evidence of two different speakers for the same lexical content based speech (a) Speaker-1 (b) Speaker-2.

it can capture the speaker-specific phonetic information which may be useful to provide additional speaker characteristics.

### 4.4.2 Experimental studies

The NIST SRE 2003 database considered for SV using sufficient train with limited test data as described in the previous chapter. The utterances of the database are short term processed with Hamming windowed frame of size 20 ms keeping a shift of 10 ms and 39-dimensional (13-base + 13-$\Delta$ + 13-$\Delta\Delta$) MFCC features are extracted. The features of the speech regions are taken by performing energy based VAD over which CMVN is applied. On the other hand, the VTC feature is extracted for the utterances of the database in a pitch synchronous manner as mentioned by (4.1). The VTC features belonging to the speech regions are considered similar to the case of MFCC features.

The i-vector based SV system explained in the previous chapter over NIST SRE 2003 database is developed using MFCC features. Switchboard corpus-2 is used for learning the background models for the i-vector based framework. To observe the effectiveness of the VTC feature on top of the existing baseline framework using MFCC features, a combination of VTC with MFCC is opted. As the VTC feature is only of single dimension, it is combined to the MFCC features at the feature level. It is

**Table 4.7:** Performance for fusion of VTC and MFCC features over i-vector based SV framework.

| Test Duration | MFCC | | MFCC + VTC | |
|---|---|---|---|---|
| | EER (%) | DCF | EER (%) | DCF |
| 10 s | 5.81 | 0.1090 | **5.56** | **0.1031** |
| 5 s | 10.52 | 0.1977 | **10.43** | **0.1937** |
| 3 s | 16.94 | 0.3100 | **15.94** | **0.2929** |
| 2 s | 22.31 | 0.4128 | **21.73** | **0.4017** |

to be noted that the VTC feature is extracted pitch synchronously, whereas MFCC is obtained by short term processing of speech signal. Therefore, VTC feature values within a frame are averaged to obtain a single value so that can be combined with MFCC at the feature level. Then the i-vector based SV system is developed with the fused features. The background models are learned using MFCC and VTC based combined features extracted from the development database for the i-vector based speaker modeling. The performance of the resultant SV system can be seen from Table 4.7 with its comparison to the baseline results obtained using MFCC features for limited test data SV. The fusion of the duo results in an improvement, which is visible for different durations of limited test data considered in this study. This improvement thereby indicates the importance of VTC feature for speaker characterization in such a scenario.

## 4.5 Summary

In this chapter, the phonetic information for speaker characterization is explored in an explicit and then in an implicit manner. Initially, a text-constrained model based framework is proposed from the context of making an explicit match of the lexical content between train and test sessions. The significance of this text-constrained model based setup is explored on a database collected from an IVRS callflow based setup and then on RedDots database having sufficient train with limited test data based framework. Different attributes of source information are also explored in the context of the text-constrained models, which show significant improvement on fusion to the conventional MFCC based vocal tract feature. The chapter later focuses on implicit exploitation of speaker-specific phonetic information in terms of VTC feature. This provides additional information for speaker characterization by showing improvement when fused to the MFCC features for SV under limited test data scenario.

# 5

# Exploring Kernel Discriminant Analysis and Combined Framework

Contents

### Overview

*In this chapter, an attempt for achieving an improved speaker characterization is made by working at the back-end of the speaker verification (SV) framework. Kernel based approaches have been found to perform well when the classes are not linearly separable. In this regard, kernel discriminant analysis (KDA) is explored for SV with limited test data based framework. It projects the i-vectors into a higher dimensional space and then performs discriminant analysis to remove the unwanted information for speaker modeling. The chapter later focuses on the proposal of a combined framework for dealing with SV under limited test data scenario by considering the explorations made in the previous chapters along with that made in this chapter.*

## 5.1 Introduction

SV systems for a practical application require limited test data based scenario as discussed under the scope of the thesis. The limited test data based SV is handled to some extent by using different attributes of source features and vocal tract constriction (VTC) feature having speaker-specific information as discussed in the previous chapters. However, there is still scope for having a better pattern recognition approach that may be useful in the overall framework. In this regard, there is a requirement at the back-end of the i-vector based SV system to focus and improve on it. With this motivation, a framework based on KDA is explored in this chapter. This may be useful for having a better channel/session compensation in the i-vector domain representation of the speakers for improved speaker characterization. The use of KDA in the i-vector based SV framework arose from the hypothesis that the channel/session information in the i-vector domain may not be linearly separable. In this regard, KDA may serve in a better way than the other existing approaches. Further, the i-vectors for limited data vary more due to the difference in capture of the phonetic content as small amount of data is considered. In such a case, KDA is expected to work in an advantageous way making a better separability across the speakers. The chapter later focuses on the proposal of a combined framework for SV with limited test data based on the explorations made throughout this thesis. This is based on combining the work explored in the previous chapters along with KDA based work conducted in this chapter. The proposed framework considers different attributes of excitation source information and VTC information along with conventional mel frequency cepstral coefficient (MFCC) features at the front-end for extraction of speaker characteristics. Additionally, KDA based

channel/session compensation is used at the back-end of i-vector based SV systems developed using different features. Finally, a score level fusion is performed to utilize the different directions considered for this work to achieve a high performance based SV system with limited test data based scenario.

The organization of this chapter is as follows: Section 5.2 describes the exploration made with KDA based work for SV with limited test data along with the results and analysis supporting its significance. In Section 5.3, the combined framework proposed using different attributes of source features, VTC and KDA is explained highlighting the final outcome of the proposed system. Finally, Section 5.4 provides the summary of the chapter.

## 5.2 Exploring kernel discriminant analysis

The field of SV has witnessed a breakthrough with the development of i-vector based speaker modeling [32]. Later, in [54, 131], the efficacy of the i-vector based SV system under short utterance conditions is explored. In [32], various channel/session compensation techniques have been carried out at the back-end, out of which linear discriminant analysis (LDA) followed by within class covariance normalization (WCCN) has provided better results. Further, in [54], LDA followed by WCCN has given comparable results for the condition of sufficient train with limited data test utterances ($\leq$ 10 s) to that obtained with Gaussian probabilistic linear discriminant analysis (GPLDA). In these works, the role of LDA is to reduce the dimension of i-vectors along with minimizing the intra-speaker variability and maximizing the separation between the speakers. Similarly, WCCN deals with reduction of session variability among the i-vectors.

It is known that LDA is a powerful feature extraction technique when classification is the task [33]. However, LDA as a dimensionality reduction technique can only transform the feature vectors onto a single hyperplane. Hence, LDA may not be a good option for many pattern recognition tasks when the data are not linearly separable [132, 133]. To address this problem, many researchers have worked on kernel based discriminant analysis techniques [134–137]. The principle of such methods is to map the data onto higher dimensional spaces where the classes are more separable, perform LDA in this space and dimensionally reduce to the desired dimensional space, thus separating the classes well. KDA has already been successfully used in various pattern recognition areas, a few of which are, face recognition [138–140], facial-expression recognition [141], hand-written digit recognition [135, 142], human activity recognition [143] and speech recognition [144, 145]. In [144] and [145], the KDA is

utilized to remove the speaker dependent part of the features in order to make them robust to speech variations for the speech recognition task. The authors of [146] use a multi-modal KDA in an SV system at the front-end feature extraction level for improving the robustness of the features extracted.

The i-vectors are found to vary with speaker, session and phonetic content of the utterance. Additionally, the variabilities are even more with short duration utterances [101]. The conventional techniques available minimize these variabilities by handling the data points linearly. It is hypothesized that, these techniques may not be suitable as the effects of the variabilities may be non-linear in nature. Further, the impact of these variabilities are expected to be more for short duration utterances due to larger variation with the effect of different phonetic content in text-independent SV. In this work, the efficacy of the i-vector based SV system is explored, when KDA is used at back-end. The proposed setup for SV using KDA is compared with the existing and the recent approaches in case of short utterances.

### 5.2.1 Kernel discriminant analysis

KDA performs a non-linear mapping from the actual feature space to a high dimensional space and then implements LDA on the mapped features. In this way, both non-linearity in the data and the class separation problem are addressed simultaneously. The mathematical formulation of KDA can be discussed as follows:

Suppose the training data points are given as $\mathcal{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n\}$ belonging to '$C$' classes and there is a non-linear mapping function $\varphi$, which transforms the data points to a space $\Gamma$. Let $\boldsymbol{S}_B^{\varphi}$, $\boldsymbol{S}_W^{\varphi}$ and $\boldsymbol{S}_t^{\varphi}$ be the between scatter matrix and within scatter matrix and the total scatter matrix, respectively, in the transformed space $\Gamma$ represented by,

$$\boldsymbol{S}_B^{\varphi} = \sum_{C} (\boldsymbol{\mu}_c - \boldsymbol{\mu})(\boldsymbol{\mu}_c - \boldsymbol{\mu})^T \tag{5.1}$$

$$\boldsymbol{S}_W^{\varphi} = \sum_{C} \sum_{x \in \mathcal{X}_c} (\varphi(\boldsymbol{x}) - \boldsymbol{\mu}_c)(\varphi(\boldsymbol{x}) - \boldsymbol{\mu}_c)^T \tag{5.2}$$

$$\boldsymbol{S}_t^{\varphi} = \sum_{i=1}^{n} (\varphi(\boldsymbol{x}_i) - \boldsymbol{\mu})(\varphi(\boldsymbol{x}_i) - \boldsymbol{\mu})^T \tag{5.3}$$

where $\boldsymbol{\mu_c}$ and $\boldsymbol{\mu}$ are the individual class mean and the overall mean respectively in transformed domains. In representation we have,

$$\boldsymbol{\mu}_c = \frac{1}{N_c} \sum_{j=1}^{N_c} \varphi(\boldsymbol{x}_j) \tag{5.4}$$

where, $N_c$ is the number of training samples under class '$c$'. Using the kernel trick, without actually computing the higher dimensional features, utilizing the dot-product of the input training features, the kernels are constructed. Gaussian radial basis function (RBF) kernel $k(\boldsymbol{x}, \boldsymbol{y}) = \exp(-\| \boldsymbol{x} - \boldsymbol{y} \|^2)/a$ or polynomial kernel $k(\boldsymbol{x}, \boldsymbol{y}) = (\boldsymbol{x}.\boldsymbol{y})^b$, where $a$, $b \in \boldsymbol{R}^+$ are examples of such kernel functions that satisfy the Mercer's condition as can be seen from the literature [135].

Let us define a kernel matrix $\boldsymbol{K}$ with $\boldsymbol{K}(i,j) = k(\boldsymbol{x}_i, \boldsymbol{x}_j)$ and a matrix $\boldsymbol{W}$ with $\boldsymbol{W}(i,j)$ as $\frac{1}{N_c}$ if $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ belong to the $c^{\text{th}}$ class or 0 otherwise. The actual problem of kernel discriminant analysis is to maximize the inter-class variance and minimize the intra-class variance which is equivalent to the eigen-value problem

$$\lambda \boldsymbol{S}_t^{\varphi} \mathbf{w} = \boldsymbol{S}_B^{\varphi} \boldsymbol{w} \tag{5.5}$$

where $\lambda$ are the eigen-values and $\mathbf{w}$ are the corresponding eigen-vectors. Here, the maximizing criteria $\lambda$ can be expressed as

$$\lambda = \frac{\boldsymbol{w}^T \boldsymbol{S}_B^{\varphi} \boldsymbol{w}}{\boldsymbol{w}^T \boldsymbol{S}_t^{\varphi} \boldsymbol{w}} \tag{5.6}$$

From the theory of reproducing kernels [147], we can express $\boldsymbol{w}$ (eigen-vector) as a linear combination of training samples in the transformed space $\Gamma$ as,

$$\boldsymbol{w} = \sum_{i=1}^{n} a_i \boldsymbol{\varphi}(\boldsymbol{x_i}) \tag{5.7}$$

By multiplying (5.5) with $\varphi^T(\boldsymbol{x_i})$ we obtain the following equation which has the same eigen-vectors as (5.5)

$$\lambda \varphi^T(\boldsymbol{x}_i) \boldsymbol{S}_t^{\varphi} \boldsymbol{w} = \varphi^T(\boldsymbol{x}_i) \boldsymbol{S}_B^{\varphi} \boldsymbol{w} \tag{5.8}$$

As mentioned in [134], using (5.6) and (5.8) $\lambda$ can thus be expressed as

$$\lambda = \frac{\boldsymbol{a}^T \boldsymbol{K} \boldsymbol{W} \boldsymbol{K} \boldsymbol{a}}{\boldsymbol{a}^T \boldsymbol{K} \boldsymbol{K} \boldsymbol{a}} \tag{5.9}$$

where $\boldsymbol{a}$ is a column vector with elements $a_i$. Once the eigen-vectors $\boldsymbol{a}$ are obtained, a test sample $\boldsymbol{x}$ can be used as,

$$(\boldsymbol{w}.\varphi(\boldsymbol{x})) = \sum_{i=1}^{n} a_i k(\boldsymbol{x}_i, \boldsymbol{x}) \tag{5.10}$$

$$(\boldsymbol{w}.\varphi(\boldsymbol{x})) = \boldsymbol{a}^T \boldsymbol{K}(:, \boldsymbol{x}) \tag{5.11}$$

where $\boldsymbol{K}(:, \boldsymbol{x}) = [k(\boldsymbol{x}_1, \boldsymbol{x}), k(\boldsymbol{x}_2, \boldsymbol{x}), \ldots, k(\boldsymbol{x}_n, \boldsymbol{x})]^T$. Let $\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_p$ be the eigen-vectors corresponding to the top $p$ eigen-values of $\lambda$. Then by defining a matrix $\boldsymbol{\Lambda}$ as $[\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_p]$, a sample $\boldsymbol{x}$ can

be represented in $p$ dimensional space by $\mathbf{\Lambda}^T\mathbf{K}(:,\mathbf{x})$. This algorithm is used in [135] for two-class problems, whereas in [134] and [148] for multi-class problems.

For obtaining a stable solution of the eigen-problem in (5.9), the denominator term $\boldsymbol{KK}$ must be non-singular. This can be ensured by adding a regularization term to $\boldsymbol{KK}$. This method was used in [148]. If we decompose the matrix $\boldsymbol{K}$, we obtain

$$\boldsymbol{K} = \boldsymbol{P\Sigma P}^T \tag{5.12}$$

where $\boldsymbol{\Sigma}$ is a diagonal matrix of sorted eigen-values and $\boldsymbol{P}$ is a matrix of normalized eigen-vectors, hence $\boldsymbol{P}^T\boldsymbol{P} = \boldsymbol{I}$. The matrix after adding the regularization term can be expressed as,

$$\boldsymbol{KK} + \epsilon\boldsymbol{I} = \boldsymbol{P\Sigma P}^T\boldsymbol{P\Sigma P}^T + \epsilon\boldsymbol{I} \tag{5.13}$$

$$\boldsymbol{KK} + \epsilon\boldsymbol{I} = \boldsymbol{P}(\boldsymbol{\Sigma}^2 + \epsilon\boldsymbol{I})\boldsymbol{P}^T \tag{5.14}$$

By substituting (5.12) and (5.14) in (5.9),

$$\lambda = \frac{\boldsymbol{a}^T\boldsymbol{P\Sigma P}^T\boldsymbol{WP\Sigma P}^T\boldsymbol{a}}{\boldsymbol{a}^T\boldsymbol{P}(\boldsymbol{\Sigma}^2 + \epsilon\boldsymbol{I})\boldsymbol{P}^T\boldsymbol{a}} \tag{5.15}$$

By substituting $\boldsymbol{b} = (\boldsymbol{\Sigma}^2 + \epsilon\boldsymbol{I})^{1/2}\boldsymbol{P}^T\boldsymbol{a}$ in (5.15),

$$\lambda = \frac{\boldsymbol{b}(\boldsymbol{\Sigma} + \epsilon\boldsymbol{I})^{-1/2}\boldsymbol{\Sigma P}^T\boldsymbol{WP\Sigma}(\boldsymbol{\Sigma} + \epsilon\boldsymbol{I})^{-1/2}\boldsymbol{b}}{\boldsymbol{b}^T\boldsymbol{b}} \tag{5.16}$$

Therefore, the $p$ eigen-vectors of the matrix $(\boldsymbol{\Sigma} + \epsilon\boldsymbol{I})^{-1/2}\boldsymbol{\Sigma P}^T\boldsymbol{WP\Sigma}(\boldsymbol{\Sigma} + \epsilon\boldsymbol{I})^{-1/2}$ will give us $p$ values of $\boldsymbol{b}$ and the corresponding $p$ values of $\boldsymbol{a}$ can be calculated.

### 5.2.2 Kernel discriminant analysis for speaker verification

The i-vectors, apart from possessing dominant speaker information, also contain information about session, channel and the phonetic content used. To eliminate the unwanted information for speaker modeling, LDA followed by WCCN or GPLDA has proven its efficacy as mentioned in [32,54]. However, it is hypothesized that these techniques may not be apropos in capturing the speaker dependent information. This is because they tend to eliminate salient information when the classes are not linearly separable. To overcome this drawback, KDA has been proposed as an alternative to the existing techniques for decreasing intra-speaker variability. In order to visualize the effect of KDA, the KDA transformed i-vectors and LDA followed by WCCN transformed i-vectors of three different

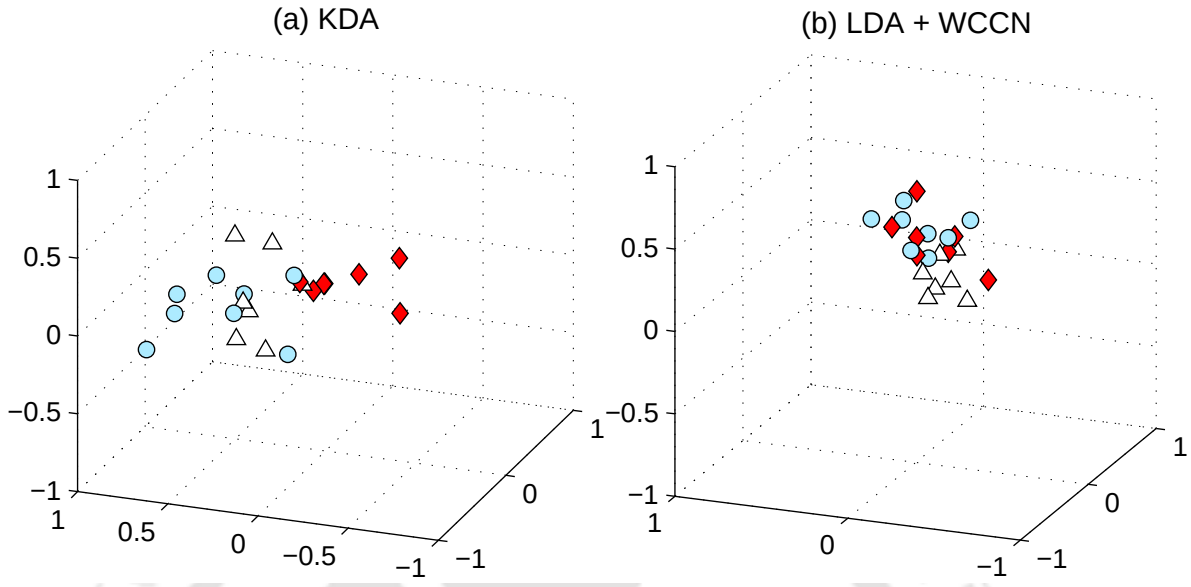


**Figure 5.1:** 3-D plots of i-vectors considering the top three dimensions obtained by performing PCA for three different speakers, showing better discrimination in the proposed KDA based framework over LDA followed by WCCN. The three different shapes depict three different speakers, whose i-vectors are plotted.

speakers of NIST SRE 2003 database are projected to 3-dimensional (3-D) space using principal component analysis (PCA) and are shown in Figure 5.1. It illustrates the effectiveness of KDA over LDA followed by WCCN based framework showing better discrimination of the i-vectors across speakers in the former. This provided the motivation to proceed with KDA based framework for i-vector based speaker modeling. The steps involved in performing KDA on top of i-vectors can be seen from Algorithm 1.

For constructing the Gram matrix in the proposed algorithm for channel/session compensation, any kernel function can be used. The Gaussian kernel function is used in our studies. The kernel function has been fine-tuned for different variances of Gaussian function by obtaining the equal error rate (EER) and detection cost function (DCF) values for each case. A unit variance found to be optimal and is chosen for the study. Figure 5.2 shows the detailed block diagrammatic representation of the proposed KDA framework for compensation at the back-end of i-vector based SV system. The descriptions related to the system development of the proposed and the existing techniques are discussed in the next subsection.

**Algorithm 1** : Proposed KDA based compensation.

1: Perform mean and length normalization of development i-vectors.
2: Construct the Gram matrix of the mean and length normalized development i-vectors using appropriate kernel function.
3: Decompose the calculated Gram matrix using the regularization technique.
4: Compute the eigen-vectors $\boldsymbol{b}$ and then $\boldsymbol{a}$ correspondingly.
5: Construct the matrix $\boldsymbol{\Lambda}$ from the vectors $\boldsymbol{a}$.
6: Perform mean and length normalization of the train i-vectors.
7: Construct the train kernel matrix ($\boldsymbol{K_{train}}$) using the normalized train, development i-vectors and appropriate kernel function.
8: Perform mean and length normalization of the test i-vectors.
9: Construct the test kernel matrix ($\boldsymbol{K_{test}}$) using the normalized test i-vector, development i-vectors and appropriate kernel function.
10: Compute the transformed train and test i-vectors as $\boldsymbol{\Lambda}^T \boldsymbol{K_{train}}$ and $\boldsymbol{\Lambda}^T \boldsymbol{K_{test}}$, respectively.
11: Perform cosine kernel scoring of a transformed test and a claimed i-vector to compare with the threshold.

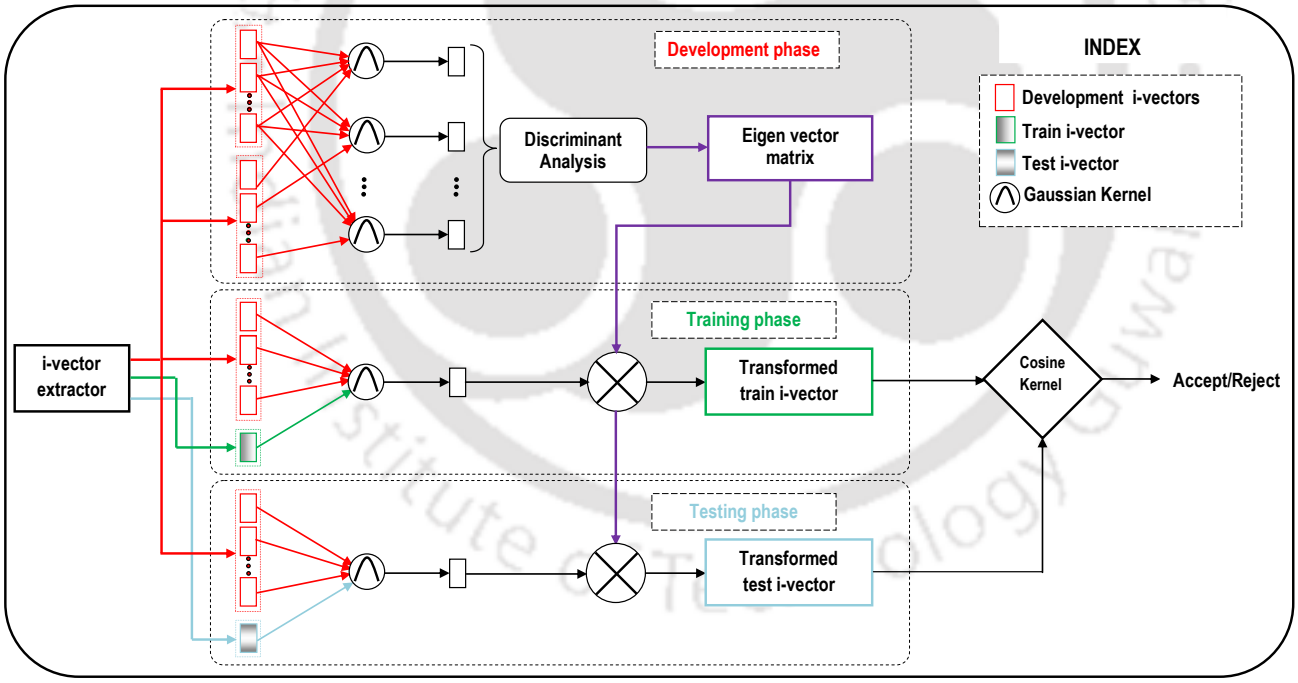


**Figure 5.2:** Block diagram of the proposed KDA based framework used in back-end of i-vector based speaker modeling.

### 5.2.3 System descriptions

The i-vector based SV system as outlined in Chapter 3 has been developed over standard NIST SRE 2003 database, which is considered for SV with limited test data based study. The utterances

are processed using 20 ms Hamming window with 10 ms shift. A 39-dimensional (13-base + 13-$\Delta$ + 13-$\Delta\Delta$) MFCC feature vector is extracted for each of the short term processed frames considering 22 logarithmically spaced filters. The features of the speech regions are retained after performing energy based voice activity detection (VAD). These feature vectors are then cepstral mean and variance normalized for having zero mean unit variance distribution [25].

Switchboard Corpus-2 database containing 1872 utterances from 417 speakers is used as development data for learning universal background model (UBM) and total variability matrix (T-matrix). A gender independent 1024-component UBM is trained using 251 utterances from male and 251 utterances from female speakers to maintain the equal amount of speech from both genders. This is followed by computation of T-matrix containing 400 columns using the zeroth and the first order statistics of the 1872 utterances from development data with respect to the UBM. The 400-dimensional i-vectors of the train, test and development data are estimated using the T-matrix and the respective sufficient statistics. An LDA followed by WCCN framework is implemented at the back-end as described in [32] for the baseline classical setup. LDA is performed on the development data i-vectors and the transformation matrix comprising the eigen-vectors corresponding to top 150 eigen-values is obtained. The WCCN matrix is calculated using the LDA transformed development data i-vectors. Then the train and the test i-vectors are further transformed using this matrix before evaluating the score via cosine kernel. Another back-end of the SV framework is made with LDA followed by GPLDA based setup for comparison as done in [128]. The GPLDA based method can decompose the i-vectors into speaker and channel components and then perform scoring based on likelihood measure. Further, the recently available method of short utterance variance normalization (SUVN) [101] is also taken into consideration for this study for comparing to the proposed approach. This method is found to be effective for compensating the phonetic mismatch among short duration i-vectors. The SUVN based technique is based on the fact that the i-vectors generated from the short utterances have more variation. This is because the phonetic content is different from one another and the acoustic space for covering the vocabulary of the language is less. Thus there is a high variation on the estimated i-vectors of the same speaker itself due to this phonetic mismatch. In order to reduce this variation, the i-vectors of the short utterances are normalized into the same space by SUVN method. Here the matrix for transforming the i-vectors is learned using a large number of utterances from the development data and then capturing the variation between full and truncated short utterance i-vectors obtained from

**Table 5.1:** Performance for different compensation techniques with dimension (D).

| Test Duration | LDA-WCCN (150-D) | | LDA-GPLDA (150-D) | | SUVN (150-D) | | KDA (150-D) | | KDA (400-D) | |
|---|---|---|---|---|---|---|---|---|---|---|
| | EER (%) | DCF | EER (%) | DCF | EER (%) | DCF | EER (%) | DCF | EER (%) | DCF |
| Full | 2.48 | 0.0474 | 1.94 | 0.0338 | - | - | 1.58 | 0.0270 | **0.81** | **0.0148** |
| 10 s | 5.81 | 0.1090 | 4.92 | 0.0880 | 5.74 | 0.1050 | 4.74 | 0.0845 | **2.80** | **0.0501** |
| 5 s | 10.52 | 0.1977 | 9.58 | 0.1803 | 9.67 | 0.1768 | 9.53 | 0.1724 | **5.96** | **0.1068** |
| 3 s | 16.94 | 0.3100 | 16.17 | 0.2975 | 15.09 | 0.2824 | 15.40 | 0.2860 | **11.20** | **0.1954** |
| 2 s | 22.31 | 0.4128 | 21.82 | 0.4068 | 19.78 | 0.3707 | 21.90 | 0.3972 | **15.85** | **0.2866** |

each speaker [101]. This method thus performs the phonetic compensation and is found to be helpful for dealing with limited data involving short utterance based SV as mentioned in [101].

The proposed KDA based compensation technique uses a Gaussian kernel for learning the non-linearity in data. The train, test and the development data i-vectors are mean and length normalized. The mean and length normalization is performed on the i-vectors to deal with the non-Gaussian behavior of the i-vectors as mentioned in [128], which helps in improving performance. The subspace for transformation is leaned using the mean and length normalized development data i-vectors as given in Algorithm 1.During training and testing, the mean and length normalized train and test i-vectors are transformed into the learned subspace. Finally, the testing is made by performing cosine kernel between the transformed train and test i-vectors. The studies conducted for this work are reported in the next subsection.

### 5.2.4 Studies and analysis of results

Table 5.1 reports the performance in terms of EER and DCF obtained for full and truncated limited data based short test utterances (2-10 s). The second column of the table shows the performance with the classical baseline method of LDA followed by WCCN on the i-vector based SV system developed over NIST SRE 2003 database. Then its comparison to the LDA-GPLDA based setup can be seen from the third column. The dimension of LDA is taken to be 150, which is found to give the optimal performance over the database. Similarly, GPLDA of 150 dimensions is considered for the study. It is observed that the LDA followed by GPLDA is better capable of handling channel/session compensation than the conventional LDA followed by WCCN approach. The SUVN technique is implemented using the short duration development data i-vectors for different duration of 2-10 s and then applied on top of the i-vectors as mentioned in [101]. As this method is purely for short utterance variance compensation, hence the studies are made only on the short duration limited test data cases. The

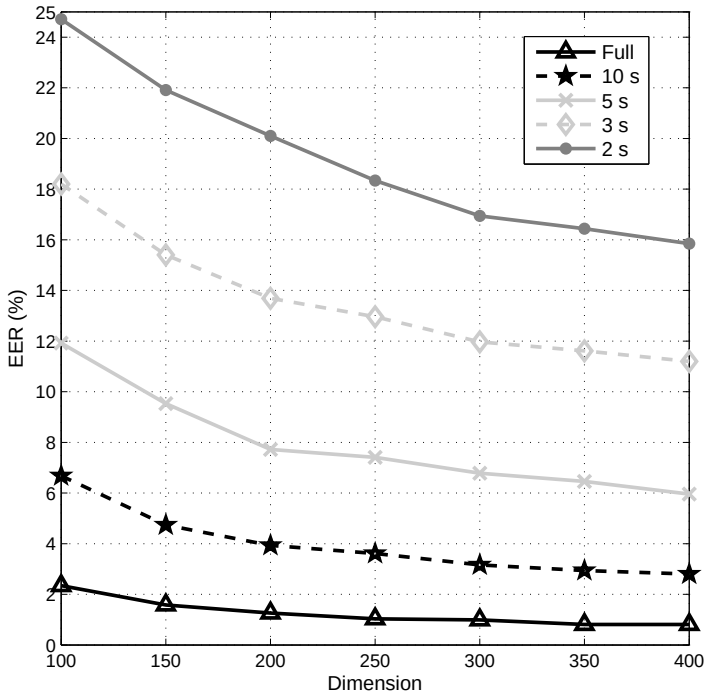


**Figure 5.3:** EER trends for different durations of test utterance and for different values of dimension in the KDA based framework.

fourth column of Table 5.1 reports the results associated with SUVN technique. It shows that the performance of SUVN based compensation is comparable to that with LDA-GPLDA based framework under 10 s and 5 s cases. However, the former outperforms the latter for limited test data cases of 3 s and 2 s showing its capability of handling the SV under limited test data scenario.

The proposed framework using KDA at the back-end is then implemented to observe its effectiveness with respect to the other methods. Initially, the KDA dimension is fixed at 150 with reference to the optimal dimension selection of LDA. Then performance is evaluated for full as well as the limited test data conditions. On comparing to the existing setups, the proposed framework is found to dominate the results. This justifies the considered hypothesis that the information such as speaker, channel and session are non-linearly separated in the i-vector domain, limiting the linear compensation approaches. Further, KDA based framework projects the i-vectors into a higher dimensional subspace and then perform discriminant analysis. Thus it is expected that the higher dimensions of KDA may be beneficial for channel/session compensation. In this regard, SV studies are extended by changing the dimension of KDA based compensation technique to observe its impact along with choosing the optimal dimension for KDA. Figure 5.3 shows the trend in performance with respect to the different

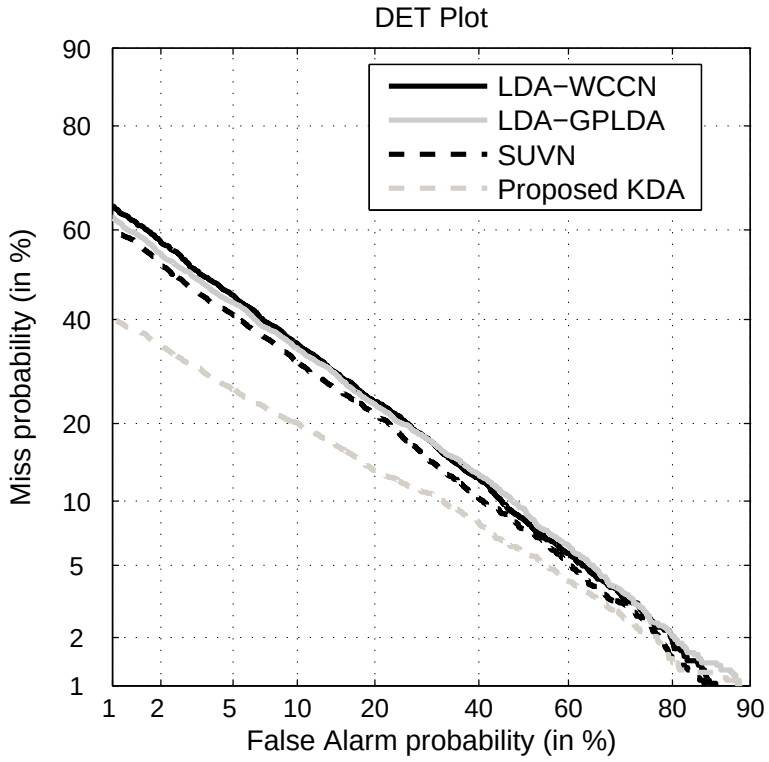


**Figure 5.4:** DET plots of different techniques with the proposed KDA based framework for 2 s of test data.

dimensions of KDA for full length and limited test data conditions. The figure clearly indicates the significant improvement obtained with the higher dimensions of KDA in the proposed framework. Although the performance with LDA/GPLDA is optimal at 150 dimensions, the KDA based results are observed to improve as the dimension is increased and found to give best around 400 dimensions. The explanation for this may be seen as KDA is capable of forming non-linear manifolds for projection unlike in the case of LDA. Hence it is able to separate necessary information and noise without much reduction in dimension. Additionally, the absolute improvements are more distinctive as the duration of test utterances are reduced. This strengthens the claim of non-linear separability of i-vectors with more mismatch in the phonetic content for limited test data and proves the capability of KDA in the i-vector based SV framework.

Figure 5.4 shows the detection error tradeoff (DET) plots for different methods and their comparison to the KDA based framework for 2 s test speech condition. This further highlights the efficacy of the proposed method over the other approaches for having better compensation in i-vector based speaker modeling. Hence the KDA based framework can be used as a prospective approach for SV under limited test data scenario.

**Table 5.2:** Performance for the fusion of VTC and MFCC features over i-vector based SV framework.

| Test Duration | MFCC | | MFCC + VTC | |
|---|---|---|---|---|
| | EER (%) | DCF | EER (%) | DCF |
| 10 s | 5.81 | 0.1090 | 5.56 | 0.1031 |
| 5 s | 10.52 | 0.1977 | 10.43 | 0.1937 |
| 3 s | 16.94 | 0.3100 | 15.94 | 0.2929 |
| 2 s | 22.31 | 0.4128 | 21.73 | 0.4017 |

## 5.3 Proposed combined framework

In this section, a combined framework is proposed with the different explorations made with respect to SV with limited test data condition. The results obtained from the previous chapters related to different attempts are briefly revisited along with KDA based technique discussed in this chapter. Then they are tied with a common thread in an SV architecture to have improved speaker characterization under limited test data based scenario.

### 5.3.1 Different explorations for speaker verification with limited test data

In order to have a common framework using the explorations made for limited test data based scenario, the works with respect to each of these explorations are mentioned briefly.

#### 5.3.1.1 Speaker-specific vocal tract constriction information

The VTC feature is found to have speaker-specific phonetic information, which is useful from the perspective of capturing definite speaker characteristics. This feature is of single dimension as discussed in Chapter 4 and is fused with MFCC features at the feature level for developing the i-vector based SV system. Table 5.2 shows the performance of VTC feature in fusion to the MFCC based conventional vocal tract features and its comparison to the baseline results obtained with MFCC features for SV with limited test data. This clearly indicates the significance of speaker-specific phonetic information for improved speaker modeling.

#### 5.3.1.2 Different attributes of source information

The features mel power difference of spectrum in subbands (MPDSS), residual mel frequency cepstral coefficient (RMFCC) and discrete cosine transform of integrated linear prediction residual (DCTILPR) are found to capture different attributes of excitation source information as discussed

**Table 5.3:** Performance for different source features and their fusion under limited duration test segments over i-vector framework.

| Test Duration | DCTILPR | | RMFCC | | MPDSS | | Source Fusion | |
|---|---|---|---|---|---|---|---|---|
| | EER (%) | DCF | EER (%) | DCF | EER (%) | DCF | EER (%) | DCF |
| 10 s | 13.91 | 0.2497 | 12.96 | 0.2466 | 17.43 | 0.3269 | **10.57** | **0.1964** |
| 5 s | 18.65 | 0.3460 | 18.88 | 0.3462 | 22.58 | 0.4250 | **11.97** | **0.2252** |
| 3 s | 22.13 | 0.4077 | 23.62 | 0.4362 | 27.60 | 0.5202 | **15.85** | **0.2854** |
| 2 s | 27.78 | 0.5198 | 27.55 | 0.5203 | 31.44 | 0.5958 | **20.19** | **0.3759** |

in Chapter 3. The MPDSS feature captures the periodicity nature, whereas the RMFCC feature contains the smoothed spectrum information. Similarly, the DCTILPR feature possesses the shape of glottal signal as investigated earlier. As the three source features contain different aspects of excitation source, each of them is important for extracting speaker characteristics. Thus each of them is considered for the development of parallel SV systems and their performances are reported in Table 5.3. Additionally, the fusion of the source features is carried out at the score level as explained in Chapter 3 and their combined performance can be viewed from the last column of Table 5.3. This shows that for limited data of very short test utterance based cases the results obtained with the fusion of three source features outperform the baseline results obtained with MFCC features. This signifies the the importance of the different attributes of excitation source features for SV with limited test data based scenario.

#### 5.3.1.3 Kernel based discriminant analysis

The KDA based framework at the back-end of i-vector based SV system, explored in this chapter has shown promising results. Further, its impact is found to be more towards limited test data based studies. KDA projects the i-vectors into a higher dimensional space and then performs discriminant analysis, which helps for better classification with reference to the i-vectors of the limited data. In this regard, after utilizing the VTC and different attributes of excitation source information in terms of three voice source features at the feature domain, it is considered for employing in each of the cases expecting some benefit for verification of a trial.

The KDA based framework is implemented with the development data i-vectors and applied for channel/session compensation on top of train and test i-vectors. This approach is evaluated for NIST SRE 2003 database with reference to the MFCC+VTC based features and the three source features considered in this work. Table 5.4 reports the performance for each of the features considered for

**Table 5.4:** Performance for different features under limited duration test segments over i-vector framework with KDA as back-end.

| Test Duration | MFCC+VTC | | DCTILPR | | RMFCC | | MPDSS | |
|---|---|---|---|---|---|---|---|---|
| | EER (%) | DCF | EER (%) | DCF | EER (%) | DCF | EER (%) | DCF |
| 10 s | 2.78 | 0.0480 | 10.16 | 0.1684 | 8.58 | 0.1602 | 14.04 | 0.2385 |
| 5 s | 5.88 | 0.1056 | 14.00 | 0.2345 | 12.65 | 0.2300 | 17.91 | 0.3114 |
| 3 s | 11.07 | 0.1942 | 18.47 | 0.3145 | 16.89 | 0.3091 | 22.45 | 0.4136 |
| 2 s | 15.49 | 0.2858 | 23.85 | 0.4319 | 20.51 | 0.3804 | 27.06 | 0.5079 |

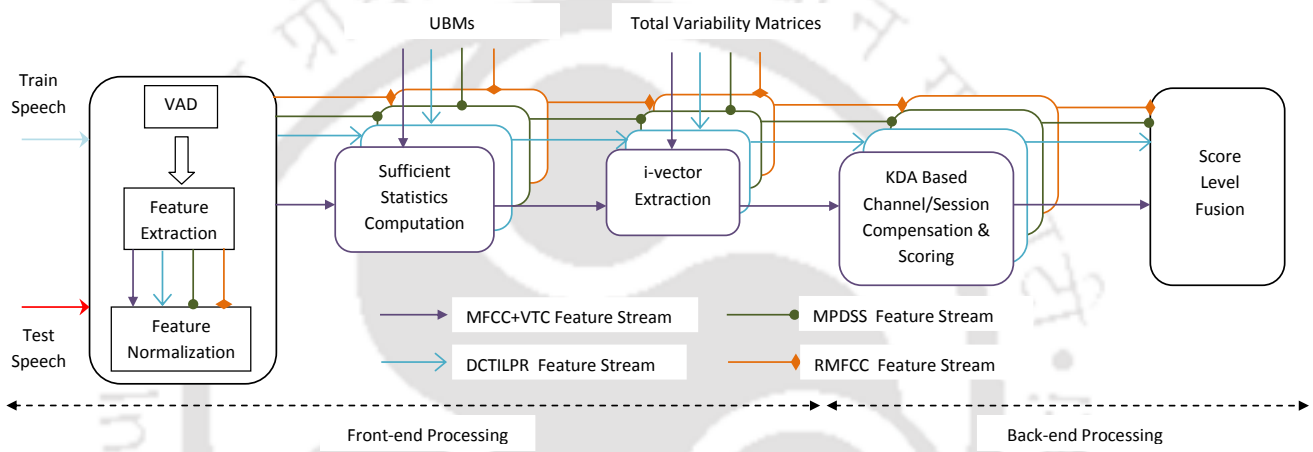


**Figure 5.5:** The block diagram of the proposed combined framework for dealing with SV using limited test data involving different explorations.

limited test data based SV. The use of KDA at the back-end of the i-vector based speaker modeling provides an improved performance than the conventional techniques. The improvements are achieved mainly because KDA transforms the features into higher dimensional space. Thereby they become better discriminative, which is useful for the i-vectors of limited test data as they are poorly estimated due to less availability of features from the speakers.

### 5.3.2 Combined framework: studies and results

This subsection discusses the combined framework involving different explorations made for limited test data based SV for practical systems. The previous subsection explained the works attempted in order to achieve an improved performance when there is a very small amount of data available for authentication of a trial. These directions include alternative features capturing speaker characteristics in addition to the conventional MFCC features, suitable pattern recognition approaches for handling limited data scenario, etc. Figure 5.5 shows a block diagram of the proposed framework that integrates

**Table 5.5:** Performance for fusion of different features under limited duration test segments over i-vector framework with KDA as back-end.

| Test Duration | Source Fusion | | Source+MFCC+VTC | |
|---|---|---|---|---|
| | EER (%) | DCF | EER (%) | DCF |
| 10 s | 6.50 | 0.1104 | **2.48** | **0.0455** |
| 5 s | 9.67 | 0.1667 | **4.47** | **0.0778** |
| 3 s | 12.92 | 0.2281 | **7.36** | **0.1319** |
| 2 s | 17.43 | 0.3269 | **11.20** | **0.1990** |

the various directions for limited test data scenario into a common platform. It shows that given a speech signal, its features are extracted with different features MFCC, VTC, DCTILPR, MPDSS and RMFCC. They are then used over the i-vector based speaker modeling to obtain the compact representation in the form of an i-vector. As the VTC feature is of single dimension, it is combined to MFCC features at the feature level as discussed earlier. Then a single SV system is developed for them. However, individual SV systems based on the source features DCTILPR, MPDSS and RMFCC are developed over i-vector based speaker modeling approach. The KDA is used at the back-end of the systems developed for different features over i-vector based speaker modeling to compensate the channel/session information. It is followed by cosine kernel scoring between the test and the claimed model i-vector to obtain the similarity score. Finally, a score level fusion is made similar to that explained in Chapter 3 for SV systems developed using different features to generate a single score for verification of an identity claim.

Table 5.5 reports the performance under score level fusion of different systems. It shows that the fusion of the three source features achieves significant improvement than that obtained with each of them. The combined score from these source features is further fused with that obtained from MFCC including VTC feature based system to yield a final fused score. The performance of the same is reported in the last column of Table 5.5. It indicates commendable improvement achieved with the proposed framework based on speaker-specific phonetic information, different/complementary features, pattern recognition technique for SV with limited test data. Figure 5.6 illustrates the DET trends for different studies under this work. It further highlights the efficacy of the combined framework developed using various explorations for 2 s case based limited test data study. Additionally, the histograms of genuine and impostor scores for different explorations and their combinations are depicted in Figure 5.7 for 2 s test data case. It clearly illustrates that the separability of genuine

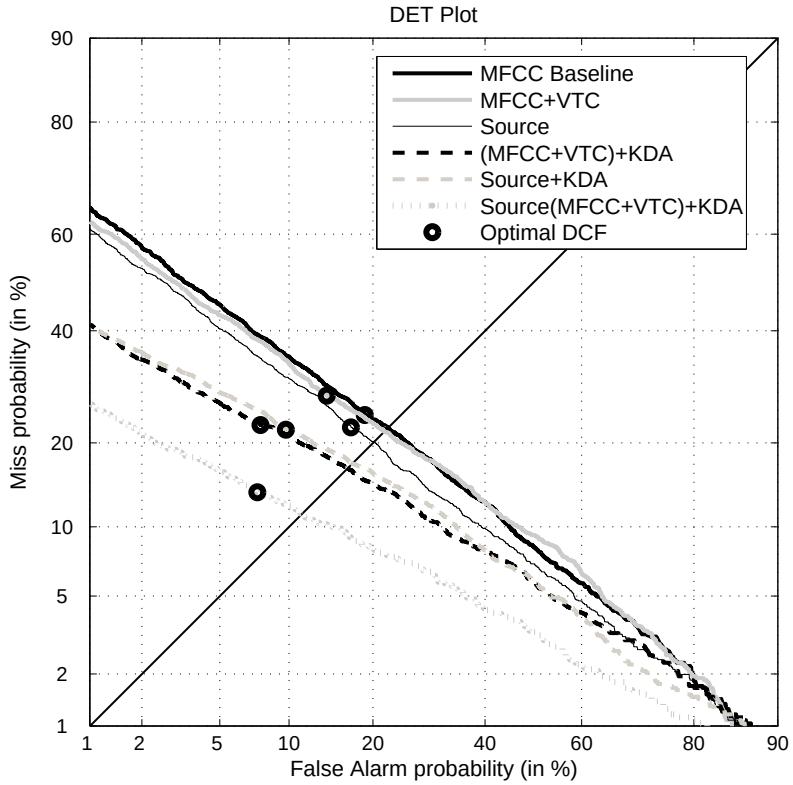


**Figure 5.6:** DET plots for different explorations and their combination for 2 s test data case.

**Table 5.6:** Extent of overlap of genuine and impostor score histograms indicating better separability with proposed framework for 2 s test data case.

| Different Explorations | Overlap (%) |
|---|---|
| MFCC | 42.90 |
| MFCC+VTC | 42.39 |
| (MFCC+VTC)+KDA | 27.81 |
| Source | 39.09 |
| Source+KDA | 28.39 |
| Source+(MFCC+VTC)+KDA | **20.27** |

and impostor scores enhances with each of the studies made for handling SV with limited test data. Table 5.6 quantifies the area of overlap for genuine and impostor scores for each of the cases considered for the histogram plots in Figure 5.7. The proposed framework based on the combination of different explorations is able to boost the separation of genuine and impostor scores. Thereby minimizing their area of overlap by a large margin from that of the baseline system.

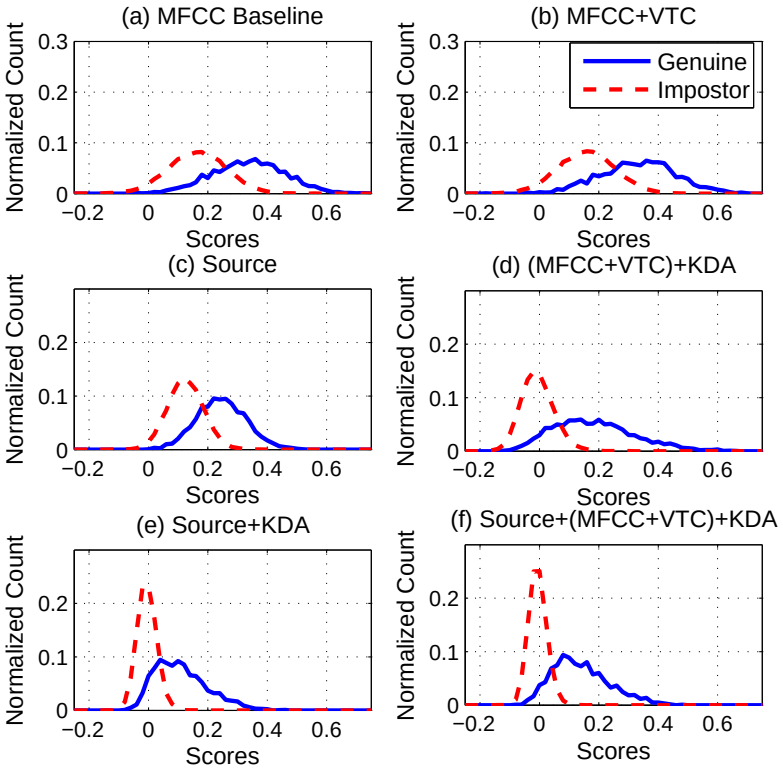


**Figure 5.7:** Distributions of genuine and impostor scores for different explorations and their combinations for 2 s test data case.

## 5.4 Summary

This chapter primarily deals with a contribution at the back-end of an SV system in terms of implementing kernel based discriminant analysis. The proposed KDA based channel/session compensation approach transforms the data points in some higher dimensional space. Then it performs discriminant analysis, where the classes are more separable. Its usage in the i-vector based SV framework provides significant improvement when compared to the baseline and other existing techniques. Further, its impact is more dominant for limited test data based scenario, thereby projecting its importance for the thesis objective. Later in the chapter, a framework is proposed with inclusion of explorations made in the former chapters based on speaker-specific phonetic information, different attributes of source information along with KDA explored in this chapter. The proposed combined framework developed using these explorations provides improved speaker characterization. Thus it shows the scope for practical SV systems under limited test data scenario.

# 6

# Investigating Different Issues for Practical Systems

Contents

### Overview

*This thesis motivates a problem statement based on speaker verification (SV) using sufficient train and limited test data from the perspective of practical systems. The previous chapters discuss regarding different directions to provide an improved performance in such a scenario and finally a common framework, including those directions. However, when we go for a field deployable system having limited data conditions, there arises many other issues in practical settings. These are quite crucial for having a robust SV system for fitting into some application. One such issue is the mismatch of speaking rate between train and test sessions. In this chapter, this issue is investigated and a framework for dealing with this issue is proposed based on prosody modification of the test speech in terms of duration modification with respect to that of the claimed speaker model speech. Another issue is the effect of speaker information when there is a large session variability. The chapter later focuses on exploring issues related to session variability and template aging and their impact in a limited data scenario from the view of application oriented systems.*

## 6.1 Duration modification for mismatch speech tempo conditions

The widespread use of technologies has given the scope for implementation of speech based person authentication systems. There have been many efforts for development of such systems that have proven to be path breaking in terms of field deployable systems. Some of them have acclaimed recognition in the scientific community [47–49]. However, the performance of many such systems is dependent under controlled test conditions. The reliability of SV systems tends to decrease severely under mismatched train-test conditions which may be in terms of language, duration, content, etc.

The speaking rate of a person depends on many factors in a practical environment, which may vary from one session to another. A mismatch in speech tempo of test speech from train speech can degrade the SV performance. This is because the speaker characteristics changes, which may be more crucial for short segments of speech used for test sessions. The studies conducted in [149] have shown that speaking rate has a strong impact on SV performance. The change in speaking rate results in variation in phone duration, aspiration, transient nature of fast speech spectra, which leads to a mismatch in the feature set corresponding to train and test. In [150], the authors proposed a speech rate classifier based on the dynamic coefficients of the feature vectors, suitable for real-time application of a speech recognition system. There are some attempts made to improve the performance

for speech with fast speaking rate by implementation of Baum-Welch codebook adaptation, adapting hidden Markov model (HMM) state transition probabilities and rule based pronunciation modification dictionaries [151]. A novel probabilistic method to estimate the speaking rate is proposed in [152], which selects the recognition model of suitable speaking rate useful for a speech recognition task. This provided the motivation for exploring the speaking rate mismatch condition for SV. From the view of application oriented systems, a mechanism to detect the mismatch in speech tempo may be helpful for compensating the same. In this regard, duration modification for modifying the speaking rate can act as a useful approach.

Duration modification, which is one of the aspects of prosody modification can be used either to slow down or speed up the speech signal. The prosody modification methods can be either in time domain or frequency domain. Some of the time-domain prosody modification approaches are overlap and add (OLA), synchronous overlap and add (SOLA) and pitch synchronous overlap and add (PSOLA) [153, 154]. Different variants of PSOLA based on the principle used are the time-domain PSOLA (TD-PSOLA), frequency domain PSOLA (FD-PSOLA) and linear predictive PSOLA (LP-PSOLA). The authors of [155] report that, prosody modification using instant of significant excitation performs better in the terms of spectral and phase distortion compared to TD-PSOLA. However, apart from the quality of prosody modified speech signal, the time constraint and simplification of the modification algorithm are also some important factors. A computationally simpler method in [156] employs epoch based prosody modification method achieving comparable performance as in [155]. This method with lower time complexity can be useful for different speech based real-time application oriented systems. Hence this motivated us to consider the same for modification of speaking rate under mismatch speech tempo condition for an SV system, where the decision on the fly has to be taken for verification of a trial.

This work proposes a novel framework that first identifies whether there exists a mismatch in speech tempo between the test and the train session. If there exists a mismatch, then it modifies the speaking rate of the former according to that of the latter by faster prosody modification approach. This mismatch in speech tempo is found by detecting the speaking rate of train and test utterances computed using excitation source information based evidence. The prosody modified test speech, that compensates the mismatch in speech tempo is taken for verification of a claim. The SV performance is expected to be better when the speech tempo corresponding to train and test speech are made same.

The novelty of this work can be viewed as bringing out the framework for speaking rate modification to compensate the mismatch in speech tempo, which is favorable for practical systems. Further, the studies related to mismatch in speech tempo are carried out under limited test data for exploring its impact in such a scenario.

The rest of the work in this section is organized as follows. Section 6.1.1 describes the faster prosody modification algorithm and its significance. In Section 6.1.2, the development of the baseline SV system and an analysis is provided for studies under mismatch speech tempo conditions. Section 6.1.3 provides the details of the proposed framework of duration modification under mismatch speech tempo conditions. The experimental results and analysis are presented in Section 6.1.4.

### 6.1.1 Faster prosody modification

This work concentrates on duration modification in terms of speaking rate for real-time SV systems, for which the time constraint is a crucial factor apart from having less spectral and phase distortion. To achieve lower computational complexity compared to [155], the faster prosody modification method proposed in [156] is employed for this framework. In the epoch based prosody modification method, the accuracy of deriving epoch locations also plays an important role. Hence, in [156], the epochs are extracted using zero frequency filtering (ZFF) based method, which has the potential to detect the epoch locations accurately from the speech signal [130]. The key steps involved in most of the prosody modification methods can be viewed as,

- Derive the epoch locations from the given speech segment.
- Derive modified epoch locations according to the given prosody modification factor.
- Modify the speech segment by considering modified epoch locations as anchor points.

In the above mentioned steps, the modified epoch locations are derived by resampling the original epoch locations. Then it searches for the nearest corresponding epoch in the original epoch location sequence [155]. This process of finding the nearest epoch in the original epoch sequence for each epoch in the resampled sequence is time consuming and involves a number of iterations. This leads to an increase in time complexity.

In [156], time complexity for prosody modification is reduced by simplifying the procedure for deriving modified epoch locations. As shown in Figure 6.1, simple time scaling of original epoch locations is performed according to the given modification factor Beta ($\beta$) instead of deriving modified epoch locations. Using the knowledge of these time scaled epoch locations, the original speech samples

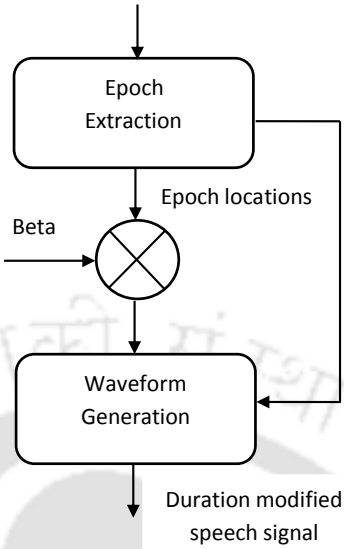


**Figure 6.1:** Block diagram showing faster prosody modification process.

are copied into a new array in a pitch synchronous way to derive the prosody modified speech. Based on the knowledge of epochs and given duration modification factor, group of speech samples corresponding to a particular pitch period is skipped to make the speech faster. For slowing down the speech signal, the group of speech samples corresponding to a particular pitch period is repeated a number of times based on the duration modification factor. It has been found that this method by employing ZFF for epoch extraction and time scaling of epochs makes the prosody modification process faster by 56.25% compared to the existing approaches keeping the required quality intact [156]. It thus makes this approach favorable for integrating in a framework of practical SV systems that involves low time complexity for deployment.

### 6.1.2 Exploring speaker verification under mismatch speech tempo conditions

This subsection describes the development of baseline SV system and the studies under mismatch speech tempo conditions.

#### 6.1.2.1 Database, preprocessing and feature extraction

The database considered in this work is NIST SRE 2003, that contains a population of 356 speakers as discussed in the previous chapters [112]. Switchboard Corpus-2 is considered as the development database for learning the background models. The speech utterances are processed in terms of blocks of 20 ms with a shift of 10 ms and 39-dimensional (13-base + 13$\Delta$ + 13$\Delta\Delta$) mel frequency cesptral

**Table 6.1:** Performance of the baseline SV system over i-vector based modeling.

| EER (%) | DCF |
|---|---|
| 1.94 | 0.0338 |

coefficient (MFCC) features are extracted for each of the Hamming windowed frame. Voice activity detection (VAD) is performed based on the energy of the speech signals. Then only the speech regions are considered upon which cepstral mean and variance normalization (CMVN) is carried out [25].

#### 6.1.2.2 Development of i-vector based baseline framework

The i-vector based SV framework as described in Chapter 3 is chosen for building the SV system [32]. A 1024 component universal background model (UBM) is built using the development data features. The sufficient statistics of the development, train and test features are extracted using the UBM and then a total variability matrix (T-matrix) of 400 columns is trained using the development data statistics. The i-vectors are then extracted using this T-matrix for train, test and development data. 150-dimensional linear discriminant analysis (LDA) followed by probabilistic linear discriminant analysis (PLDA) is used at the back-end for channel/session compensation and scoring [33, 128]. The trials are made as per the evaluation plan of NIST SRE 2003. Table 6.1 shows the performance of the baseline system in terms of equal error rate (EER) and detection cost function (DCF).

#### 6.1.2.3 Studies under mismatch speech tempo conditions

To study the effect of speaking rate on SV, an experimental setup is designed. Under this, the duration modification is made for the test speech using faster prosody modification technique. The duration of the test speech is modified for various factors and SV system is evaluated using the i-vector based framework for each case to observe the trend. Table 6.2 shows the performance of the SV system when the speaking rate is made faster compared to that of the train speech by different factors from 0.8 to 0.4 in the intervals of 0.2. Mismatch in speech tempo beyond 0.4 times faster speaking rate is not considered as it does not have any significance under practical scenario. It is observed that the performance tends to fall exponentially as the mismatch in speech tempo is more. Additionally, another set of experiments are conducted by making the speaking rate of the test speech slower to that of the train speech. The results associated with this study are reported in Table 6.3. Although, there is a performance degradation observed under this case, it is comparatively less to that observed

**Table 6.2:** Performance under mismatch speech tempo for faster test speech conditions.

| Mismatch Factor ($\beta$) | EER (%) | DCF |
|---|---|---|
| 0.8 | 2.53 | 0.0421 |
| 0.6 | 3.30 | 0.0595 |
| 0.4 | 5.56 | 0.1019 |

**Table 6.3:** Performance under mismatch speech tempo for slower test speech conditions.

| Mismatch Factor ($\beta$) | EER (%) | DCF |
|---|---|---|
| 1.2 | 2.12 | 0.0388 |
| 1.4 | 2.21 | 0.0403 |
| 1.6 | 2.39 | 0.0430 |
| 1.8 | 2.66 | 0.0484 |

for faster test speech. Thus, the studies confirmed that the speaking rate has a great impact in the performance of SV systems. This motivated for developing a framework to detect the mismatch in speech tempo of the test to the train speech on which prosody modification in terms of duration modification may be useful for compensation.

### 6.1.3 Proposed framework of duration modification for mismatch speech tempo

The architecture of the proposed framework may be viewed from Figure 6.2 that is aspired for having a practical system, where the test conditions are uncontrolled. As Figure 6.2 shows, first the mismatch in speech tempo of test speech from that of the train speech is checked, for which it is necessary to determine the speaking rate of the test utterance precisely. The speaking rate of an utterance may be defined as number of syllables spoken per second [157]. The accuracy of speaking rate detection algorithm depends on the robustness of the method employed for syllable nucleus detection. Basically, the vowels form the syllable nucleus with its prominent formant structure with respect to syllable onset or coda. Considering the application in practical environments, a speech rate algorithm is developed for this purpose. Two significant attributes of speech, Hilbert envelope of linear prediction residual (HE-LPR) and zero frequency filtered signal (ZFFS) are used, that have proven their effectiveness for detection of vowel-like regions (VLRs) under degraded condition [22]. The large errors in LPR due to significant excitation are manifested as peaks in the HE-LPR. Since syllable nuclei are for longer duration with sharp glottal closures, so maximum value over a 5 ms

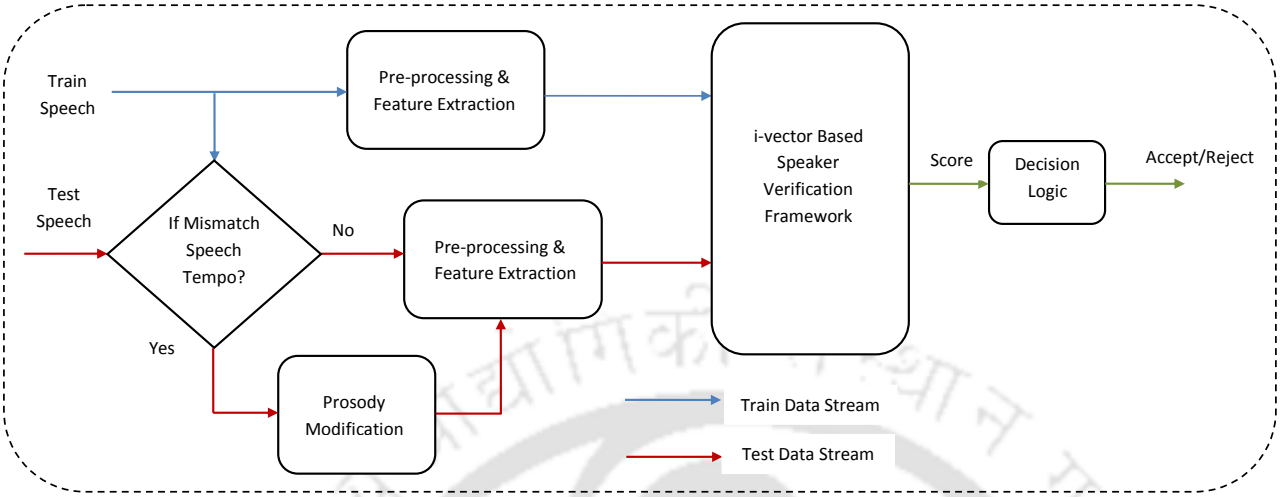


**Figure 6.2:** Block diagram of the proposed framework for SV under mismatch speech tempo condition.

window with one sample shift of HE-LPR is considered. This is expected to have a better correlation with syllable structure. The corresponding contour is shown in Figure 6.3 (b) for the speech signal of Figure 6.3 (a), which is taken from the NIST SRE 2003 database over telephone channel. The slopes of positive zero crossings, i.e. the first order derivatives of ZFFS are proven to represent the strength of excitation, which is an important cue from the perspective of speech production [158]. Thus, its second order derivatives represent the change in strength of excitation which, is depicted in Figure 6.3 (c). The combined smoothed evidence is shown in Figure 6.3 (d), where the peaks correlate well with syllable nucleus position compared to that of the individual evidence as can be seen by comparing to the dotted lines. By using a peak detection algorithm with a proper threshold of peak-to-dip ratio and distance between the peaks, the syllable nuclei can be detected automatically for a given speech utterance and thereby the speaking rate. This method is validated over a set of speech segments from the NIST SRE 2003 database considered for this work.

The difference in speaking rate of the test speech with respect to the train speech is found as mentioned above. This is followed by computation of the mismatch factor, which is taken as the prosody modification factor ($\beta$). The speaking rate of the test speech is then modified with respect to that of the train speech of the claimed speaker using the faster prosody modification process to have the compensation in mismatch speech tempo. The remaining framework for SV is similar to that explained for the baseline framework for verification of a trial using i-vector based speaker modeling.

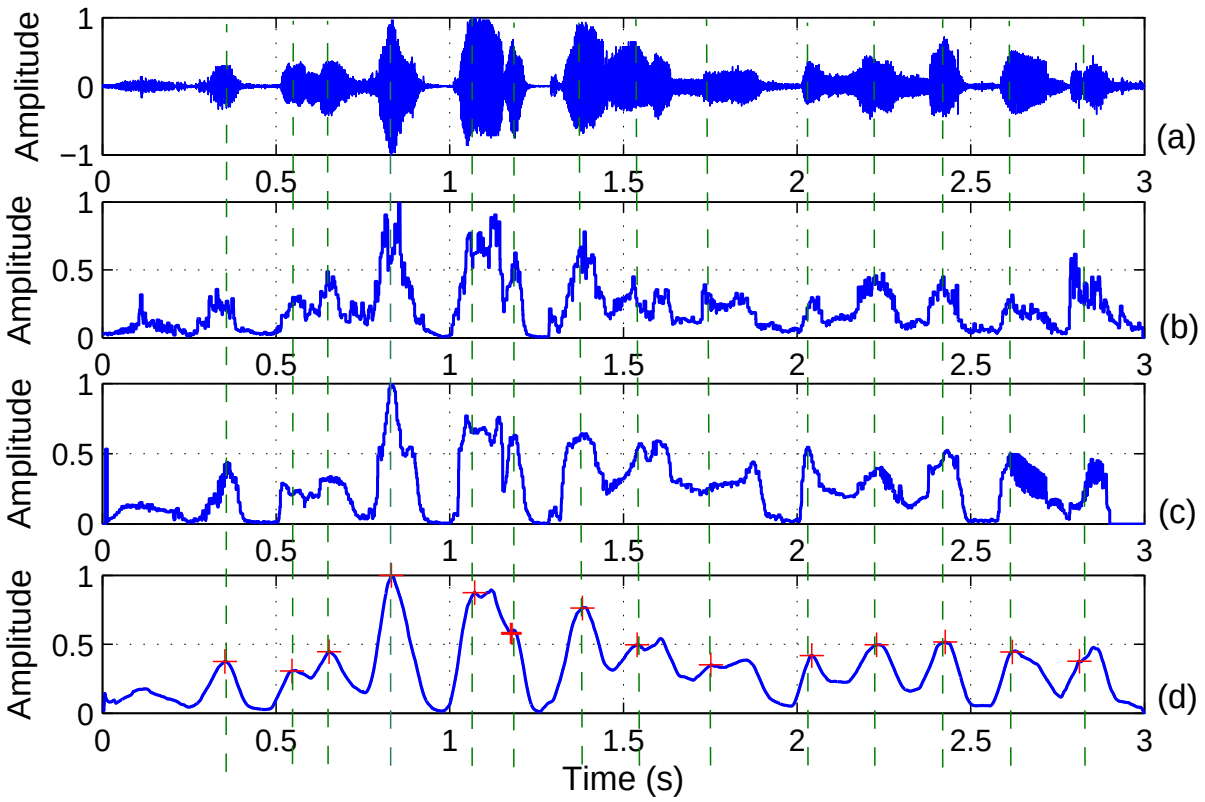


**Figure 6.3:** (a) Speech segment (b) HE-LPR evidence (c) Evidence from ZFFS (d) Combined evidence with detected syllable nucleus.

### 6.1.4 Experimental studies and analysis

To study the effectiveness of the proposed framework, another experimental setup is designed, where the train speech examples of the speakers are kept at different speaking rate. Each of these cases is then considered for speaker modeling. Unlike the previous study, for this current study the speaking rate of the test speech is modified with reference to that of the train speech by faster prosody modification approach and the SV experiments are conducted. This in turn will signify the impact of duration modification under a match case of speaking rate. The observations are first made under faster speaking rate, the results of which can be seen from Table 6.4 for matched speaking rate of train and test set. On comparing it to the results obtained under mismatch condition from Table 6.2 significant improvements are visible. Further, the results of the studies under the proposed framework for slower speaking rate under match condition are mentioned in Table 6.5. This also shows some amount of improvement from the results under mismatch case given in Table 6.3. Figure 6.4 shows the comparison of SV performance in terms of EER for different speaking rate of speakers under match

**Table 6.4:** Performance under duration modified speech for faster speaking rate under train-test match conditions.

| Prosody Modification Factor ($\beta$) | EER (%) | DCF |
|---|---|---|
| 0.8 | 1.97 | 0.0354 |
| 0.6 | 2.30 | 0.0427 |
| 0.4 | 3.21 | 0.0608 |

**Table 6.5:** Performance under duration modified speech for slower speaking rate under train-test match conditions.

| Prosody Modification Factor ($\beta$) | EER (%) | DCF |
|---|---|---|
| 1.2 | 2.08 | 0.0380 |
| 1.4 | 2.08 | 0.0379 |
| 1.6 | 2.03 | 0.0373 |
| 1.8 | 2.01 | 0.0365 |

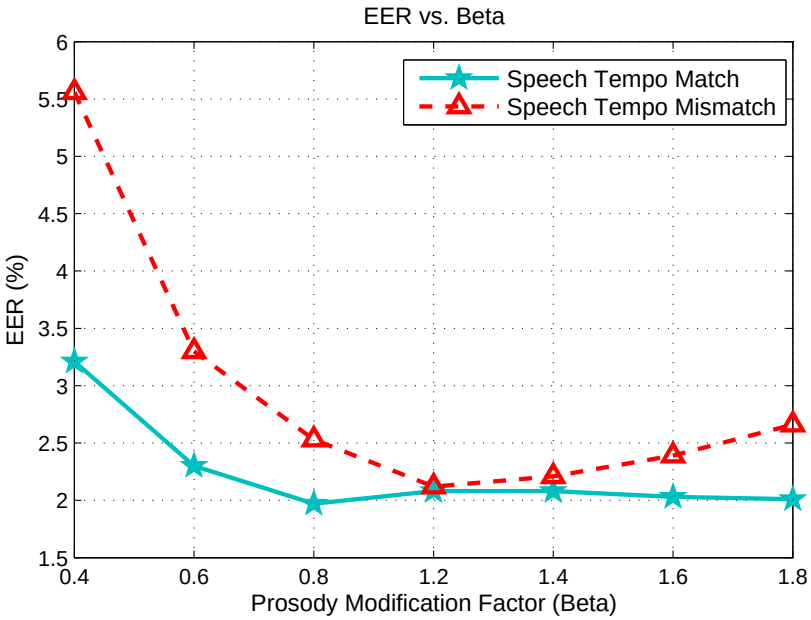


**Figure 6.4:** EER vs. Beta ($\beta$) trend for sufficient test data condition.

and mismatch conditions. It is observed that the improvements are much more prominent when the speech is spoken with faster speaking rate to that of the original speech. On the other hand, the same is comparatively lower for the slower speaking rate case.

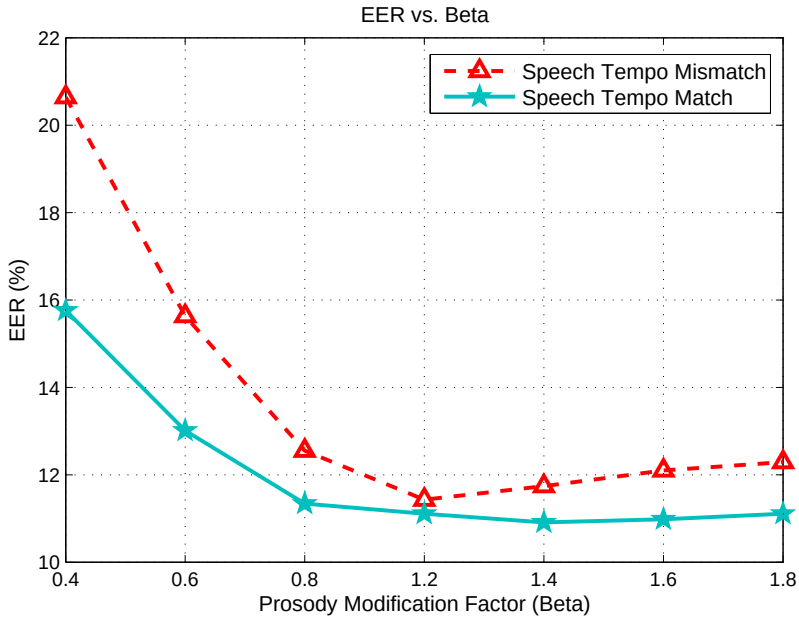


**Figure 6.5:** EER vs. Beta ($\beta$) trend for 5 s test data condition.

From the perspective of practical application oriented systems, limited test data is preferred, which is discussed in the previous chapters of the thesis. In this regard, SV with sufficient train and limited test data based scenario is explored under mismatch speech tempo condition. Limited test data case of 5 s is considered for the study and is kept at different speaking rate as done for the earlier studies under sufficient test data. Additionally, the duration modification is applied for having a match in speech tempo of the train and the test speech to observe the trend for each of the cases considered. Figure 6.5 shows the trends of EER for mismatch and match condition of speech tempo for limited test data of 5 s duration. This shows that the mismatch in speech tempo based issue also has an impact under limited test data scenario, which can be compensated by the proposed framework having duration modification. The mismatch in speech tempo condition is more likely in the range 0.6 to 1.6 of the speaking rate in a practical scenario as the speaker can have a mismatch by speaking in the stated range. In this range, if we observe the performance trend for sufficient train and limited test data condition, then it shows that it is almost similar for faster speaking rate. However, the effect is more for speech with slower speaking rate. This may be because when dealing with limited test data, a slight variation of speaking rate affects the performance as only a small portion of speaker's speech is available for validation of a claim. On the other hand, in case of sufficient test condition when slow speech is considered then that does not affect much as in the case of limited test data.

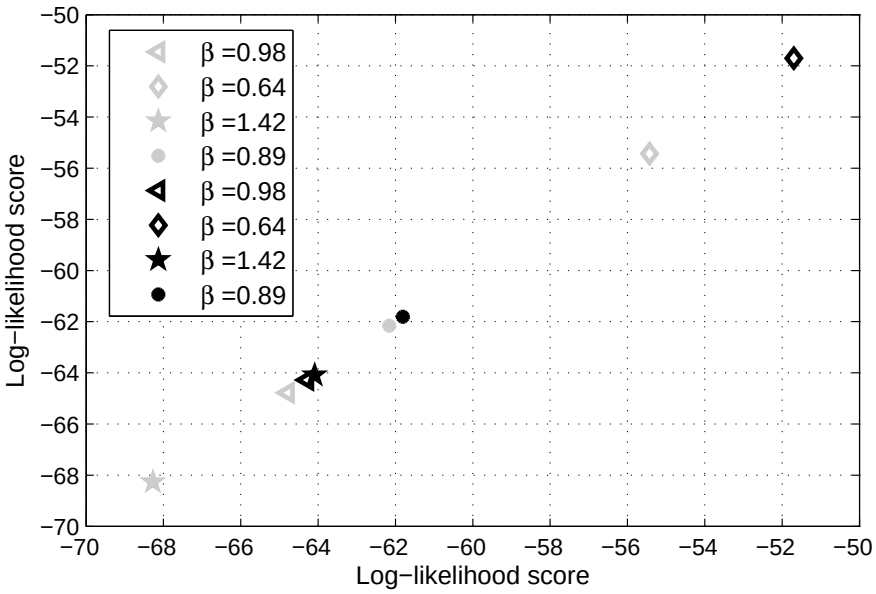


**Figure 6.6:** Log-likelihood score trend: The gray color denotes log-likelihood scores of test speech without duration modification and the corresponding black color denotes the same after duration modification for different mismatch factors ($\beta$).

In order to validate the significance of duration modification in a practical database, few examples of a small set of speakers at varying speaking rate are collected for a fixed phrase short utterance during their test sessions. Each speaker has three examples of the short fixed phrase for training the speaker models. The MFCC features are extracted for the phrases from train session and an 8-component Gaussian mixture model (GMM) is built for each speaker. The GMM based framework is used here for simplicity in validating the impact of duration modification for practical data. Given a test speech having a mismatch in speaking rate from the train examples, its MFCC features are extracted without and with duration modification. Then the corresponding log-likelihood is calculated with respect to the speaker model. Figure 6.6 shows the trend of log-likelihood scores denoting higher scores for duration modified speech that compensates the mismatch in speaking rate. Further, it can be noted that the impact of duration modification is more, when there is a large mismatch in speaking rate. This henceforth signifies the impact of speaking rate match for SV, which is pivotal for practical application oriented systems.

This section advocates the mismatch in speech tempo as one of the crucial factors of a practical system. Then it proposes a novel framework for compensating the mismatch in speech tempo based on faster prosody modification method. The studies are reported on simulated as well as a practical database, which show the significance of the proposed framework.

## 6.2 Session variability and template aging for speaker verification

The current achievements in the field of SV have found wide spread use in various application oriented services. These application oriented services mainly focus on limited data based scenario for recognizing speakers due to the constraint of time involved, which can provide feasibility in field deployment as discussed earlier. However, for deployable systems with a regular use over a long period of time, the effect of session variability and template aging may result in performance degradation.

Several works have been done in the past to address the issue of session variability. In [159], the authors explicitly model session variability by generating a session dependent factor in a low dimensional subspace. The efficacy of this approach is proved for the NIST database in a text-independent framework, which clearly showed the significance of session variation for SV performance. Another way to handle this session variation is to have session compensation techniques to reduce the effect of session variability. There are different approaches for session compensation, some of which are joint factor analysis (JFA), LDA and nuisance attribute projection (NAP) [33, 35, 160]. These approaches are found to enhance SV performance by providing session compensation. The authors in [103] have made a comparison of different approaches used for compensating session variability in SV. The work reported in [104] proposes an approach based on maximum-likelihood linear regression (MLLR) adaptation that transforms multiple recognition models and phone classes for session variability normalization, which improves the SV performance. Thus, the impact of session variation is found to be very crucial for SV systems.

The aging phenomenon in different biometrics has been an interesting aspect for dealing with cutting edge technologies from a practical system point of view [161]. Considering speech biometric based systems, the aging effect of speaker models has not been addressed to a large extent. The studies of [162] carried out on 22 speakers data collected for three sessions with a gap of 1-2 months show that time lapse in test session degrades the performance to an extent. In [163], the authors have made studies on long term aging data over 18 speakers for 30-60 years span that show the genuine scores of speakers are affected severely than that of the impostor scores with the aging of the speaker templates. The work in [164] reports that the error rate doubles when the train and the test sessions have an interval of more than a month. In [165], the author conducts a study for exploring the aging effect for data collected for an interval of four years. The studies report that the amount of degradations in performance is gradually more for the trials having a larger time interval from training.

The limited exploration in the area of template aging is mainly due to the lack of availability in databases having large session variation from a sizeable population of speakers. The recently made available data as a part of RedDots project has opened the doors towards exploring template aging for fixed phrase short utterances [53]. In this current work, the effect of session variability is addressed by the creation of a speaker model with session variated three templates (first, middle and last sessions) and then testing by the remaining templates. This framework for creation of speaker models by data having a larger session variation is expected to perform better than that of the baseline due to the consideration of the session variability. Further, template aging studies are conducted with the creation of speaker models by two approaches. The first one is based on the creation of speaker models with the first three sessions and the latter is using the last three sessions. It is hypothesized that there may be a significant difference in speaker characteristics from the first three sessions to the last three sessions that is collected over a span of one year, which can be critical from the perspective of a practical system. The novelty of this work lies in investigating the effect of session variability and template aging for fixed phrase short utterances having limited data and proposing frameworks to deal with such issues to some extent. This knowledge can be utilized for a practical SV system under regular use for deployment. It is also to be noted that although the thesis focuses on SV using sufficient train with limited test data based scenario, the studies conducted for session variability and template aging are on fixed phrase short utterances. This is because the RedDots database, which is having trials with large session variation, are based on fixed phrase based SV framework. Thus these studies are reported on a scenario, where limited train and test data is involved.

The work is compiled in the following order. Section 6.2.1 explains the development of baseline SV system on RedDots database. In Section 6.2.2, the proposed framework for session variability study is reported, which is found to exploit the session variation information for speaker modeling. Section 6.2.3 highlights how the template aging characteristics can be observed on data having a large session variation. Section 6.2.4 provides a discussion on the conducted study for session variability and template aging, which is pivotal for a system with practical implementation.

### 6.2.1 Development of baseline speaker verification system on RedDots database

This subsection describes the baseline system developed on the RedDots database. The baseline framework is based on the i-vector modeling as front-end and PLDA as the back-end module for SV architecture [32, 128].

#### 6.2.1.1 Database, preprocessing and feature extraction

The RedDots database contains a population of 49 male and 13 female speakers totaling to 62 speakers having 572 sessions in total [53]. It has four different parts out of which the Part I contains 10 fixed phrases that are common for all the speakers. The speakers of Part II and Part III have 10 unique phrases, each of which are assigned and user chosen, respectively. On the other hand, the Part IV of the database contains free text phrases that are unique across different sessions. For evaluation of Part IV, there are two enrollment conditions, which are text-dependent (TD) and text-prompted. As discussed earlier in Chapter 4, the text-prompted based enrollment condition is referred to as the text-independent for keeping uniformity with the thesis framework. The sessions $2^{nd}$, $4^{th}$ and $6^{th}$ are used for speaker modeling in each category of the database, except for text-independent condition of the Part IV. In text-independent based enrollment condition, around 10 different fixed phrases are used for the creation of speaker models.

The RedDots database is collected over a span of one year. Hence there is a large session variability involved across the trials of each speaker. Figure 6.7 shows the histograms for number of sessions per speaker for the RedDots database for all the four parts. It is clearly visible that most of the speakers have more than 10-15 sessions of fixed phrase short utterances. Only, Part III of the RedDots database has relatively less number of speakers, who do not have a large number of sessions. Additionally, the database has three different trial conditions as discussed in Chapter 4. However, only *Impostor Correct* condition is considered in this study. As Part II and Part III do not have the *Impostor Correct* trial condition due to the involvement of speaker-specific fixed phrases, they are not considered here. The studies are reported considering Part I and Part IV TD category of the RedDots database.

The utterances from RedDots database are processed with blocks of 20 ms with a shift of 10 ms for every Hamming windowed frame. 39-dimensional (13-base + 13$\Delta$ + 13$\Delta\Delta$) MFCC features are extracted for each short term processed frame. Energy based VAD is performed to select the region of interest. Then the features of those regions are normalized with CMVN technique [25].

#### 6.2.1.2 i-vector and PLDA based framework for SV

The i-vector with PLDA based framework is used for the development of the baseline SV system. The RSR2015 database is used as development data for building the UBM, T-matrix and PLDA model [52]. The male and female subsets of the development data are processed and two gender depen-

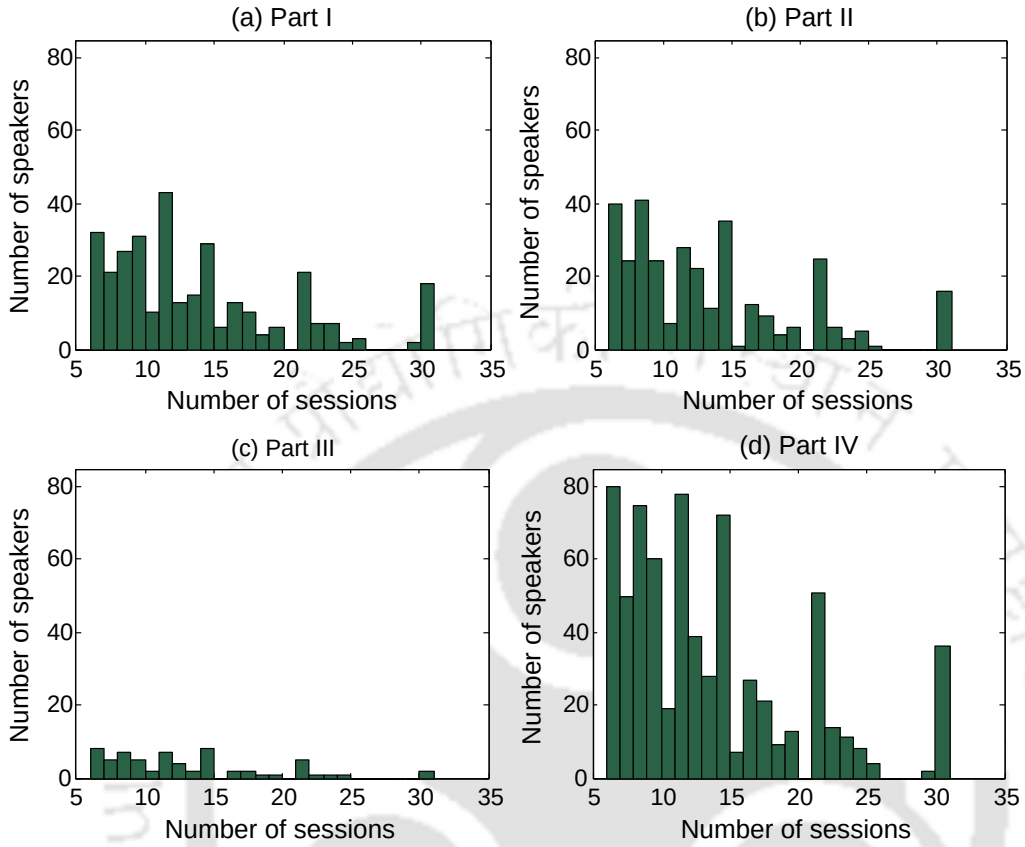


**Figure 6.7:** Histograms depicting number of sessions per speaker for male subset of RedDots database.

**Table 6.6:** Baseline system performance on RedDots database.

| RedDots Subset | Part I | | Part IV TD | |
|---|---|---|---|---|
| | EER (%) | DCF | EER (%) | DCF |
| Male | 9.50 | 0.0442 | 9.50 | 0.0433 |
| Female | 14.20 | 0.0524 | 13.73 | 0.0514 |

dent UBMs of 512 mixtures are built. Then the sufficient statistics are extracted for the development data as well as for the RedDots database. Two gender dependent T-matrices of 150 dimensions are trained using development data statistics. The i-vectors of male and female subset are then extracted using respective T-matrix for the mentioned database. Two 100-dimensional PLDA models are trained using the development data i-vectors of male and female subsets. They are used for channel/session compensation and scoring. The classification of a trial is performed according to the evaluation procedure of RedDots database [53]. Table 6.6 shows the performance of the baseline system in terms of EER and DCF for male and female subsets of Part I and Part IV TD of the database.

**Table 6.7:** Performance on RedDots database under implicit exploitation of session variability.

| RedDots Subset | Part I | | Part IV TD | |
|---|---|---|---|---|
| | EER (%) | DCF | EER (%) | DCF |
| Male | 6.91 | 0.0324 | 6.67 | 0.0308 |
| Female | 9.15 | 0.0409 | 8.47 | 0.0387 |

### 6.2.2 Proposed framework for session variability study

The evaluation of RedDots database is made according to the evaluation procedure, where sessions $2^{nd}$, $4^{th}$ and $6^{th}$ from each speaker are used for modeling and remaining for testing. The sessions $2^{nd}$, $4^{th}$ and $6^{th}$ are relatively former sessions compared to the sessions beyond 15-20 of each speaker. Hence they may have produced a lot of mismatch in speaker characteristics as can be known from the literature review discussed. To overcome this mismatch in session variation and to observe the effectiveness of it, an experimental setup is designed. In this setup, the first, middle and the last sessions of each speaker are taken for speaker modeling. However, these sessions are used as test sessions in the baseline setup. Therefore, the train sessions under baseline setup, i.e. $2^{nd}$, $4^{th}$ and $6^{th}$ sessions of speakers are used for testing in the experimental setup which is currently designed for session variation study. The performance for this setup is then evaluated with i-vector PLDA based framework. Table 6.7 shows the associated results from this study. It shows a significant improvement in performance than that of the baseline system. This improvement is attributed due to the implicit exploitation of session variability across the trials of each speaker. This is made by choosing the three sessions with large session variation for modeling the speakers. From the perspective of practical systems, regular testing is expected from a population of speakers for having access of some services. In such a scenario, this kind of approach may be adopted over some sizeable interval to have an impact on SV performance by addressing session variability in an implicit manner.

### 6.2.3 Proposed framework for template aging study

The session variation in speakers is found to have an important aspect that reflects in system performance. As the RedDots database is collected over a period of almost one year, the effect of template aging is observed over it across all the speakers. Hence this information may be useful for a practical system point of view. To study the effect of template aging of the speakers two experimental setups are designed and explained in this subsection.

**Table 6.8:** Performance on RedDots database considering first three sessions for modeling.

| RedDots Subset | Part I | | Part IV TD | |
|---|---|---|---|---|
| | EER (%) | DCF | EER (%) | DCF |
| Male | 7.53 | 0.0376 | 7.65 | 0.0364 |
| Female | 13.56 | 0.0476 | 11.32 | 0.0447 |

**Table 6.9:** Performance on RedDots database considering last three sessions for modeling.

| RedDots Subset | Part I | | Part IV TD | |
|---|---|---|---|---|
| | EER (%) | DCF | EER (%) | DCF |
| Male | 6.88 | 0.0326 | 6.57 | 0.0311 |
| Female | 10.25 | 0.0398 | 9.36 | 0.0376 |

#### 6.2.3.1 First three sessions

In this study, to observe the aging effect on the RedDots database, the speaker models are generated considering the first three sessions of the speakers. It is to be mentioned that sessions $2^{nd}$, $4^{th}$ and $6^{th}$ are used for the baseline framework. The trials under the baseline setup that contained the first three sessions are replaced with earlier train sessions, i.e. with $2^{nd}$, $4^{th}$ and $6^{th}$ sessions of the speakers. In this way, the SV system is built over i-vector PLDA platform for this study. Table 6.8 shows the performance of the same for the male and female subsets. It can be observed that the results are comparable to that obtained from the baseline, with slight deviation in both the subsets.

#### 6.2.3.2 Last three sessions

In the second set of studies for observing the effect of template aging, the last three sessions of the speakers are considered for speaker modeling. The performance of the system is evaluated in a similar manner as mentioned in the study based on the first three sessions of the speakers. The results under this study using the last three sessions of speakers for modeling can be seen from Table 6.9. These results show that they are having significant improvement over the baseline system, which is observed by comparing to Table 6.6 for both the male and female subsets. Additionally, the performance is far better from that obtained with training the speaker models with the first three sessions, which can be observed by comparing to Table 6.8. Thus, the aging of speaker templates shows an interesting trend that can be useful for exploitation in a practical system.

### 6.2.4 Discussion

In this subsection, a discussion is made over the results obtained under session variability and template aging studies. The studies under session variability showed that if the train templates of a speaker are having a larger session lapse among them, then that can help for achieving improved performance. Similarly, the study for template aging shows that if comparatively later sessions of a speaker are taken for modeling, that gives better performance than that obtained using earlier sessions. Both these studies are beneficial for practical systems, as an update of speaker models may be done in the system. Further, we compare the performance of the session variability study shown in Table 6.7 to the template aging study done by considering the last three sessions for speaker modeling, which is shown in Table 6.9. It can be observed that the results obtained from them are more or less comparable. This trend conveys that although there exists a lesser session gap in the last three sessions of a speaker, some vast speaker characteristics may have evolved with aging of the speaker templates. These characteristics are retained in comparatively later sessions, which have a large gap from the earliest sessions. It has a similar impact in the system performance if the speaker templates are chosen with more session variation among them for modeling. Additionally, another important fact can be the learning ability of the users. The RedDots database is collected through some application interface using mobile devices. This in turn affects the SV performance as the speakers may get acquainted with the system and the phrases over time, thereby struggle less to fix their pronunciation. If we assume that the speaker fixes his/her pronunciation after a few sessions, the last three sessions are closer from most of the sessions in terms of pronunciation than the first three sessions. This makes the large session variation and template aging as two very interesting areas to explore together in coming years with more precise investigation into it.

Figure 6.8 shows the detection error tradeoff (DET) curves for different studies based on session variability and template aging on Part I of the male speaker subset of RedDots database. The plots clearly show that the baseline system performance is comparatively closer to that obtained from modeling the speakers with the first three sessions. On the other hand, the performance that is obtained with modeling the speakers with first, middle and last sessions closely resembles to that obtained with modeling the speakers with the last three sessions. The both approaches show improved results compared against the baseline. This highlights the matter of discussion, showing many interesting aspects of session variability and template aging that can be utilized for practical realizable systems.

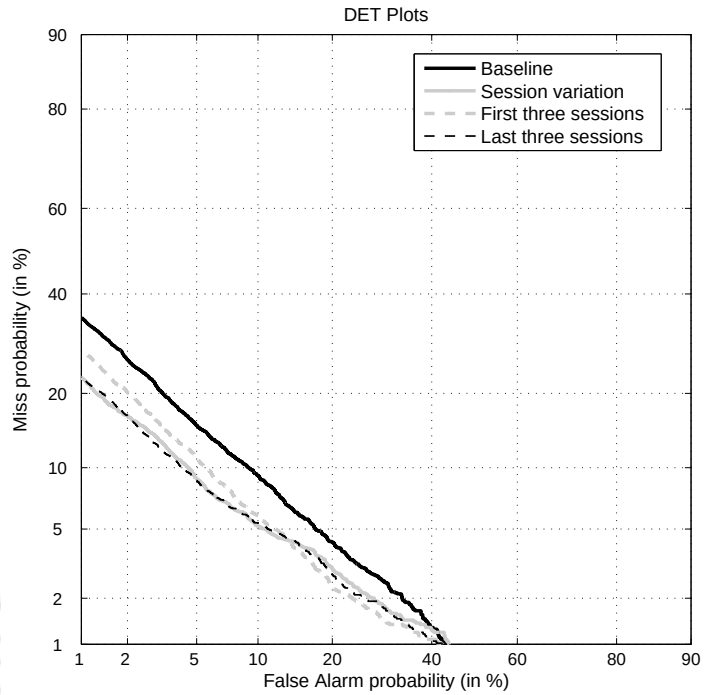


**Figure 6.8:** DET plots for different studies on male subset of Part I of RedDots database.

## 6.3 Summary

In this chapter, some issues that act as hurdles for a practical SV system are investigated along with possible suggestions for handling those. SV with limited test data based scenario is expected from the perspective of practical systems. Hence there are several issues, which can come across during the implementation of systems in practice. One such issue, which is addressed in this work is the mismatch in speaking rate between train and test sessions of the speakers. A novel framework involving faster prosody modification approach is proposed to modify the prosody of the test speech in terms of duration modification whenever there is a mismatch in speaking rate. It is found to be beneficial for compensating the mismatch in speech tempo and thereby provides improved SV performance. Further, its significance under sufficient train and limited test data based framework is also investigated. This exploration is followed by another investigation, when there is a large session variability present in the data. The effects of session variability and template aging are explored on RedDots database, which contains data collected over a period of about a year from a sizeable population. The session variability and template aging related studies indicate that these additional issues can be utilized in real-time SV systems to get the benefit out of it in regular sizeable intervals.

# 7

# Summary and Conclusions

Contents

### Overview

*This chapter summarizes the prime contributions of the thesis towards having a framework based on various explorations to deal with speaker verification (SV) using sufficient train and limited test data. Based on the contributions and different investigations, it provides some insights for future directions, which are also discussed.*

## 7.1 Summary

The thesis focuses on its objective towards development of an SV system using sufficient train and limited test data from the view of application oriented systems. The need of limited data has come across from providing user comfort and efficient decision delivery. Initially, different scenarios are investigated for having limited data setup in a text-independent SV system. The explorations project sufficient train with limited test data based framework as the most favorable for practical systems. This is because one time training with sufficient data can be made by the users of a system. However, the test condition has to involve limited data of short segments. The regular testings should provide comfort to the users in terms of demanding a less amount of speech data and efficient verification of a trial. Further, studies depict that the performance for segments with duration less than 10 s deteriorates exponentially. Therefore, the test segments of duration less than or equal to 10 s are considered for this work. In this direction, several explorations are made, which has significance towards the limited test data based scenario. Then all of them are put together for having an improved combined system. Firstly, the voice source features which are having alternative/complementary information from that of the conventional vocal tract features are explored. They are found to have different attributes of excitation source information, which are useful for capturing definite speaker characteristics. Then, text-constrained model based SV framework is proposed and significance of speaker-specific phonetic information in terms of the vocal tract constriction (VTC) feature is demonstrated. After having these explorations, kernel discriminant analysis (KDA) is implemented at the back-end of the i-vector based SV system for channel/session compensation. The KDA is found to perform well and particularly it is more significant for limited test data. Finally, all the explorations made with respect to different attributes of source information, speaker-specific phonetic information and KDA are tied up with a common thread to build a combined framework to handle the SV with limited test data. Additionally, some issues that come across practical systems are investigated as the problem definition of limited

test data based SV is from the perspective of such systems. These issues include mismatch in speech tempo between train and test sessions, effect of session variability and template aging for larger session gap present in the test sessions.

(i) **Different attributes of source information:**

SV with limited test data becomes challenging as there is very less data available from the speakers for validation of a claim. Thus, there lies a scope for exploring different features for extracting alternative/complementary speaker information. The traditional SV systems use mel frequency cepstral coefficient (MFCC) based vocal tract features for speaker characterization. However, the excitation source information is not considered while dealing with MFCC features. The voice source features contain definite knowledge of excitation source in terms of modeling the glottal excitation signal generated from the voiced sound units. This information is present in different representations, some of which can be viewed as glottal pulse shape, fundamental frequency, etc. The literature has shown that the voice source features although perform poorer compared to conventional vocal tract features, their significance is more under limited data condition. This makes explorations with source features as a prospective direction for SV with limited data.

The linear prediction (LP) residual of speech that mostly contains excitation source information does not contain second order relations as they are already extracted by LP analysis. Therefore, when compact representation of a source feature by some signal processing method is obtained, it may not capture all the aspects of the source. Additionally, the noise like structure of the LP residual itself creates difficulty in capturing source information. This signifies the importance for exploration of different source features. In this direction, three features mel power difference of spectrum in subbands (MPDSS), residual mel frequency cepstral coefficient (RMFCC) and discrete cosine transform of integrated linear prediction residual (DCTILPR) are explored. The MPDSS feature is obtained as subband version interpretation of spectral flatness measure, whereas the RMFCC feature is obtained by frame based processing of LP residual and provides source information using cepstral analysis. Similarly, the DCTILPR feature is obtained by pitch synchronous analysis, which provides compact representation of source information using DCT. Although LP residual is common for extraction of these features, the investigations through experimental analysis show that they contain different attributes of

source characteristics. The features MPDSS, RMFCC and DCTILPR are found to capture periodicity information, segmental level smoothed spectrum information and shape of glottal signal, respectively.

As the source features are found to contain different attributes, their fusion is carried out expecting a better performance. The combination of the source features with one another and overall fusion of three provides commendable improvement than that with individual case. It is also observed that the performance obtained by the fusion of three source features for very short segment of test speech for 3 s and 2 s case, outperforms the baseline results obtained with MFCC features. This signifies the importance of the different aspects of excitation source information for SV with limited test data. Further, the score level fusion of the three source features is done with MFCC features, which enhances the results by a large margin for each of the cases considered for limited test data studies.

(ii) **Text-constrained speaker models and VTC information:**

The work explored with respect to the excitation source features and their different attributes is found to be useful for SV with limited test data based scenario. Another direction for improving the performance in such a scenario is the utilization of acoustic-phonetic information. This is motivated from different works that show consideration of acoustic-phonetic knowledge from a speech signal provides an additive information for speaker modeling. In this regard, firstly, a text-constrained model based setup is proposed that utilizes speaker-specific phonetic information in an explicit manner. It consists of a combination of a user chosen phrase and a common fixed phrase for training the speaker models. The same is expected from the speakers during testing. By introduction of a user chosen phrase along with a common fixed phrase some amount of constraint is put on the text and hence it is referred to as the text-constrained model. This kind of framework, provides a lexical content match between train and test sessions that helps in SV under limited data condition. The studies with respect to text-constrained models are then extended to sufficient train and limited test data condition on Part IV of RedDots database. This database has two different enrollment conditions, namely text-dependent and text-prompted. Here the former refers to enrollment with three instances of a fixed phrase and a test phrase having the same lexical content. On the contrary, the latter deals with enrollment made by a set of around 10 different fixed phrases. A phrase having different lexical

content from the phrases used for training is considered during testing. The text-prompted terminology of the database is considered as text-independent for keeping uniformity in the thesis. The text-constrained model based framework is explored in case of text-independent enrollment condition. It is made by replacing one of the phrases taken for training by a phrase of the speaker having the same lexical content to that of the test phrase. This setup provides improvement over the baseline showing its effectiveness. Thus it implies the importance of explicit utilization of the phonetic match for speaker modeling. The different attributes of source information discussed previously have been also explored for the text-constrained based framework in order to demonstrate their significance.

The work later focuses on utilization of the phonetic information in an implicit manner. The text-constrained based framework puts some constraint on the users in the text-independent based framework. However, this scenario may not be available all the time. In such cases, to utilize the phonetic information in an implicit manner, the VTC based feature is used. It captures the level of constriction resulted in the vocal tract while producing different sound units. The level of constriction is expected to be distinct for each speaker as the size and shape of vocal tract are unique for each individual. With this motivation VTC based feature is combined with MFCC features, which results into an improvement over the baseline. This is attributed as it captures some additional knowledge of the speakers in terms of speaker-specific phonetic information. Thus these explorations show the importance of the phonetic information for speaker modeling, which is utilized in an explicit and then in an implicit manner.

(iii) **Kernel discriminant analysis and combined framework:**

The contributions towards having improved performance after exploring different attributes of source information and then with utilization of speaker-specific phonetic information, another direction is investigated. The observations from i-vector based speaker modeling show that the i-vectors of the short utterances vary much due to large variation in the phonetic content. Therefore, the conventional pattern recognition techniques for channel/session compensation could not distinctly separate the i-vectors across different classes in such a scenario. In this direction, KDA is explored, which transforms the feature vectors into a higher dimensional space and then performs discriminant analysis. The KDA is used at the back-end of i-vector based speaker modeling for channel/session compensation. It performs better than the other

existing techniques like linear discriminant analysis (LDA) followed by within class covariance normalization (WCCN) and probabilistic linear discriminant analysis (PLDA). Further, it is observed that the KDA is more useful while dealing with limited test data scenario. This signifies the importance of KDA to compensate different variabilities for SV systems under test data conditions.

The work then focuses on bringing out a framework combining all the explorations made to have improved speaker characterization for SV with limited test data. In this final framework, VTC feature is extracted and combined with conventional MFCC features at the feature level to build a single SV system. Three parallel systems are built using the different attributes of excitation source information based on source features MPDSS, RMFCC and DCTILPR. Then KDA based channel/session compensation is applied at the back-end of each system developed with the stated features. Finally, a score level fusion is made for all the systems to complete the combined framework with the different explorations made. The proposed combined system is able to handle SV using limited test data to a large extent enhancing the system performance.

(iv) **Different issues:**

The first three working modules of the thesis concentrate on improving the performance of SV with limited test data from the view of practical systems. During implementation of such systems for regular use, there comes different issues that certainly bring down the performance. Hence, these issues are very much crucial and their addressing is necessary. In this regard, some of these issues are investigated and discussed based on experimental studies.

The first issue deals with a mismatch in speech tempo between train and test sessions. It is observed that when there is a mismatch in speaking rate between train and test sessions, it affects the performance. In order to handle this issue a framework is proposed, where the prosody modification in terms of duration modification of the test speech is performed according to the speaking rate of the claimed speaker. The faster prosody modification algorithm which is found to be efficient for duration modification is used for this purpose. The SV studies are then made with the modified test speech that helps in improving the system performance. Further, the SV using sufficient train and limited test based scenario is evaluated for mismatch in speech tempo conditions. Then the proposed framework is used for compensating the mismatch, which is also found to be effective for such a scenario. Additionally, the studies apart from conducting

on a simulated database for a larger set, extended to demonstrate for a small set collected on a practical setting involving short phrases.

Another issue that comes into the picture when there is a lot of session variability present in the test sessions. For a field deployable system, which is used over a long period of time like a year or so, there exists a large session gap of the test examples from the train sessions. It thereby directly hampers the system performance. The RedDots database contains data collected over a year period, which makes it possible for the study with session variability. A framework is proposed to update the speaker models when there is a large session gap to compensate the session variation. It is performed in an implicit manner by considering the first, middle and the last sessions of a speaker for a fixed phrase based limited data scenario involving short utterances. Further, the studies related to template aging are conducted. The experimental studies show that there is a variation in performance when the speaker models are trained with first three sessions to that obtained by considering the last three sessions. This occurs when there is session variability over a period of a year or so. It is hypothesized that the speaker characteristics are evolved over time to some extent. Moreover, the learning ability of the speakers increase to a system as they go on testing the same regularly. Thus updating the speaker models with later sessions may also help for speaker characterization in a practical scenario. In this way, some issues that occur during the development and implementation of an SV system are investigated and discussed.

## 7.2 Contributions

The prime contributions of the thesis may be viewed as,

(i) Bringing out an SV framework having sufficient train with limited test data suitable for application oriented systems.

(ii) Exploring different attributes of source information and demonstrating their significance for SV with limited test data on fusion.

(iii) Proposal of a text-constrained model based framework and explicit/implicit utilization of speaker-specific phonetic information for speaker modeling.

(iv) Exploring kernel based discriminant analysis and its scope for SV with limited test data.

(v) A common framework involving different attributes of excitation source features, speaker-

specific phonetic information in terms of VTC feature and suitable pattern recognition technique as KDA at the back-end for dealing with the limited test data scenario.

(vi) Investigating some issues related to SV with limited test data from the perspective of practical systems: mismatch speech tempo, session variability and template aging.

## 7.3 Future work directions

The work carried out in this thesis with respect to SV using sufficient train and limited test data provides directions for future work. Some of them may be seen as,

(i) The exploration with respect to the source features showed that they carry different attributes of excitation source information. It is observed that the nature of complementary information is more for DCTILPR feature, which is extracted in a pitch synchronous manner. The features MPDSS and RMFCC are extracted by short term processing of the speech signal. In this regard, the work can be extended to extract MPDSS and RMFCC features in a pitch synchronous manner to observe their effectiveness against DCTILPR feature.

(ii) The current investigations consider the energy of speech signal to detect the regions with sufficient voice activity. The features belonging to the detected regions are taken for modeling, which also applies for the case of source features. Hence, there lies a scope to detect glottal activity regions and then consider the extraction of a source feature. These regions are expected to be useful for capturing the excitation source information. Additionally, the phase component of the features can be explored, which can contribute towards speaker characterization.

(iii) The thesis proposes a text-constrained model based framework that puts some constraint on the text to be produced for having a match of phonetic information. This work can be extended to provide the match of sound units between train and test sessions without putting any constraint on the users. It may be done by detecting the sound units from the test speech and then picking up those sound units from the train speech. The detected units from the train speech can be used for speaker modeling for having a better match to the test sessions.

(iv) The thesis explores VTC evidence for consideration of speaker-specific phonetic information in an implicit manner. It is expected that other evidence that capture speaker characteristics, mainly for vowel-like sound units can be investigated. In this regard, evidence like vowel roundness, frontness can be explored, which may have unique speaker information.

(v) The kernel based discriminant analysis is used at the back-end for compensation of different variabilities like channel/session information, which is found to be effective for SV studies under limited test data based scenario. However, variabilities that occur in such a scenario are not quantified and investigated in detail. This can be explored in future to have specific compensation and normalization techniques with respect to each of the variability aspects.

(vi) From the view of practical systems, the thesis investigates some of the issues like the mismatch in speech tempo, session variability and template aging. The impact of session variability and template aging shows the potential for speaker confidence estimation. This can be crucial for practical systems, particularly for limited data scenario. Further, the studies under degraded conditions are not explored in this thesis. Such a scenario can be investigated to observe the impact of excitation source, speaker-specific phonetic information and pattern recognition approaches in adverse conditions. Moreover, issues like spoofing attacks are also of concern as they have significance towards application oriented systems.

# Bibliography


[1] D. O'Shaughnessy, "Speaker recognition," *IEEE ASSP Magazine*, vol. 3, no. 4, pp. 4–17, October 1986.

[2] J. H. L. Hansen and T. Hasan, "Speaker recognition by machines and humans: A tutorial review," *IEEE Signal Processing Magazine*, vol. 32, no. 6, pp. 74–99, November 2015.

[3] M. Hèbert, "Text-dependent speaker recognition," *Springer-Verlag Heidelberg*, pp. 743–762, 2008.

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

[5] J. R. Deller, J. H. L. Hansen, and J. G. Proakis, *Discrete-Time Processing of Speech Signals*. IEEE press, 1999.

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

[7] J. Makhoul, "Linear prediction: A tutorial review," *Proceedings of the IEEE*, vol. 63, no. 4, pp. 561–580, April 1975.

[8] X. Huang, A. Acero, and H. W. Hon, *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*. Prentice-Hall, New Jersey, 2001.

[9] H. Hermansky, "Perceptual linear prediction (PLP) analysis for speech," *Journal of the Acoustic Society of America*, vol. 87, pp. 1738–1752, 1990.

[10] M. Plumpe, T. Quatieri, and D. Reynolds, "Modeling of the glottal flow derivative waveform with application to speaker identification," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 7 (5), pp. 569–586, 1999.

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

[12] T. Kinnunen and P. Alku, "On separating glottal source and vocal tract information in telephony speaker verification," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) 2009*, April 2009, pp. 4545–4548.

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

[14] H. Hermansky, "Should recognizers have ears?" *Speech Communication*, vol. 25, no. 13, pp. 3 – 27, 1998.

[15] T. Kinnunen, "Joint acoustic-modulation frequency for speaker recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2006*, vol. 1, May 2006, pp. 665–668.

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

[17] A. Adami, R. Mihaescu, D. Reynolds, and J. Godfrey, "Modeling prosodic dynamics for speaker recognition," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) 2003*, vol. 4, April 2003, pp. IV–788–91.

[18] K. Bartkova, D.L.Gac, D. Charlet, and D. Jouvet, "Prosodic parameter for speaker identification," in *International Conference on Spoken Language Processing (ICSLP) 2002*, September 2002, pp. 1197–1200.

[19] A. Benyassine, E. Shlomot, H. Y. Su, D. Massaloux, C. Lamblin, and J. P. Petit, "Itu-t recommendation g.729 annex b: a silence compression scheme for use with g.729 optimized for v.70 digital simultaneous voice and data applications," *IEEE Communications Magazine*, vol. 35, no. 9, pp. 64–73, September 1997.

[20] V. Hautamki, M. Tuononen, T. Niemi-laitinen, and P. Frnti, "Improving speaker verification by periodicity based voice activity detection," in *12th Int. Conf. Speech and Computer (SPECOM)*, October 2007.

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

[22] S. R. M. Prasanna and G. Pradhan, "Significance of vowel-like regions for speaker verification under degraded conditions," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 8, pp. 2552–2565, November 2011.

[23] G. Pradhan and S. R. M. Prasanna, "Speaker verification by vowel and nonvowel like segmentation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 4, pp. 854–867, April 2013.

[24] T. Kinnunen and P. Rajan, "A practical, self-adaptive voice activity detector for speaker verification with noisy telephone and microphone data," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) 2013*, May 2013, pp. 7229–7233.

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

[26] B. S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *The Journal of the Acoustical Society of America*, vol. 55, no. 6, pp. 1304–1312, 1974.

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

[28] D. A. Reynolds and R. Rose, "Robust text-independent speaker identification using gaussian mixture speaker models," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 1, pp. 72–83, January 1995.

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

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

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

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

[33] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*. John Wiley and Sons, 2000.

[34] A. O. Hatch, S. Kajarekar, and A. Stolcke, "Within-class covariance normalization for SVM-based speaker recognition," in *International Conference on Spoken Language Processing (ICSLP) 2006*, 2006, pp. 1471–1474.

[35] W. M. Campbell, D. E. Sturim, D. A. Reynolds, and A. Solomonoff, "SVM based speaker verification using a GMM supervector kernel and NAP variability compensation," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) 2006*, May 2006.

[36] D. Garcia-Romero, X. Zhou, and C. Y. Espy-Wilson, "Multicondition training of gaussian plda models in i-vector space for noise and reverberation robust speaker recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2012*, March 2012, pp. 4257–4260.

[37] T. Hasan and J. H. L. Hansen, "Acoustic factor analysis for robust speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 4, pp. 842–853, April 2013.

[38] ——, "Maximum likelihood acoustic factor analysis models for robust speaker verification in noise," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 2, pp. 381–391, February 2014.

[39] P. Kenny, V. Gupta, T. Stafylakis, P. Ouellet and J. Alam, "Deep neural networks for extracting baum-welch statistics for speaker recognition," in *Speaker Odyssey 2014*, 2014.

[40] F. Richardson, D. Reynolds, and N. Dehak, "Deep neural network approaches to speaker and language recognition," *IEEE Signal Processing Letters*, vol. 22, no. 10, pp. 1671–1675, October 2015.

[41] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, "Score normalization for text-independent speaker verification systems," *Digital Signal Processing*, vol. 10, no. 13, pp. 42–54, 2000.

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

[43] L. Heck and M. Weintraub, "Handset-dependent background models for robust text-independent speaker recognition," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) 1997*, vol. 2, April 1997, pp. 1071–1074.

[44] R. Dunn, T. Quatieri, D. Reynolds, and J. Campbell, "Speaker recognition from coded speech and the effects of score normalization," in *Conference Record of the Thirty-Fifth Asilomar Conference on Signals, Systems and Computers, 2001*, vol. 2, November 2001, pp. 1562–1567.

[45] N. Brummer, L. Burget, J. Cernocky, O. Glembek, F. Grezl, M. Karafiat, D. 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 Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 7, pp. 2072–2084, September 2007.

[46] D. Chakrabarty, S. R. M. Prasanna, and R. K. Das, "Development and evaluation of online text-independent speaker verification system for remote person authentication," *International Journal of Speech Technology*, vol. 16, no. 1, pp. 75–88, 2013.

[47] K. A. Lee, A. Larcher, H. Thai, B. Ma, and H. Li, "Joint application of speech and speaker recognition for automation and security in smart home," in *Interspeech 2011*, 2011, pp. 3317–3318.

[48] S. Dey, S. Barman, R. K. Bhukya, R. K. Das, Haris B C, S. R. M. Prasanna, and R. Sinha, "Speech biometric based attendance system," in *National Conference on Communications (NCC) 2014, IIT Kanpur*, 2014.

[49] R. Ramos-Lara, M. Lpez-Garca, E. Cant-Navarro, and L. Puente-Rodriguez, "Real-time speaker verification system implemented on reconfigurable hardware," *Journal of Signal Processing Systems*, vol. 71, no. 2, pp. 89–103, 2013.

[50] Rohan Kumar Das, S. Jelil, and S. R. M. Prasanna, "Development of multi-level speech based person authentication system," *Journal of Signal Processing Systems*, pp. 1–13, 2016. [Online]. Available: http://dx.doi.org/10.1007/s11265-016-1148-z

[51] "NIST SRE Evaluations 1999-2016, NIST USA."

[52] A. Larcher, K. A. Lee, B. Ma, and H. Li, "Text-dependent speaker verification: Classifiers, databases and RSR2015," *Speech Communication*, vol. 60, pp. 56–77, 2014.

[53] K. A. Lee, A. Larcher, W. Guangsen, K. Patrick, N. Brummer, D. van Leeuwen, H. Aronowitz, M. Kockmann, C. Vaquero, B. Ma, H. Li, T. Stafylakis, J. Alam, A. Swart, and J. Perez, "The RedDots data collection for speaker recognition," in *Interspeech 2015, Dresden, Germany*, 2015, pp. 2996–3000.

[54] A. Kanagasundaram, R. Vogt, D. Dean, S. Sridharan, , and M. Mason, "i-vector based speaker recognition on short utterances," in *Interspeech 2011*, 2011.

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

[56] L. Rabiner and B. Juang, *Fundamentals of speech recognition*. Prentice-Hall, Inc. Upper Saddle River, NJ, USA, 1993.

[57] D. Hosseinzadeh and S. Krishnan, "On the use of complementary spectral features for speaker recognition," *EURASIP Journal on Advances in Signal Processing*, vol. 2008, no. 1, 2007.

[58] H. Li, B. Ma, K. A. Lee, H. Sun, D. Zhu, K. C. Sim, C. You, R. Tong, I. Karkkainen, C. L. Huang, V. Pervouchine, W. Guo, Y. Li, L. Dai, M. Nosratighods, T. Tharmarajah, J. Epps, E. Ambikairajah, E.-S. Chng, T. Schultz, and Q. Jin, "The i4u system in nist 2008 speaker recognition evaluation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2009*, April 2009, pp. 4201–4204.

[59] M. Sahidullah and T. Kinnunen, "Local spectral variability features for speaker verification," *Digital Signal Processing*, vol. 50, pp. 1–11, 2016.

[60] M. Athineos and D. P. W. Ellis, "Autoregressive modeling of temporal envelopes," *IEEE Transactions on Signal Processing*, vol. 55, no. 11, pp. 5237–5245, November 2007.

[61] S. O. Sadjadi and J. H. L. Hansen, "Mean hilbert envelope coefficients (mhec) for robust speaker and language identification," *Speech Communication*, vol. 72, pp. 138–148, 2015.

[62] C. Kim and R. M. Stern, "Power-normalized cepstral coefficients (pncc) for robust speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2012*, March 2012, pp. 4101–4104.

[63] J. Gudnason and M. Brookes, "Voice source cepstrum coefficients for speaker identification," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) 2008*, 2008, pp. 4821–4824.

[64] N. Zheng, T. Lee, and P. C. Ching, "Integration of complementary acoustic features for speaker recognition," *IEEE Signal Processing Letters*, vol. 14, no. 3, pp. 181–184, March 2007.

[65] D. Pati and S. R. M. Prasanna, "A comparative study of explicit and implicit modelling of subsegmental speaker-specific excitation source information," *Sadhana*, vol. 38, no. 4, pp. 591–620, 2013.

[66] ——, "Speaker information from subband energies of linear prediction residual," in *National Conference on Communications (NCC), 2010*, January 2010, pp. 1–4.

[67] A. G. Ramakrishnan, B. Abhiram, and S. R. M. Prasanna, "Voice source characterization using pitch synchronous discrete cosine transform for speaker identification," *JASA Express Letters*, vol. 137, pp. EL469–EL475, 2015.

[68] K. Vijayan, P. R. Reddy, and K. S. R. Murty, "Significance of analytic phase of speech signals in speaker verification," *Speech Communication*, vol. 81, pp. 54 – 71, 2016, phase-Aware Signal Processing in Speech Communication.

[69] W. Chan, N. Zheng, and T. Lee, "Discrimination power of vocal source and vocal tract related features for speaker segmentation," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 15 (6), pp. 1884–1892, 2007.

[70] F. Soong and A. Rosenberg, "On the use of instantaneous and transitional spectral information in speaker recognition," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) 1986*, vol. 11, April 1986, pp. 877–880.

[71] I. Magrin-Chagnolleau, G. Durou, and F. Bimbot, "Application of time-frequency principal component analysis to text-independent speaker identification," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 6, pp. 371–378, September 2002.

[72] N. Malayath, H. Hermansky, S. Kajarekar, and B. Yegnanarayana, "Data-driven temporal filters and alternatives to gmm in speaker verification," *Digital Signal Processing*, vol. 10, no. 1, pp. 55 – 74, 2000.

[73] C. Jankowski Jr, T. Quatieri, and D. Reynolds, "Measuring fine structure in speech: Application to speaker identification," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) 1995*, vol. 1, 1995, pp. 325–328.

[74] M. Grimaldi and F. Cummins, "Speaker identification using instantaneous frequencies," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 6, pp. 1097–1111, 2008.

[75] M. S. Deshpande and R. S. Holambe, "Speaker identification based on robust am-fm features," in *2nd IEEE International Conference on Emerging Trends in Engineering and Technology (ICETET), 2009*, 2009, pp. 880–884.

[76] N. E. Huang, Z. Shen, S. R. Long, M. C. Wu, H. H. Shih, Q. Zheng, N.-C. Yen, C. C. Tung, and H. H. Liu, "The empirical mode decomposition and the hilbert spectrum for nonlinear and non-stationary time series analysis," *Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, vol. 454, no. 1971, pp. 903–995, 1998.

[77] N. E. Huang and S. S. Shen, *Hilbert-Huang transform and its applications*. World Scientific, 2005, vol. 5.

[78] E. Ambikairajah, "Emerging features for speaker recognition," in *6th International Conference on Information, CommunicationsSignal Processing 2007*, December 2007, pp. 1–7.

[79] P. Rose, *Forensic speaker identification*. Taylor & Francis London, New York, 2002.

[80] M. J. Carey, E. S. Parris, H. Lloyd-Thomas, and S. Bennett, "Robust prosodic features for speaker identification," in *Fourth International Conference on Spoken Language, 1996 (ICSLP '96)*, vol. 3, October 1996, pp. 1800–1803.

[81] J. Markel, B. Oshika, and A. Gray, "Long-term feature averaging for speaker recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 25, no. 4, pp. 330–337, August 1977.

[82] 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, 2008.

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

[84] G. Doddington, "Speaker recognition based on idiolectal differences between speakers," in *European Conference on Speech Communication and Technology (Eurospeech)*, 2001.

[85] W. D. Andrews, M. A. Kohler, J. Campbell, J. J. Godfrey, and J. Hernandez-Cordero, "Gender-dependent phonetic refraction for speaker recognition," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) 2002*, vol. 1, May 2002, pp. I–149–I–152.

[86] K. Leung, M. Mak, M. Siu, and S. Kung, "Adaptive articulatory feature-based conditional pronunciation modeling for speaker verification," *Speech Communication*, vol. 48, no. 1, pp. 71–84, 2006.

[87] M. Li and W. Liu, "Speaker verification and spoken language identification using a generalized i-vector framework with phonetic tokenizations and tandem features," in *Interspeech 2014, Singapore*, 2014.

[88] T. Fu, Y. Qian, Y. Liu, and K. Yu, "Tandem deep features for text-dependent speaker verification," in *Interspeech 2014, Singapore*, 2014, pp. 1327–1331.

[89] T. N. Sainath, B. Kingsbury, and B. Ramabhadran, "Auto-encoder bottleneck features using deep belief networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2012*, March 2012, pp. 4153–4156.

[90] K. Vesel, M. Karafit, F. Grzl, M. Janda, and E. Egorova, "The language-independent bottleneck features," in *IEEE Spoken Language Technology Workshop (SLT) 2012*, December 2012, pp. 336–341.

[91] S. X. Zhang, M. W. Mak, and H. Meng, "Speaker verification via high-level feature based phonetic-class pronunciation modeling," *IEEE Transactions on Computers*, vol. 56, no. 9, pp. 1189–1198, September 2007.

[92] N. Fatima and T. F. Zheng, "Vowel-category based short utterance speaker recognition," in *International Conference on Systems and Informatics (ICSAI 2012)*, 2012.

[93] N. Scheffer and Y. Lei, "Content matching for short duration speaker recognition," in *Interspeech 2014, Singapore*, 2014.

[94] A. Larcher, P. Bousquet, K. A. Lee, D. Matrouf, H. Li, and J.-F. Bonastre, "i-vectors in the context of phonetically-constrained short utterances for speaker verification," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2012*, March 2012, pp. 4773–4776.

[95] A. Larcher, K. A. Lee, B. Ma, and H. Li, "Phonetically-constrained PLDA modeling for text-dependent speaker verification with multiple short utterances," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2013*, May 2013, pp. 7673–7677.

[96] R. W. Aldhaheri and F. E. Al-Saadi, "Robust text-independent speaker recognition with short utterance in noisy environment using svd as a matching measure," *Journal of King Saud University - Computer and Information Sciences*, vol. 17, pp. 25–44, 2004.

[97] P. Kenny, T. Stafylakis, P. Ouellet, M. J. Alam, and P. Dumouchel, "Plda for speaker verification with utterances of arbitrary duration," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2013*, May 2013, pp. 7649–7653.

[98] A. K. Sarkar, D. Matrouf, P. M. Bousquet, and J. F. Bonastre, "Study of the effect of i-vector modeling on short and mismatch utterance duration for speaker verification," in *Interspeech 2012*, 2012.

[99] V. Hautamki, Y.-C. Cheng, P. Rajan, and C.-H. Lee, "Minimax i-vector extractor for short duration speaker verification," in *Interspeech 2013*, 2013.

[100] A. Kanagasundaram, D. Dean, J. Gonzalez-Dominguez, S. Sridharan, D. Ramos, and J. Gonzalez-Rodriguez, "Improving short utterance based i-vector speaker recognition using source and utterance-duration normalization techniques," in *Interspeech 2013*, 2013.

[101] A. Kanagasundaram, D. Dean, S. Sridharan, J. Gonzalez-Dominguez, J. Gonzalez-Rodriguez, and D. Ramos, "Improving short utterance i-vector speaker verification using utterance variance modelling and compensation techniques," *Speech Communication*, vol. 59, pp. 69–82, 2014.

[102] T. Hasan, R. Saeidi, J. H. L. Hansen, and D. A. van Leeuwen, "Duration mismatch compensation for i-vector based speaker recognition systems," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2013*, May 2013, pp. 7663–7667.

[103] M. McLaren, R. Vogt, B. Baker, and S. Sridharan, "A comparison of session variability compensation approaches for speaker verification," *IEEE Transactions on Information Forensics and Security*, vol. 5, no. 4, pp. 802–809, December 2010.

[104] A. Stolcke, S. S. Kajarekar, L. Ferrer, and E. Shrinberg, "Speaker recognition with session variability normalization based on mllr adaptation transforms," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 7, pp. 1987–1998, September 2007.

[105] J. Villalba and E. Lleida, "Preventing replay attacks on speaker verification systems," in *International Carnahan Conference on Security Technology (ICCST) 2011*, pp. 1–8.

[106] ——, "Detecting replay attacks from far-field recordings on speaker verification systems," in *Biometrics and ID Management*. Springer, 2011, pp. 274–285.

[107] Z. Wu, S. Gao, E. S. Chng, and H. Li, "A study on replay attack and anti-spoofing for text-dependent speaker verification," in *Asia-Pacific Signal Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2014.

[108] Z. Wu, T. Kinnunen, N. Evans, J. Yamagishi, C. Hanilc i, M. Sahidullah, and A. Sizov, "Asvspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge," in *Interspeech 2015, Dresden, Germany*, 2015.

[109] B. Fauve, N. Evans, and J. Mason, "Improving the performance of text-independent short duration svm- and gmm-based speaker verification," in *In: The Speaker and Language Recognition Workshop (Odyssey 2008), Stellenbosch, South Africa,*, January 2008.

[110] P. Krishnamoorthy, H. S. Jayanna, and S. R. M. Prasanna, "Speaker verification under limited data condition by noise addition," *Expert Systems and Applications (Elsevier)*, vol. 38, pp. 13 487–13 490, 2011.

[111] W. M. Fisher, G. R. Doddington, and K. M. Goudie-Marshall, "The DARPA Speech Recognition Research Database: Specifications and Status," in *Proceedings of DARPA Workshop on Speech Recognition*, 1986, pp. 93–99.

[112] *"The NIST Year 2003 Speaker Recognition Evaluation Plan", NIST , February 2003.*

[113] K. S. R. Murty, V. Boominathan, and K. Vijayan, "Allpass modeling of lp residual for speaker recognition," in *International Conference on Signal Processing and Communications (SPCOM) 2012, IISc Bangalore*, July 2012, pp. 1–5.

[114] S. Hayakawa, K. Takeda and F. Itakura, "Speaker identification using harmonic structure of lp-residual spectrum," *Biometric personal Aunthentification, Lecture notes, Springer, Berlin*, vol. 1206, pp. 253–260, 1997.

[115] A. H. Gray Jr. and J. D. Markel, "A spectral-flatness measure for studying the autocorrelation method of linear prediction of speech analysis," *IEEE Trans. on Acoustic Speech and Signal Process.*, vol. ASSP-22, no. 3, pp. 207–217, March 1974.

[116] P. Thèvenaz and H. Hügli, "Usefulness of the lpc-residue in text-independent speaker verification," *Speech Communication*, vol. 17, no. 12, pp. 145–157, 1995.

[117] A. P. Prathosh, T. V. Ananthapadmanabha, and A. G. Ramakrishnan, "Epoch extraction based on integrated linear prediction residual using plosion index," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 21, issue 12, pp. 2471–2480, 2013.

[118] T. V. Ananthapadmanabha, A. P. Prathosh, and A. G. Ramakrishnan, "Detection of closure-burst transitions of stops and affricates in continuous speech using plosion index," *Journal of the Acoustical Society of America*, vol. 135, no. 1, pp. 460–471, 2014.

[119] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The DET curve in assessment of detection task performance," in *Eurospeech 1997*, Rhodes, Greece, 1997, pp. 1895–1898.

[120] Sarfaraz Jelil, Rohan Kumar Das, Rohit Sinha, and S. R. M. Prasanna, "Speaker verification using gaussian posteriorgrams on fixed phrase short utterances," in *Interspeech 2015, Dresden, Germany*, 2015, pp. 1042–1046.

[121] *"The NIST Year 2010 Speaker Recognition Evaluation Plan", NIST , 2010.*

[122] M. J. Alam, P. Kenny, and V. Gupta, "Tandem features for text-dependent speaker verification on the reddots corpus," in *Interspeech 2016*, 2016, pp. 420–424.

[123] H. Zeinali, H. Sameti, L. Burget, J. ernock, N. Maghsoodi, and P. Matjka, "i-vector/hmm based text-dependent speaker verification system for reddots challenge," in *Interspeech 2016*, 2016, pp. 440–444.

[124] A. K. Sarkar and Z.-H. Tan, "Text dependent speaker verification using un-supervised hmm-ubm and temporal gmm-ubm," in *Interspeech 2016*, 2016, pp. 425–429.

[125] G. Wang, K. A. Lee, T. H. Nguyen, H. Sun, and B. Ma, "Joint speaker and lexical modeling for short-term characterization of speaker," in *Interspeech 2016*, 2016, pp. 415–419.

[126] J. Ma, S. Irtza, K. Sriskandaraja, V. Sethu, and E. Ambikairajah, "Parallel speaker and content modelling for text-dependent speaker verification," in *Interspeech 2016*, 2016, pp. 435–439.

[127] T. Kinnunen, M. Sahidullah, I. Kukanov, H. Delgado, M. Todisco, A. K. Sarkar, N. B. Thomsen, V. Hautamki, N. Evans, and Z.-H. Tan, "Utterance verification for text-dependent speaker recognition: A comparative assessment using the reddots corpus," in *Interspeech 2016*, 2016, pp. 430–434.

[128] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems," in *Interspeech 2011*, August 2011, pp. 249–252.

[129] B. D. Sarma and S. R. M. Prasanna, "Analysis of vocal tract constrictions using zero frequency filtering," *IEEE Signal Processing Letters*, vol. 21, no. 12, pp. 1481–1485, December 2014.

[130] K. S. R. Murty and B. Yegnanarayana, "Epoch extraction from speech signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 8, pp. 1602–1613, November 2008.

[131] Rohan Kumar Das and S. R. M. Prasanna, *Speaker Verification for Variable Duration Segments and the Effect of Session Variability.* Lecture Notes in Electrical Engineering: Springer, 2015, ch. 16, pp. 193–200.

[132] M. Sugiyama, "Local fisher discriminant analysis for supervised dimensionality reduction," in *23rd international conference on Machine learning.* ACM, 2006, pp. 905–912.

[133] J. Ye and S. Ji, "Discriminant analysis for dimensionality reduction: An overview of recent developments," *Biometrics: Theory, Methods, and Applications. Wiley-IEEE Press, New York*, 2010.

[134] G. Baudat and F. Anouar, "Generalized discriminant analysis using a kernel approach," *Neural computation*, vol. 12, no. 10, pp. 2385–2404, 2000.

[135] B. Schölkopf and K.-R. Mullert, "Fisher discriminant analysis with kernels," *in Proceedings of IEEE Workshop on Neural Networks for Signal Processing IX*, 1999.

[136] G. McLachlan, *Discriminant analysis and statistical pattern recognition.* John Wiley & Sons, 2004.

[137] V. N. Vapnik, *Statistical learning theory.* John Wiley & Sons, 1998.

[138] Q. Liu, R. Huang, H. Lu, and S. Ma, "Face recognition using kernel-based fisher discriminant analysis," in *fifth IEEE International Conference on Automatic Face and Gesture Recognition*, 2002, pp. 197–201.

[139] S. Zafeiriou, G. Tzimiropoulos, M. Petrou, and T. Stathaki, "Regularized kernel discriminant analysis with a robust kernel for face recognition and verification," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 23, no. 3, pp. 526–534, 2012.

[140] D. Zhou and Z. Tang, "Kernel-based improved discriminant analysis and its application to face recognition," *Soft Computing*, vol. 14, no. 2, pp. 103–111, 2010.

[141] X. Zhao and S. Zhang, "Facial expression recognition based on local binary patterns and kernel discriminant isomap," *Sensors*, vol. 11, no. 10, pp. 9573–9588, 2011.

[142] J. Yang, A. F. Frangi, J. Y. Yang, D. Zhang, and Z. Jin, "KPCA plus LDA: a complete kernel Fisher discriminant framework for feature extraction and recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 2, pp. 230–244, 2005.

[143] A. M. Khan, Y.-K. Lee, S. Lee, and T.-S. Kim, "Human activity recognition via an accelerometer-enabled-smartphone using kernel discriminant analysis," in *5th IEEE International Conference on Future Information Technology (FutureTech), 2010*, 2010, pp. 1–6.

[144] H. Erdogan, "Subspace kernel discriminant analysis for speech recognition," in *COST278 and ISCA Tutorial and Research Workshop (ITRW) on Robustness Issues in Conversational Interaction*, 2004.

[145] H. Choi, R. Gutierrez-Osuna, S. Choi, and Y. Choe, "Kernel oriented discriminant analysis for speaker-independent phoneme spaces," in *19th IEEE International Conference on Pattern Recognition (ICPR) 2008*, 2008, pp. 1–4.

[146] M. S. Kim, I. L. H. Yang, and H. J. Yu, "Kernel multimodal discriminant analysis for speaker verification," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2010*, 2010, pp. 4498–4501.

[147] N. Aronszajn, "Theory of reproducing kernels," *Transactions of the American mathematical society*, vol. 68, no. 3, pp. 337–404, 1950.

[148] D. Cai, X. He, and J. Han, "Speed up kernel discriminant analysis," *The VLDB Journal*, vol. 20, no. 1, pp. 21–33, Feb 2011.

[149] T. Matsui and S. Furui, "Comparison of text-independent speaker recognition methods using vq-distortion and discrete/continuous hmm's," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 3, pp. 456–459, July 1994.

[150] F. Martinez, D. Tapias, and J. Alvarez, "Towards speech rate independence in large vocabulary continuous speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 1998*, vol. 2, May 1998, pp. 725–728.

[151] M. A. Siegler and R. M. Stern, "On the effects of speech rate in large vocabulary speech recognition systems," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP) 1995*, vol. 1, May 1995, pp. 612–615.

[152] H. Yasuda and M. Kudo, "Speech rate change detection in martingale framework," in *12th International Conference on Intelligent Systems Design and Applications (ISDA)*, November 2012, pp. 859–864.

[153] R. E. Crochiere, "A weighted overlap-add method of short-time fourier analysis/synthesis," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 28, no. 1, pp. 99–102, 1980.

[154] S. Roucos and A. M. Wilgus, "High quality time-scale modification for speech," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) 1985*, vol. 10, 1985, pp. 493–496.

[155] K. Rao and B. Yegnanarayana, "Prosody modification using instants of significant excitation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 3, pp. 972–980, May 2006.

[156] B. Sharma and S. R. M. Prasanna, "Faster prosody modification using time scaling of epochs," in *Annual IEEE India Conference (INDICON) 2014*, 2014, pp. 1–5.

[157] N. Morgan and E. Fosler-Lussier, "Combining multiple estimators of speaking rate," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 1998*, vol. 2, 1998, pp. 729–732.

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

[159] R. Vogt and S. Sridharan, "Explicit modelling of session variability for speaker verification," *Computer Speech & Language*, vol. 22, no. 1, pp. 17–38, 2008.

[160] P. Kenny, "Joint factor analysis of speaker and session variability: Theory and algorithms," CRIM, Montreal, Tech. Rep. CRIM-06/08-13, 2005.

[161] A. Lanitis, "A survey of the effects of aging on biometric identity verification," *International Journal of Biometrics*, vol. 2, no. 1, pp. 34–52, 2010.

[162] H. Beigi, "Effects of time lapse on speaker recognition results," in *16th International Conference on Digital Signal Processing 2009*, July 2009, pp. 1–6.

[163] F. Kelly, A. Drygajlo, and N. Harte, "Speaker verification with long-term ageing data," in *Biometrics (ICB), 2012 5th IAPR International Conference on*, March 2012, pp. 478–483.

[164] J. P. Campbell, W. Shen, W. M. Campbell, R. Schwartz, J. F. Bonastre, and D. Matrouf, *Forensic speaker recognition.* IEEE Signal Processing Magazine, 2009, vol. 26, no. 2.

[165] Y. Matveev, *Speech and Computer: 15th International Conference, SPECOM 2013, Pilsen, Czech Republic, September 1-5, 2013. Proceedings.* Springer International Publishing, 2013, ch. The Problem of Voice Template Aging in Speaker Recognition Systems, pp. 345–353.

# List of Publications

## Journal Publications

1. **Rohan Kumar Das** and S. R. M. Prasanna, "Exploring Different Attributes of Source Information for Speaker Verification with Limited Test Data", *Journal of Acoustic Society of America (JASA)*, vol. 140, no. 1, pp. 184-190, July 2016.

2. **Rohan Kumar Das**, Sarfaraz Jelil and S. R. M. Prasanna, "Development of Multi-Level Speech based Person Authentication System", *Journal of Signal Processing Systems, Springer*, vol. 88, pp. 259-271, September 2017.

3. **Rohan Kumar Das**, Akhil Babu Manam and S. R. M. Prasanna, "Exploring Kernel Discriminant Analysis for Speaker Verification with Limited Test Data", *Pattern Recognition Letters (PRL), Elsevier*, vol. 98, pp. 26-31, October 2017.

4. **Rohan Kumar Das** and S. R. M. Prasanna, "Speaker Verification from Short Utterance Perspective: A Review", *IETE Technical Review* (Accepted on $15^{th}$ July 2017).

5. **Rohan Kumar Das** and S. R. M. Prasanna, "Investigating Text-independent Speaker Verification Systems Under Varied Data Conditions", *IEEE Transactions on Information Forensics and Security* (Under Revision).

6. **Rohan Kumar Das**, Bidisha Sharma and S. R. M. Prasanna, "Significance of Duration Modification for Speaker Verification Under Mismatch Speech Tempo Condition", in *International Journal of Speech Technology, Springer* (Under Revision).

7. **Rohan Kumar Das**, Sarfaraz Jelil and S. R. M. Prasanna, "Exploring Text-constraint Models and Source Information for Long-enrollment with Short-test Speaker Verification", *Circuits, Systems and Signal Processing (CSSP), Springer* (Under Review).

### Book Chapters

1. **Rohan Kumar Das** and S. R. M. Prasanna, "Speaker Verification for Variable Duration Segments and the Effect of Session Variability", *Lecture Notes in Electrical Engineering, Springer*, vol. 347, Chapter 16, pp. 193-200, 2015.

### Conference Publications

1. **Rohan Kumar Das**, Abhiram B , S. R. M. Prasanna and A. G. Ramakrishnan,, "Combining Source and System Information for Limited Data Speaker Verification", in *Interspeech 2014*, Singapore, September 2014.

2. **Rohan Kumar Das**, Debadatta Pati and S. R. M. Prasanna, "Different Aspects of Source Information for Limited Data Speaker Verification", in $21^{st}$ *National Conference on Communications (NCC) 2015*, IIT Bombay, February 2015.

3. **Rohan Kumar Das**, Sarfaraz Jelil and S. R. M. Prasanna, "Significance of Constraining Text in Limited Data Text-independent Speaker Verification", in *International Conference on Signal Processing and Communications (SPCOM) 2016*, IISc Bangalore, June 2016.

4. **Rohan Kumar Das**, Sarfaraz Jelil and S. R. M. Prasanna, "Exploring Session Variability and Template Aging in Speaker Verification for Fixed Phrase Short Utterances", in *Interspeech 2016*, San Francisco, September 2016.

5. **Rohan Kumar Das** and S. R. M. Prasanna, "Text-independent Speaker Verification with Limited Test Data from the Perspective of Practical Systems", in $2^{nd}$ *Doctoral Consortium, Interspeech 2016*, ICSI, Berkeley, California, September 2016.

6. **Rohan Kumar Das**, "Incorporating Source Features, Acoustic-phonetic Information and Suitable Pattern Recognition Approach for Limited Test Data Speaker Verification", in $3^{rd}$ *Doctoral Consortium, Interspeech 2017*, KTH Sweden, Stockholm, Sweden, August 2017.

**Other related publications during thesis work**

**Journal Publications**

1. Rajib Sharma, S. R. M. Prasanna, Ramesh K. Bhukya and **Rohan Kumar Das**, "Analysis of the Intrinsic Mode Functions for Speaker information" , *Speech Communication, Elsevier*, vol. 91, pp. 1-16, July 2017.

2. **Rohan Kumar Das**, Hrishikesh Dutta, Sukumar Nandi and S. R. M. Prasanna, "An Overview of Digital Audio Steganography", *IET Signal Processing* (Under Review).

3. **Rohan Kumar Das**, Sarfaraz Jelil and S. R. M. Prasanna, "Multi-style Speaker Recognition Database in Practical Conditions", *International Journal of Speech Technology, Springer* (Under Revision).

**Conference Publications**

1. Subhadeep Dey, Sujit Barman, Ramesh K. Bhukya, **Rohan Kumar Das**, Haris B. C., S. R. M. Prasanna and Rohit Sinha, "Speech Biometric Based Attendance System"', in $20^{th}$ *National Conference on Communications (NCC) 2014*, IIT Kanpur, February 2014.

2. Ramesh K., S. R. M. Prasanna and **Rohan Kumar Das**, "Significance of Glottal Activity Detection and Glottal Signatures for Text-Dependent Speaker Verification", in *International Conference on Signal Processing and Communications (SPCOM) 2014*, IISc Bangalore, July 2014.

3. Sarfaraz Jelil, **Rohan Kumar Das**, K. Amitab, F. Pyrtuh, L. J. Singh and S. R. M. Prasanna, "Exploring Speaker Modeling Techniques for Short Pass-Phrase Based Person Authentication System", in *International Conference on Computing and Communication Systems (I3CS'15)*, NEHU Shillong, April 2015.

4. Sarfaraz Jelil, **Rohan Kumar Das**, Rohit Sinha and S. R. M. Prasanna, "Speaker Verification Using Gaussian Posteriorgrams on Fixed Phrase Short Utterances", in *Interspeech 2015*, Germany, September 2015.

5. Ashutosh Pandey, **Rohan Kumar Das**, Nagaraj Adiga, Naresh Gupta and S. R. M. Prasanna, "Significance of Glottal Activity Detection for Speaker Verification in Degraded and Limited Data Condition" in *IEEE TENCON 2015*, Macau, November 2015.

6. Deepshikha Mahanta, Anupama Paul, Ramesh K. Bhukya, **Rohan Kumar Das**, Rohit Sinha and S. R. M. Prasanna, "Warping Path and Gross Spectrum Information for Speaker Verification under Degraded Condition", in $22^{nd}$ *National Conference on Communications (NCC) 2016*, IIT Guwahati, March 2016.

7. Anupama Paul, **Rohan Kumar Das**, Rohit Sinha and S. R. M. Prasanna, "Countermeasures to Handle Record and Replay Attacks in Practical Speaker Verification Systems", in *International Conference on Signal Processing and Communications (SPCOM) 2016*, IISc Bangalore, June 2016.

8. Salil Mamodiya, Lav Kumar, **Rohan Kumar Das** and S. R. M. Prasanna, "Exploring Acoustic Factor Analysis for Limited Test Data Speaker Verification", in *IEEE TENCON 2016*, Singapore, November 2016.

9. Akhil Babu Manam, Tummala Sai Revanth, **Rohan Kumar Das** and S. R. M. Prasanna, "Speaker Verification using Acoustic Factor Analysis with Phonetic Content Compensation in Limited and Degraded Test Conditions", in *IEEE TENCON 2016*, Singapore, November 2016.

10. Kuruvachan K. George, **Rohan Kumar Das**, Sarfaraz Jelil, K. Arun Das, C. Santhosh Kumar, S. R. M. Prasanna and Ashish Panda, "AMRITATCS-IITGUWAHATI Combined System for the Speakers in the Wild (SITW) Speaker Recognition Challenge", in *IEEE TENCON 2016*, Singapore, November 2016.

11. Sarfaraz Jelil, **Rohan Kumar Das**, S. R. M. Prasanna and Rohit Sinha, "Role of Voice Activity Detection Methods for the Speakers in the Wild Challenge", in $23^{rd}$ *National Conference on Communications (NCC) 2017*, IIT Madras, March 2017.

12. Sarfaraz Jelil, **Rohan Kumar Das**, S. R. M. Prasanna and Rohit Sinha, "Spoof Detection Using Source, Instantaneous Frequency and Cepstral Features", in *Interspeech 2017*, Stockholm, Sweden, August 2017.

13. Nagendra Kumar, **Rohan Kumar Das**, Sarfaraz Jelil, Dhanush B. K., H. Kashyap, K. S. R. Murty, S. Ganapathy, Rohit Sinha and S. R. M. Prasanna, "IITG-Indigo System for NIST 2016 SRE Challenge", *Interspeech 2017*, Stockholm, Sweden, August 2017.

14. **Rohan Kumar Das** and S. R. M. Prasanna, "Investigating Text-independent Speaker Verification from Practically Realizable Perspective", in *IEEE TENCON 2017*, Penang, Malaysia, November 2017 (Accepted).