APPROACHES FOR ROBUST TEXT-DEPENDENT SPEAKER VERIFICATION UNDER DEGRADED CONDITIONS

Ramesh Kumar Bhukya

APPROACHES FOR ROBUST TEXT-DEPENDENT SPEAKER VERIFICATION UNDER DEGRADED CONDITIONS

A

Thesis submitted

for the award of the degree of

DOCTOR OF PHILOSOPHY

By

RAMESH KUMAR BHUKYA

DEPARTMENT OF ELECTRONICS AND ELECTRICAL ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY GUWAHATI

GUWAHATI - 781 039, ASSAM, INDIA

April 2019

Certificate

This is to certify that the thesis entitled "**Approaches for Robust Text-Dependent Speaker Verification Under Degraded Conditions**", submitted by **RAMESH KUMAR BHUKYA** (126102001), a research scholar in the *Department of Electronics and Electrical Engineering, Indian Institute of Technology Guwahati*, for the award of the degree of **Doctor of Philosophy**, is a record of an original research work carried out by him under my supervision and guidance. The thesis has fulfilled all requirements as per the regulations of the institute and in my opinion has reached the standard needed for submission. The results embodied in this thesis have not been submitted to any other University or Institute for the award of any degree or diploma.

Dated:

Guwahati.

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

To

Lord Sri Venkateswara Swamy

for His blessings

&

My Mother

Late RATNAMMA, for whom I could not do anything

&

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

for his guidance and inspiration

&

My dear wife **Saahithi**

for the inspiration that you gave and the sacrifies you have made

during this period

&

My **parents** and **family members**

for their endless love, support, encouragement and blessings

Acknowledgements

I am obliged to GOD for His divine guidance and blessings. I solely dedicate my thesis to Lord Sri Venkateswara Swamy.

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

First and foremost, I express my deepest and sincerest gratitude to my research supervisor, **Prof. S. R. Mahadeva Prasanna** for providing me an opportunity to work under his guidance. It is very difficult to describe my feelings in words to acknowledge my supervisor for his continuous guidance, inspiration, encouragement, constant motivation and support throughout the doctoral studies and in all aspects. I am very much thankful to him for transforming me from an unstructured form to a structured form in every aspect of my life. His insightful comments and weekly-based research discussions immensely helped me in improving the quality of my research work and the manuscripts. It would be completely impossible for me to bring the research as well as the thesis to this level without his suggestions and motivation to collaborate with my senior colleagues and their personal support and advice at the initial stage of my research work. I am wholeheartedly grateful to him for providing me the immense facilities and creating an excellent atmosphere for doing research in the EMST and Signal Informatics Laboratory and the freedom of work he has given to me.

I am thankful to my doctoral committee members **Prof. Samarendra Dandapat, Prof. Rohit Sinha** and **Dr. Suresh Sundaram** for their encouragement and valuable suggestions on my work. Their constructive assessment during my seminars helped me to improve the work. I would like to thank faculty members and the office staff of the Department of Electronics and Electrical Engineering, IIT Guwahati, for their direct and indirect help in carrying out this research work. I am very much thankful to Mr. Sanjib Das (Scientific Officer) for his enormous help whenever required. I would like to express my sincere gratitude to Dr. L N Sharma (Technical Officer) for his kind help, support and maintaining the EMST laboratory very smoothly.

I wish to express my warm and sincere special thanks to my seniors Dr. Gayadhar Pradhan, Dr. Haris B. C., Dr. Rajib Sharma, Dr. Biswajit Dev Sarma and Dr. Banriskhem K. Khonglah, for making me understand different concepts during the initial period for their personal support and advice on various technical discussions and suggestions at the initial stage of my research work.

I convey my gratitude to Mr. Sarfaraz Jelil, Mr. Vikram CM and Mr. Sishir Kalita for carefully correcting my mistakes in the papers. I would also like to thank my seniors Dr. Deepak K. T., Dr. S. Shahnawazuddin, Dr. Nagaraj Adiga, Dr. Abhishek Sharma, Dr. Venkata Ramana Kasi, Dr. Subhadeep Dey, Dr. Sunil Y, Dr. Suman Deb and Mr. Hari Krishna Veldandi for their help at different times. In my daily work, I have been blessed with a friendly and cheerful group of lab mates. I am fortunate to have friends Dr. Ramanjaneyulu Unnava, Dr. K Suresh, Dr. Ajay Kumar Maddirala, Dr. Nagendra Kumar, Mr. Subhasis Mandal, Mr. Balaji Rao Katika, Dr. Rajesh Tripathy, Dr. Anurag Singh, Mr. Sree Ram Ganji, Mr. Akhilesh Dubey, Mr. Eedara Prabhakar, Himakshi Chaudhary, Protima Nomo Sudro, Vineeta Das and Sukanya Biswas who helped me in different times. I am thankful to my friends Biju Pegu, Mawsumi Debnath, Meghamallika Sarma, Dr. Bidisha Shrama, Dr. Rohan Kumar Das, Abhishek Dey, Sujit Barman, Chafidul, Soumar, Nazibur, Deepshikha Mahanta, Anupama Paul, Irani, Pallavi, Shikha Baghel, Mrinmoy Bhattacharjee, Moakala Tzudir, Saswati Rabha, Tilendra Choudhary, and all other members in the EMST, Signal Informatics and Signal Processing Laboratory. I would again like to thank Sarfaraz for reading the thesis patiently and carefully for making corrections related to writing.

I am thankful to my wife for her sacrifice and support. I am also thankful to Dr. Sunil Kumar, Machepalli B V Sri Harsha and Mr. Krishna Pavan Inala for their love and support at every stage of my ups and downs too, and I thank the people around me.

I attribute this achievement to my parents and parents-in-law for their constant blessings, support, silent prayers for my success and moreover, making me stand in this position, for which I will be indebted to them forever. Especially, my sincere thanks to my younger brother Rajesh Kumar Bhukya, who continuously cared for me for my well being since childhood and making me strong enough every time to fight back in life during my struggles and ups and downs too, and my younger brother's wife Usha, also always stands for me, no matter whatever happens. Further, I would offer my sincere thanks to MHRD, Govt. of India for providing the fellowship to carry out my PhD thesis work. Finally, I would like to thank my well-wishers and friends for their care and love that made my days beautiful in the beautiful campus.

Ramesh Kumar Bhukya

Indian Institute of Technology Guwahati

April 2019

Abstract

The primary goal of this thesis work is to develop a robust *Text-Dependent Speaker Verification* (TDSV) system by exploring different approaches for achieving improved performance under clean and degraded conditions. To achieve this, in this thesis, three different suitable directions are explored.

First, a TDSV system is developed by using combined temporal and spectral speech enhancement for enhancing speech regions embedded in background noise followed by energy based *end point detection* (EPD). Another direction is developed based on robust features extracted from *Hilbert Spectrum* (HS) of the speech signal, constructed from its *Amplitude Modulation - Frequency Modulation* (AM-FM) components, or *Intrinsic Mode Functions* (IMFs) obtained using a non-linear and non-stationary data analysis technique called *Modified Empirical Mode Decomposition* (MEMD), in characterizing the speakers. The third direction is the use of an improved *vowel-like region onset point* (VLROP) and *vowel-like region end point* (VLREP) detection method for better end point detection. The evidences extracted from excitation source and vocal tract system information is used for this method. The *dominant resonant frequency* (DRF) is used to eliminate spurious vowel-like regions (VLRs) in the background noise. *Foreground speech segmentation* (FSS) is used to remove spurious VLRs in the background speech region. For better localization of the speech region *glottal activity detection* (GAD) is explored and also used to detect the sonorant consonants and missed VLRs. To include an unvoiced consonant, *obstruent region detection* (OBS) is done at the begin of first VLR and at the end of last VLR. Finally, *speech duration knowledge* (SDK) is used to further refine the output of begin and end points. Finally, all these proposed methods are put together for having improved system performance.

The proposed approaches are experimented for the TDSV task, considering three different sentences, each of the RSR2015 and the IITG databases for validating the experimental findings. To evaluate the practical utility of the proposed features, they are tested not only for clean speech, but also under Babble noise of varying strength, and environmental noise.

The major contributions of this thesis are as follows:

- Investigating the combined temporal and spectral speech enhancement technique followed by energy based end point detection for TDSV under degraded conditions.
- Investigating the HS features in combination with MFCCs for TDSV task.
- A method for begin and end points detection using some speech specific knowledge. This method uses both excitation source and vocal tract system information to reduce spurious detection.
- Combining different noise robust approaches at different stages of a TDSV to improve the system performance.

Keywords: Text-Dependent Speaker Verification, Speech Enhancement, MEMD, IMFs, Hilbert Spectrum, End Point Detection.

Contents

List of Figures

List of Tables

List of Acronyms

AM-FM	Amplitude Modulation - Frequency Modulation
CCA	Canonical Correlation Analysis
CMS	Cepstral Mean Normalization
CMVN	Cepstral Mean and Variance Normalization
CTI	Computer Telephone Interface
CVN	Cepstral Variance Normalization
DET	Detection Error Tradeoff
DRF	Dominant Resonant Frequency
DCT	Discrete Cosine Transform
DCTILPR	Discrete Cosine Transform of Integrated Linear Prediction Residual
DFT	Discrete Fourier Transform
DTW	Dynamic Time Warping
ED	Euclidean Distance
EPD	End Point Detection
EEMD	Ensemble Empirical Mode Decomposition
EER	Equal Error Rate
EMD	Empirical Mode Decomposition
FAR	False Acceptance Rate
FE	Feature Extraction
FFT	Fast Fourier Transform
FOGD	First Order Gaussian Differentiator
F_0	Fundamental Frequency
FPGA	Filed Programmable Gate Array
FRR	False Rejection Rate

GAD	Glottal Activity Detection
GAs	Glottal activity Regions
GCIs	Glottal Closure Instants
GFD	Glottal Flow Derivatives
GMM	Gaussian Mixture Models
HE	Hilbert Envelope
HE of LP	Hilbert Envelope of LP residual
HiLAM	Hierarchical multi-Layer Acoustic Model
HMM	Hidden Markov Models
HNGD	Hilbert Envelope of Numerator Group Delay
HS	Hilbert Spectrum
ICEEMDAN	Improved Complete Ensemble Empirical Mode Decomposition
IE	Identification Error
ILPR	Integrated LP residual
IMFs	Intrinsic Mode Functions
IR	Identification Rate
ISDN-PRI	Integrated Services Digital Network - Primary Rate Interface
i-vector	Identity Vector
IVR	Interactive Voice Response
KLT	Karhuen-Loeve Transform
LDA	Linear Discriminant Analysis
LP	Linear Prediction
LPCs	Linear Predictive Coefficients
LPCCs	Linear Predictive Cepstral Coefficients
LSFs	Linear Spectral Frequencies
MDA	Multi-band Demodulation Analysis
mDCF	Minimum Detection Cost Function
MFCCs	Mel-Frequency Cepstral Coefficients
MEMD	Modified Empirical Mode Decomposition
MMSE	Minimum Mean Square Estimation

MMSE-LSA	MMSE-Log Spectral Amplitude
MPDSS	Mel Power Difference of Spectrum in Sub-bands
NIST	National Institute of Standards and Technology
OBS	Obstruent Region Detection
OWP	Optimal Warping Path
PBX	Private Branch Exchange
PLPs	Perceptual Linear Prediction
PSTN	Public Switched Telephone Network
RMFCCs	Residual Mel Frequency Cepstral Coefficients
SDK	Speech Duration Knowledge
SI	Speaker Identification
SNR	Signal to Noise Ratio
SR	Speaker Recognitions
SREs	Speaker Recognition Evaluations
STFT	Short-time Fourier Transform
STSA	Short-time Spectral Amplitude
SV	Speaker Verification
SVD	Singular Value Decomposition
TDSV	Text-Dependent Speaker Verification
TEO	Teager Energy Operator
TISV	Text-Independent Speaker Verification
T-matrix	Total Variability Matrix
UBM	Universal Background Model
VAD	Voice Activity Detection
VLRs	Vowel-Like Regions
VLREP	Vowel-Like Region End Point
VLROP	Vowel-Like Region Onset Point
VMD	Variational Mode Decomposition
VoIP	Voice over Internet Protocol
VOP	Vowel Onset Point

VTCCs	Vocal Tract Mel Cepstral Coefficients
WCCN	Within Class Covariance Normalization
ZFF	Zero Frequency Filtering
ZFFS	Zero Frequency Filter Signal

List of Symbols

$\boldsymbol{\Delta\Delta}$	Acceleration (Delta Delta) coefficients
$\boldsymbol{\zeta_k, \gamma_k}$	a priori SNR and a posteriori SNR, respectively
$\boldsymbol{\Delta F_K, \Delta E_K}$	Absolute deviation of instantaneous frequency & energy feature set obtained from K^{th} IMFs of all speech frames of a speech signal
$\boldsymbol{x(t), X(w)}$	An arbitrary continuous-time signal, and its Fourier transform
$\boldsymbol{F_K^i, E_K^i}$	An arbitrary K-dimensional feature vectors extracted from the i^{th} speech frame
$\boldsymbol{\Delta F_K^i, \Delta E_K^i}$	An arbitrary K-dimensional feature vectors extracted from K^{th} IMFs of the i^{th} speech frame in absolute deviation of instantaneous frequency & energy feature set
$\boldsymbol{\sigma F_{K-1}^i, \sigma E_{K-1}^i}$	An arbitrary K-1 dimensional correlation feature vectors extracted from K^{th} IMFs of the i^{th} speech frame in correlation between instantaneous frequencies & energies feature set
$\boldsymbol{\Upsilon_K^i}$	An arbitrary K-dimensional feature vectors extracted from K^{th} IMFs of the i^{th} speech frame in instantaneous energy weighted instantaneous frequency feature set
$\boldsymbol{U_1, V_1}$	Canonical variate pair, for two feature vectors of length N_f
$\boldsymbol{\sigma F_K^i, \sigma E_K^i}$	Correlation between instantaneous frequencies and energies obtained from K^{th} number of IMFs of the i^{th} speech frame from a speech
$\boldsymbol{\sigma F_{K-1}, \sigma E_{K-1}}$	Correlation between instantaneous frequencies & energies feature set obtained from K-1 number of IMFs of all speech frames of a speech
$\boldsymbol{C_M, C_F}$	Cost of a Miss (a genuine claim rejected), and cost of a False Alarm (an imposter claim is accepted) assign to the event
$\boldsymbol{C_{\xi^s}}$	Cost parameter of DCF at any given value of the decision threshold

$\boldsymbol{\xi^s, \xi_0^s}$	Decision threshold for verifying the claim in TDSV system, and a particular value of it's score
$\boldsymbol{S}(K)$	DFT magnitude spectrum of the desired speech
$\boldsymbol{d(n)}$	Difference the speech signal $s(n)$
$\boldsymbol{s}(n)$	Digital speech signal/utterance
$\boldsymbol{\Delta}$	Dynamic (Delta) coefficients
$\boldsymbol{e(n)}$	Error introduced between the desired speech $s(n)$ and the estimated speech $\hat{s}(n)$
$\boldsymbol{i_k}$	Estimated location of speech
$\boldsymbol{\hat{S}_K}$	Estimated speech from degraded speech
$\boldsymbol{a_k(n)}$	Estimating the instantaneous amplitude envelope of any discrete time signal $s(n)$
$\boldsymbol{f_k(n)}$	Estimating the instantaneous frequency of any discrete time signal $s(n)$
$\boldsymbol{\hat{s}(n)}$	Estimation of the speech signal
$\boldsymbol{r_m(n), h_k(n)}$	Final residue, and the K^{th} IMFs of the speech signal $s(n)$
$\boldsymbol{w_f(n)}$	Fine weight function
$\boldsymbol{w_g(n)}$	Gross weight function
$\boldsymbol{h_e(n)}$	Hilbert envelope of LP residual
$\boldsymbol{l_h(n)}$	Hilbert transform of LP residual
$\boldsymbol{H}[\cdot]$	Hilbert transform operator
$\boldsymbol{\Upsilon_K^i}$	Instantaneous energy weighted instantaneous frequency feature vector obtained from K^{th} IMFs of the i^{th} speech frame of a speech signal
$\boldsymbol{\Upsilon_K^i}$	Instantaneous energy weighted instantaneous frequency feature vector obtained from K^{th} IMFs of all speech frames of a speech signal
$\boldsymbol{i_v}$	i-vector representation of speech utterance
$\boldsymbol{a_k}$	LP filter coefficients
$\boldsymbol{l(n)}$	LP residual signal
$\boldsymbol{M}$	Maximum number of IMFs to be extracted after which the MEMD is terminated

$\boldsymbol{F_K^i, E_K^i}$	Mean instantaneous frequency & energy feature vectors obtained from K^{th} IMFs of the i^{th} speech frame of a speech signal
$\boldsymbol{F_K, E_K}$	Mean instantaneous frequency & energy feature set obtained from K^{th} number of IMFs of all speech frame of a speech signal
$\boldsymbol{\hat{L}}$	MMSE-LAS estimator
$\boldsymbol{\hat{M}}$	MMSE-STSA estimator
$\boldsymbol{S_G}$	Number of genuine/true claims that are accepted in TDSV system
$\boldsymbol{S_M}$	Number of genuine/true claims that are rejected in TDSV system
$\boldsymbol{S_F}$	Number of imposter/false claims that are accepted in TDSV system
$\boldsymbol{S_I}$	Number of imposter/false claims that are rejected in TDSV system
$\boldsymbol{g}(k)$	Numerator group delay function
$\boldsymbol{m(p)}$	Non-negative weighting coefficient of the warping path
$\boldsymbol{u_d, v_d}$	Optimized and maximized canonical correlation coefficient between U_1 and V_1
$\boldsymbol{y(n)}$	Output of zero frequency resonator
$\boldsymbol{P_T}$	Probability of genuine/true claims
$\boldsymbol{P_M, P_F}$	Probability of Miss, and Probability of Flase Alarm
$\boldsymbol{\Gamma\{V_K^i\} = [cv_1^i, cv_2^i, ..., cv_K^i]^T}$	Refined K-dimensional feature vector extracted from the i^{th} speech frame after applying DCT on cv_K^i
$\boldsymbol{cG_K}$	Refined Sum Log Squared Amplitude feature set
$\mathbf{Sc} = \{\mathbf{sc_1, sc_2, ..., sc_S}\}$	Set of DTW scores obtained by verifying the claims of the test utterances against the speaker model utterances
$\boldsymbol{\lambda}$	Slope parameter
$\boldsymbol{H(K)}$	Spectral gain function of the estimator
$\boldsymbol{w_2(n)}$	Tapering window
$\boldsymbol{S_t(z)}$	Temporally enhanced speech signal
$\boldsymbol{X_{D1}^{N_f}}$	The entire feature set of $N_f \times D_1$ size
$\boldsymbol{Y_{D2}}$	The feature set of dimension D_2, obtained from N_f speech frames
$\boldsymbol{H(f,n)}$	The instantaneous amplitude envelope $a_k(n)$, and frequency $f_k(n)$ of $h_k(n)$

$\boldsymbol{f_k^i(n), \lvert a_k^i(n)\rvert^2}$	The instantaneous frequencies and energies respectively of $h_k^i(n)$
$\boldsymbol{S_{UBM}}$	The mean super-vector of the UBM data set, obtained by building a GMM on the combined feature set of all its utterances
$\boldsymbol{S_u}$	The mean super-vector of any speech utterance obtained from S_{UBM}
$\boldsymbol{N}$	The number of detected VLRs
$\boldsymbol{N_f, D_1}$	The rows correspond to feature vectors of dimension D_1, obtained from N_f speech frames
$\boldsymbol{\hat{y}(n)}$	The speech signal obtained after removing the trend in the zero frequency filtered signal
$\boldsymbol{h_k^i(n)}$	The i^{th} speech frame of the K^{th} IMF of speech signal $s(n)$
$\boldsymbol{s^i(n)}$	The i^{th} speech frame of $s(n)$
$\boldsymbol{V_i}$	The i^{th} VLROP in the number of speech samples
$\boldsymbol{DTW(\phi_X(p), \phi_Y(p))}$	The shortest time spectral distortion
$\boldsymbol{C}$	The VLROP having the maximum average euclidean distance of the speech utterance
$\boldsymbol{M_\phi}$	The warping path normalizing factor
$\boldsymbol{R_w(z)}$	The weighted LP residual
$\boldsymbol{DTW_\phi(X, Y)}$	Total accumulated cost distance of warping path
$\boldsymbol{N_K}$	The total number of epochs located in speech frame
$\boldsymbol{\xi}$	The values of the spectral components of the enhanced speech
$\boldsymbol{T_v}$	Total Variability matrix
$\boldsymbol{\hat{i}_t r, \hat{i}_t s}$	Two processed i-vectors, corresponding to two utterances (one from the train data set, other from the test data set)
$\boldsymbol{S_1^A(n), S_2^A(n)}$	Two utterances of speaker A from n number of speakers
$\boldsymbol{S_1^B(n), S_2^B(n)}$	Two utterances of speaker B from n number of speakers
$\boldsymbol{w(n)}$	Total weight function is obtained by multiplying $w_g(n)$ with $w_f(n)$
$\boldsymbol{\hat{y}(n)}$	Zero frequency filtered signal
$\boldsymbol{w_1(n)}$	Zero time window

1

Introduction

Contents

Objective of the Thesis

The objective of this thesis work is to develop a robust text-dependent speaker verification (TDSV) system by exploring approaches to obtain better performance under clean and degraded conditions. To achieve this, three different techniques are proposed. First, by using combined temporal and spectral enhancement method, the speech regions embedded in background noise are enhanced. Further, using enhanced speech signals, the end points are detected for the development of TDSV system. Second approach involves the exploration of features extracted from Hilbert Spectrum (HS) of the Intrinsic Mode Functions (IMFs) obtained from Modified Empirical Mode Decomposition (MEMD) of the speech signal. It is noticed that only first few IMFs are useful and necessary for characterizing speaker information and achieving enhanced performance in combination with Mel frequency cepstral coefficients (MFCCs). The third one utilizes the speech-specific knowledge for the robust end point detection, which is suitable for both clean and degraded conditions. Finally, a combined system is developed by exploiting the evidences from these three proposals and a common TDSV framework is implemented.

1.1 Speaker Verification System

Speech is the primary mode of communication for human beings. The speech signal can be easily obtained which makes it a very attractive signal at low cost for the scientific community to use in different human-machine applications [2–4]. Presence of speaker specific information makes speech useful as biometric feature to recognize or authenticate a person based on his/ her voice samples. The biometric feature contains a multitude of information like the speaker's age, height, emotion, style of speech delivery, accent, health and physiological disorders, identity, vocabulary usage, etc. [2–4]. There are many services and applications in which speaker verification could be used [5,6]. Some of the common tasks are banking over telephone network, database access services, telephone shopping [7], identity verification for access to a website [8], voice mail [9], and judicial tasks such as surveillance of suspects and other forensic applications [2–4, 10, 11] etc.

Among different applications involving speech, *speaker verification* (SV) has expanded remarkably over the years since its inception. SV refers to a technology that enables the machines to recognize persons using his/ her speech samples. SV technology is one of the leading areas in speech processing [2–4, 12–14]. SV systems are practically more distinguished and emerging area for having deployable systems with real-world applications. Based on the constraint imposed on the text content of the speech utterance used, SV systems can be classified into *text-independent speaker verification* (TISV) and *text-dependent speaker verification* (TDSV) systems [5, 15–17]. In a TDSV system, the text is fixed and the users have to utter the same text during training and testing. Therefore, a small amount of data involving a speech utterance of 1-3 secs is used. On the other hand, the TISV system does not put any restriction on text content of the speech utterance during training as well as testing speech. TISV system requires 2-3 minutes of speech data during training and 15-45 secs data during testing phase. In TDSV, the system takes the user's speech utterance and the identity claim as input to the system and decides whether the input speech utterance belongs to the claimed user or not. In this thesis work, all experimental studies are presented for a TDSV system suited for the deployable systems under clean and degraded conditions.

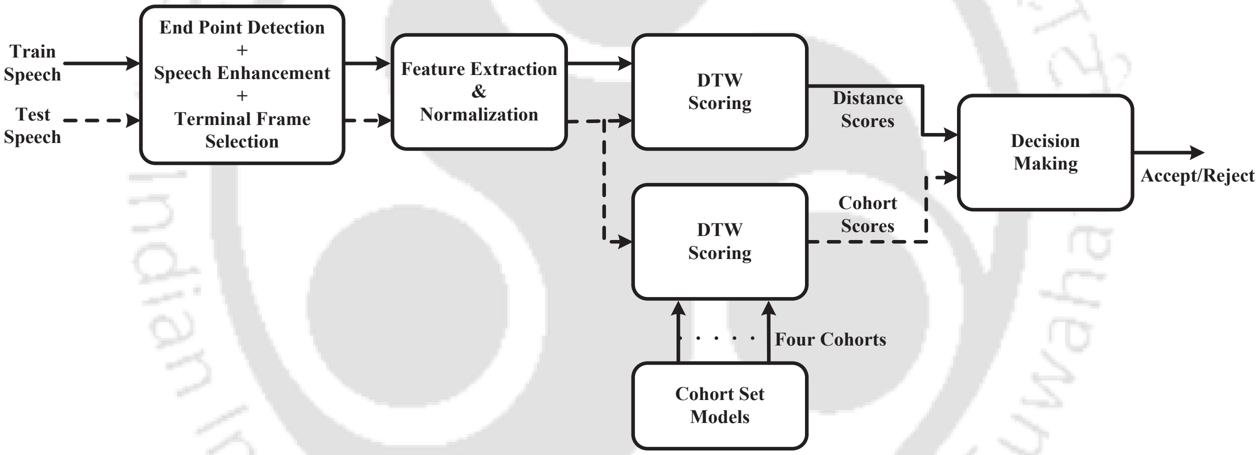


Figure 1.1: Modular representation of the TDSV system.

1.2 Text-Dependent Speaker Verification

The recent developments in the area of TDSV system has shown different possibilities for the biometric application oriented systems. A generic block diagram of the SV system is shown in the Figure 1.1. The overall system comprises of two phases, namely, the training (enrollment) phase and the testing (verification) phase. The training phase of a TDSV system accepts the speech data from the target speakers and creates the corresponding (target models) speaker models. The testing phase accepts the test speech utterances and their corresponding speaker claim as inputs to the system, compares against each of the claims and makes an accept/ reject decision as an output. The different blocks involved in the development of enrollment and verification of a TDSV system are shown individually. The purpose of each block and the most commonly used methods to understand are described in the following sections.

1.2.1 End Point Detection

This is an active research topic in the field of TDSV, because it is a prime factor that influences the system performance under degraded conditions. During training and testing phases of a SV system, speech data is first processed through an *end point detection* (EPD) used to remove silence portions in the begin and end of speech [18–20]. The task of removing silence portions from raw speech is traditionally performed with the help of *Energy based* EPD which involves marking the starting and end points of the speech utterance. The presence of background noise, background speech and other environmental conditions makes it difficult to locate the begin and end points. Therefore, each degraded speech conditions have their own type of characteristics depending on accuracy, complexity and robustness. Most of the SV systems use short-term energy based EPD methods because of their simplicity in implementation and less computational complexity. This method is determined for each analysis of speech and non-speech frames having frame energy above certain threshold are considered as speech frames and if not then that frame is discarded. However, under degraded conditions this method may fail and degrade the SV performance. Therefore, in order to handle the degraded speech conditions, different feature selection methods are proposed. Some of them are periodicity, statistical based EPD, vowel and non-vowel like onset and offset point speech region selection and pattern recognition based EPD methods [12, 21–24]

1.2.2 Speech Enhancement

Speech enhancement is basically exploited for removing the effect of low level SNRs at the signal level before feeding to end point detection. Nowadays, people use mobiles, telephones, internet,etc, as a communication devices almost as a primary need. The basic demands of customers are high coverage and quality. However, under degraded conditions, it can seriously damage the service quality. Hence, speech enhancement aims to recognize the identity of the speaker by enhancing the overall perceptual quality and intelligibility of degraded speech using several signal processing techniques [25, 26]. The noisy speech is first segmented and then windowed by Hamming window. Then *Discrete Fourier Transform* (DFT) of the segmented and noisy speech is taken and given to the estimation block to estimate the noise region during speech pause and find the noise spectrum. If the noise estimation is too low, unwanted noise residual will be recognizable, if the noise estimate is too high, speech will be distorted. Hence, the speech enhancement block is used to enhance the noisy speech spectrum to generate clean speech spectrum and then we take *inverse Fast Fourier Transform* (FFT) to obtain the enhanced speech [27, 28].

1.2.3 Feature Extraction

From the detected speech regions, it is necessary to extract relevant features from the raw speech at reduced data rate. The most successful short-term MFCC features are used to capture the changes in vocal tract as well as F_0 information more accurately and highly reflected in the final system performance. The basic role of FE is to compute relevant features from speech which carry speaker-specific information. There are various types of spectral features used for extracting speaker information like *Mel frequency cepstral coefficients* (MFCCs) [29], *linear prediction coefficients* (LPCs), *linear prediction cepstral coefficients* (LPCCs), *linear spectral frequencies* (LSFs) and *perceptual linear prediction* (PLP). The above mentioned features contains significant speaker specific attributes of the speech utterance [30, 31]. The voice source features that characterize the glottal source of a speaker also carry significant speaker-specific information [32, 33]. There are several voice source features used based on linear prediction (LP) residual, residual phase, modulation spectrum, wavelet analysis and parametric glottal flow model parameters. In some cases, modulation frequency method is used to obtains the information regarding speaker rate and style of speaking as a feature for SV system [34, 35]. Some prosodic features like fundamental frequency (F_0) in combination with spectral features are more effective, particularly in degraded speech conditions [36].

The FE process is described as follows: For each 20 ms speech frame, Hamming window is applied to reduce the edge effect while taking the *discrete cosine transform* (DCT) on the speech utterance. For each speech frame, MFCCs are computed using 22 logarithmically spaced Mel filterbank. The speech utterance is parameterized with 13-base MFCCs excluding 0^{th} cepstral coefficient which is used as a feature vector [37]. To improve the system performance, the dynamic coefficients (Δ) and the acceleration coefficients ($\Delta\Delta$) of the MFCCs are computed using the two preceding and two succeeding feature vectors from the current feature vector [38]. Thus, the final feature vector is of 39 dimensions with 13-base MFCCs, 13 ΔMFCCs and 13 $\Delta\Delta$MFCCs. Feature normalization aims to scale the feature vectors for suppressing the effect of unwanted variabilities. After FE, it is essential to normalize the features in order to nullify the common offset for channel/session compensation, which gives better speaker discrimination. Moreover, the *cepstral variance normalization* (CVN) is performed on top of *cepstral mean subtraction* (CMS) to normalize the feature vectors used to remove the channel effect. The obtained signal dependent parameters are called as *cepstral mean-variance normalization* (CMVN) to normalize the features in order to fit zero mean and unit variance distribution [39]. The extracted speaker-specific normalized features can be used either directly or further processed by several methods for speaker modeling for improving the system performance.

1.2.4 Speaker Modeling and Feature Matching

The dynamic programming algorithm provides a many-to-many mapping between train and test speech utterances while minimizing the overall accumulated distance. This minimum accumulated distance is called as *Dynamic Time Warping* (DTW) distance [40,41]. To find the best match between the test feature vector compared with reference to the claimed model reference templates, the DTW algorithm is used. The dynamic warping path minimizes the total distance between the train and test feature vectors effectively due to it's elastic distance measure.

Unlike sophisticated techniques, DTW is not *data hungry*, and can be implemented quickly, which makes it a suitable choice for conducting large sets of experiments. The DTW algorithm finds an *optimal warping path* (OWP) between two feature sets, based on *Euclidean distance* (ED) between train and test feature vectors. The final DTW distance or score is the accumulation of the ED over the OWP. By applying some constraints, the possible paths are limited to a certain limit making the computation feasible and limit the possibility of infinite paths making computation more efficient. Therefore, the DTW score obtained by matching two feature sets corresponding to two different speech utterances reflects the similarity between the two speech utterances. A lesser score indicates higher similarity between the two utterances, and vice versa. Thus, for a fixed phrase/ sentence, if the two speech utterances belong to the same speaker, a low DTW score is expected, whereas a higher score is expected when the utterances belong to different speakers.

1.2.5 Performance Metrics

The performance evaluation of the SV system is based on two standard metrics - *Equal Error Rate* (EER), and *minimum Detection Cost Function* (*m*DCF)-which are mostly used to evaluate the performance of the TISV and TDSV systems [42]. Let $\mathrm{S}c = \{\mathrm{s}c_1, \mathrm{s}c_2, ..., \mathrm{s}c_S\}$ represent the entire set of DTW scores obtained after matching the testing utterances vs. the training utterances. The set $\mathrm{S}c$

is normalized so that $\{\mathrm{sc}_i \in [0,1] \mid i = 1,2,...,S\}$. Let $\xi \in [0,1]$ be the decision threshold, above which the claim of the test utterance against the speaker is accepted. Let S_G denote the number of genuine/ true claims that have been accepted, S_I denote the number of imposter/ false claims that have been rejected, S_M denote the number of genuine/ true claims that have been rejected and S_F denote the number of imposter/ false claims that have been accepted. Therefore, $S = S_G + S_I + S_M + S_F$. We may now define EER and mDCF as follows :

EER (%) : For any given threshold, ξ, the *Miss Rate* is given as $\mathrm{MR} = \frac{S_M}{S_G+S_M} \times 100$ %. The *False Alarm Rate* is given as $\mathrm{FAR} = \frac{S_F}{S_I+S_F} \times 100$ %. At a particular threshold, $\xi = \xi_0$, MR = FAR. This error is known as the EER. In other words, probability of false acceptance and probability of false rejection are equal i.e., EER = MR = FAR, at $\xi = \xi_0$.

m**DCF :** For any given threshold, ξ, the *Probability of Miss* is given as $P_{Miss|Tar} = \frac{S_M}{S_G+S_M}$. The *Probability of False Alarm* is given as $P_{FA|Non-tar} = \frac{S_F}{S_I+S_F}$. Two parameters, C_{Miss} and C_{FA}, assign costs to the event of a *miss* (a genuine claim rejected), and that of a *false alarm* (an imposter claim is accepted) respectively. Also, an *a priori* probability P_{Tar} is assigned, which assumes that out of all the test claims against a speaker, only a fraction P_{Tar} are genuine claims. The cost parameter of the mDCF, at any given ξ, is given by,

$$C_\xi = C_{Miss} \times P_{Miss|Tar} \times P_{Tar} + C_{FA} \times P_{FA|Non-tar} \times (1 - P_{Tar}) \tag{1.1}$$

The mDCF is then given by,

$$m\mathrm{DCF} = \min_{\xi \in [0,1]} C_\xi \tag{1.2}$$

In this work, $C_{Miss} = 10$, $C_{FA} = 1$, and $P_{Tar} = 0.01$ are considered respectively. The minimum of these values are used to evaluate the SV system performance as a measure.

1.3 Motivation for the Present Work

The speech processing community has witnessed significant advances with the development of various tools and techniques suitable for modeling speaker characteristics. These advances in the field of SV system have opened doors towards practical deployable systems using speech based person authentication system with attendance as an application. Multimodal biometric person authentication system has been carried out using SV system [43]. The practical deployment of speech biometric based authentication system has been developed using latest techniques for controlling different household activities that involve various ergonomic constraints including duration of speech utterances with the knowledge of speech and speaker [44].

In [38], an automatic speech biometric based attendance system is developed for marking the student attendance over online telephone based network system and is one of the emerging application-oriented biometric system. The online real-time SV system is implemented using *field programmable gate array* (FPGA) with low-cost and less complexity which allows better performance in terms of execution time and speech duration [45]. Recently, a multi-level SV system is developed which consists of three different modules that are combined in a consecutive manner. The multi-level SV system is ported over a telephone based network through an *interactive voice response* (IVR) system for aiding remote person authentication of a claimed identity. This system is developed and deployed in the institute level successfully [46]. All these research and development works showcase the importance of biometric based SV system as an emerging area for most frequently used biometric features which have got success in practical deployment of the system.

The student attendance system using SV is implemented to be accessible over an online telephone network for marking student attendance using a few pre-decided mobile phones, which is handled by an IVR system that connects the SV system to the external world. The IVR system also guides the users in the enrollment phase as well as testing processes with the help of voice prompts interface. The IVR system call flow is developed using the open source Asterisk software on Linux platform installed in a computer server that consists of Intel Core-i7 processor having 8 cores, 4 GB RAM integrated with *computer telephone interface* (CTI) card. Asterisk software enables the computer to an *integrated services digital network-primary rate interface* (ISDN-PRI) line and used to handle telephone channel calls through the CTI card. The IVR system is provided on a voice-server, which runs the Asterisk software to handle incoming calls and make outgoing calls to and from other voice over internet protocol services or *public switched telephone network* (PSTN). In order to make the system interactive, voice prompts are played at regular intervals to guide and engage the user. When a user calls to access the attendance system, these prompts guide the user through out the process.

Based on the application and requirements, several modifications were made in the call-flow for marking student attendance system. These modifications include introduction of several modules in the call-flow to deal the issues which came on successful practical implementation. The modifications were made in response to the user feedback and care was taken to increase the users efficacy. This system is developed for departmental M.Tech and Ph.D. students is used on a regular basis for attendance purpose at our IITG institute level [38]. Hence, the registered students were allowed to call to voice-server with the designated toll-free number through which the IVR system call-flow helps for enrollment as well as regular testings in a practical scenario to facilitate authentication of a user calling from any remote location [46]. In this thesis work, all studies are presented for text-dependent methods based on template/ model sequence matching techniques. The TDSV systems generally perform better because it can tightly model the characteristics of the specific phonetic content contained in the speech utterance. To make suitable system for practical applications, the system should be designed such that, it should be language, sensor and session independent. This system requires fixed set of text to be spoken and less amount of speech duration, i.e, typically 1-3 seconds during train and test sessions. This kind of system provides more comfort in practical deployment for the users and has potential to deal with less duration of the speech utterances. This made our research proposal of TDSV system with more prominent limited test data and captivating within the context of the proposed system deployment under practical environments.

1.3.1 Motivation Towards Text-Dependent Speaker Verification

National Institute of Standards and Technology (NIST) *Speaker Recognition Evaluations* (SREs) provide the benchmark of the TISV systems in every challenge in terms of accuracy and robustness. These improvements have been strongly supported and motivated by the international bench marking events organized by NIST and large amount of challenging data for evaluations. These evaluations are held annually and the data provided for the challenge vary from year to year [47]. It can be noticed that the duration of the speech utterances are changed from NIST SRE 2004 on-wards to include short utterances of speech data having 10 seconds duration and less [48]. NIST-SREs challenges intentionally provide corrupted database which include different scenarios like channel, language and handset mismatches. Always, these type of databases may not provide accurate system performance. Hence, all these databases altogether depicts biometric based SV system as a prime need for application oriented systems.

Apart from NIST SREs databases, there have been other fixed utterances recent database released for making studies in practical scenarios. The RSR2015 database especially designed for TDSV system with fixed utterances. In recent days environmental conditions are also focusing their research on

TDSV system with fixed set of phrases during train and test speech segments. This type of practical system absolutely depicts fixed phrases of speech segments as a prime need for practical oriented systems.

Apart from the NIST SRE databases, TDSV system provides comfort to the users by using less amount of speech data during train and test phases. However, TDSV can be seen as a subcase of TISV. Based on this, historically, a succession of improvements in the field of *text-independent* benefited the *text-dependent* scenario. Most of the recent developments in TISV case are data intensive, speaker modeling that involve large amount of data for background modeling and other statistical computations. Thankfully, such an effort is supported by the ready supply of speech database from NIST and LDC. To translate these speaker modeling techniques to TD scenario, the difficulty is the non-availability of such a large database collected in TD mode. This limitation has been solved to some extent by the release of recent database on TDSV named as RSR2015 database [14]. The RSR2015 database outlined with TDSV modality has three different parts specific towards text-dependent, short commands and digit sequence based fixed phrases. The RSR2015 database first part is especially dedicated to TDSV framework using 30 short fixed phrases, whereas the second part deals with short commands of 2-3 words which can be used for smart home security application purposes. The third part of the database considers SV using randomly uttered digit sequence based speech utterances for experimental studies. Another database has been collected with reference to the TDSV studies in the EEE department at IITG institute. The database considers TDSV framework, where a fixed set of speech utterances are common across all the users.

To demonstrate and show the fair comparison between the standard and practically collected database, in this thesis, two databases are considered. First, the speech utterances of the RSR2015 database recorded in clean conditions with the same speakers and channels. On the other hand, IITG database is recorded in a natural environment in a somewhat open hall where users could roam around, walk in and out, while using the mobile phones to give their attendances, with possibly their friends talking around. Further, the users have to call to the voice-server using the predefined mobile phones, thus incorporating channel effects in the speech utterances. These databases project the significance of TDSV system in practical deployable scenarios and its scope for application oriented systems in future. Different evaluations are organized with these databases with novel ideas and techniques can be useful for addressing different issues towards practical deployable systems.

1.3.2 Issues in the Development of a TDSV System

The state-of-the-art SV system produces good performance when the speech utterance is of high quality and free from environmental conditions. However, the speech utterances are affected by different degradation like background noise, background speech, speech from other speakers, sensor and channel mismatches, lip smacks, heavy breathing, chirping noise, emotional conditions and other environmental conditions, resulting in degraded speech. To develop a robust TDSV system under degraded conditions, three different directions are identified from the perspective of deployable systems. These directions can be used to have their effect on the overall accuracy of the TDSV system.

- The performance of a TDSV system depends on the accuracy and robustness of end point detection methods [49–51]. However, practical speech recordings always have varying amounts of background noise and this leads to inaccuracy in detecting end points using energy based threshold methods. Therefore, it is necessary to develop end point detection methods that are robust to degraded speech conditions.

- It is required to detect suitable speech regions and further enhance the system performance. Both human and speech processing systems are more effective in clean speech conditions. This is due to the high quality and intelligibility of speech. However, in practical scenarios, the speech utterances are affected by different types of degradation. In such case the end point detection module alone may not be sufficient to improve the perceptual quality of the degraded speech. Hence there is a need to enhance the speech components present in the degraded speech and improve the quality and intelligibility of the speech.

- The performance of the SV system is found to degrade significantly under degraded conditions which makes it important to look for features representing speaker-specific information. The MFCCs capture the vocal tract information embedded in the speech signal. Features which capture other aspects of speaker information, like the glottal source, long-term information (prosody, modulation spectrum, etc.) have been explored. Recently, alternative signal processing approaches like AM-FM components or the IMFs of the speech signal, as obtained from *Empirical Mode Decomposition* (EMD), *Ensemble Empirical Mode Decomposition* (EEMD), *Improved Complete Ensemble Empirical Mode Decomposition with Adaptive Noise* (ICEEMDAN) and other methods, in characterizing speaker-specific information.

- Several varieties of noise with recorded speech data makes noise/ noisy speech modeling techniques critical for the SV task. The duration of speech utterance in TDSV case is around 2-3 seconds, long non-speech regions may degrade the performance with too many environmental conditions as we don't have good techniques to remove noise. There is a need to focus on non-speech segments detection so that most of the irrelevant information can be thrown out in the pre-processing level. Therefore, there is a definite need for putting effort towards exploring different approaches for handling the field deployable systems.

1.4 Organisation of the Thesis

In the light of issues mentioned in this chapter, the this thesis is organized into seven chapters. The content of each chapter is summarized as follows:

- The **Chapter 1** is dedicated on introducing speaker verification, a system used for authenticating a person from the speech utterance. The chapter discusses the modular representation of the work, the motivation behind the implementation and some issues present in the practical deployment. Finally, the chapter ends with organization of the thesis.

- The **Chapter 2** reviews the literature on existing methods related to the thesis work. The review is based on three possible directions for handling the TDSV framework under clean and degraded speech conditions. These directions can be distinguished as use of begin and end point detection, speech enhancement and exploration of alternative/ complementary features for speech recognition approaches. Further, based on the review of different approaches, the organization of the thesis is presented which is suitable for TDSV task.

- The **Chapter 3** aims to use enhanced speech for the detection of end points under degraded conditions. There are different combination of experimental studies conducted for the speech enhancement under degraded conditions. Among these, end points detected from the speech signal enhanced by the combination of temporal and spectral enhancement techniques gives an improvement in the performance of TDSV system.

- The **Chapter 4** investigates the utility of HS of the speech signal, constructed from its AM-FM components or IMFs obtained using a non-linear and non-stationary data analysis technique called MEMD. The HS features extracted from the instantaneous frequencies and energies of the

IMFs obtained from MEMD for improving the performance of TDSV system. These extracted features are tested individually and in conjunction with the MFCCs for TDSV task, using DTW technique.

- The **Chapter 5** attempts to detect begin and end points using some speech specific information for the removal of non-overlapping speech as well as non-speech background degradation. Instead of energy based end point detection. Here, vowel-like regions (VLRs), dominant resonant frequency (DRF), and foreground speech segmentation (FSS), glottal activity detection (GAD), dominant aperiodic regions detection (OBS) and speech duration knowledge (SDK) are used to detect begin and end points accurately.

- The **Chapter 6** combine several approaches presented in the earlier chapters in different stages of the TDSV system. First the speech signal is enhanced using temporal and spectral enhancement techniques. On enhanced speech, robust end point detection algorithm using speech specific knowledge is applied. From the detected speech regions, robust features extracted from HS of the IMFs obtained from MEMD of speech features augmented with MFCCs shows further improvement in the development of TDSV system under low SNR cases.

- Finally, the **Chapter 7** summarizes the works presented in this thesis, highlights the main contributions of the work and gives some future directions made possible by the present work for researchers.

2

End Point Detection, Speech Enhancement and Feature Extraction - A Review

Contents

Objective

This chapter gives a review of different works related to speech processing under different scenarios that are relevant from the perspective of practical deployment of TDSV. There are many methods presented in the literature depending on the approach taken to counter the degradation effect. Three different directions are use of end point detection, exploration of speech enhancement methods and feature extraction required for the final deployment of the TDSV system. Finally, based on various merits and demerits of the different methods in the literature, an alternative direction will be proposed which may give a better system accuracy than the existing methods. The chapter concludes by laying out the framework involving the stated directions into a common platform for having a practical deployment of the TDSV system.

2.1 Introduction

The motivation of the thesis, as discussed in Chapter 1 projects TDSV system from the perspective of the practical field deployable systems. This chapter mainly focuses on reviewing and projecting the scope towards future work direction on deployable TDSV system. In this regard, a brief review is presented by addressing different approaches for TDSV system under degraded speech conditions. Several efforts have been made starting from the development of front-end signal analysis in the pre-processing stage to back-end stage using exploration of different features. This chapter reviews some suitable directions from the perspective of practical deployment of the TDSV based framework. These directions are mainly divided into three different categories based on the issues faced in the development of the system. The first direction is explored for selection of speech frames using EPD methods. The TDSV system performance mainly depends on the accurate detection of end points of the speech utterance, especially under degraded speech conditions. The second issue focuses on the speech enhancement techniques used for removing the degradation effects and improve the quality and intelligibility of the speech. However, when *Signal to Noise Ratio* (SNR) drops to lower than 5 dB or when noise contains complex audio events with high background noise, then some alternative features are used to enhance the system performance further. Finally, the augmentation of MFCC features with different types of alternative noise robust features are explored for the TDSV system to enhance the performance.

2.2 Methods for End Point Detection

This section describes the various methods used for the detection of begin and end points of speech utterances under clean and degraded speech conditions. The detection of begin and end points of a speech utterance is used to remove the silence regions, background noise and speech regions from the raw speech present in the begin and end portions of speech utterance. The accurate detection of end points speeds-up the implementation process while suppressing the unwanted redundancies and transform the speech signal into feature vector space. This plays a major role in practical deployable system where acoustic environment changes continuously.

2.2.1 Energy Based End Point Detection

The main objective of the EPD method is to detect the presence of speech regions embedded in various types of non-speech and background noises. In this section, in the pre-processing stage, the raw speech is first processed through EPD method to detect speech regions from non-speech regions during training and testing phases. Although EPD performs well in clean speech, but difficult task in different degraded conditions. Speech is divided into short segments of 20 ms frame size overlapped with a frame shift of 10 ms using a sliding Hamming windowed technique. Hamming windowed frames are computed for each speech frame using the normalized speech. The decision of speech vs non-speech is obtained in each speech frame. The frames having energy value greater than certain threshold are considered as speech otherwise non-speech frames. This method works well in clean speech conditions. This method fails when the background speech has larger amplitude than speech and under low SNR conditions, and pass through energy threshold. This is obtained through energy threshold where each frame energy compared with predefined threshold.

Similarly, *short-time zero crossing rate*, *fullband and lowband energy*, *multiplication of upper and lower signal envelope*, *periodicity*, *spectral entropy*, *long-term spectral divergence* are explored. New methods are proposed such as perceptual wavelet packet transform and the Teager energy operator. These threshold based methods detects well in clean speech or high SNR conditions. However, under low SNR conditions, detection performance drops drastically. In such cases, these threshold based methods in combination with spectrum-based features such as entropy, cepstral features and linear discriminant analysis to MFCCs are computed. The energy and likelihood ratio based detection methods are obtained to discriminate speech and non-speech by using *Gaussian Mixture Models* (GMMs).

2.2.2 End Point Detection Using Speech-Specific Knowledge

In the pre-processing stage, special care is required to avoid threshold based methods. EPD using speech-specific knowledge is used due to threshold constraints for the detection of end points under degraded conditions. Therefore, *vowel-like regions* (VLRs) knowledge helps in overcoming the drawbacks present in multiple thresholds followed in most of the techniques. In [49], a method using *vowel onset points* (VOPs) is explored to detect begin and end points accurately. This method was found to be robust as compared to the energy based EPD due to high amplitude, longer duration, less zero crossing rate and periodicity of the speech. Different approaches like the difference in the energy of each peaks and their corresponding valleys in the amplitude spectrum [52], the pitch information [53], the wavelet scaling coefficients [54], the *Hilbert envelope* (HE) of the *linear prediction* (LP) residual [55], the spectral peaks, the modulation spectrum energies [56] and the spectral energy present in the glottal closure region [57] are explored.

Several methods such as energy, zero-crossing rate and spectral flatness may fail to detect VOPs and VEPs properly and also increase the number of spurious ones. Hence, the periodicity information obtained using HE of the LP residual which is relatively less affected by various degradations. A *zero frequency filtering* (ZFF) approach is used for detection of epochs [58].The zero frequency resonator exploits signal energy around zero frequency region and attenuates all other information. The strength of excitation obtained from ZFF signal (ZFFS) has been explored to better discriminate the voiced-unvoiced transitions [59]. Both the evidences incorporate the excitation source information. By combining the evidences obtained from the HE of the LP residual and ZFFS to accurately detect VOP and VEP evidences and reduce most of the spurious regions in speech [51].

On top of all these approaches, vowels, semivowels and diphthongs considered as VLRs [24, 51, 57] and rest of the sound units are *non-vowel like regions* (non-VLRs) such as frication, burst, voice bar, nasals are used to further improve the end points detection. VLRs are high energy regions and non-VLRs are relatively low energy regions. These VLRs are detected based on *vowel-like region onset points* (VLROPs) and *vowel-like region end points* (VLREPs). Detection of non-VLRs is around the VLROP and VLREP. Using these two events one can detect begin and end points using glottal cycle associated with VLR and marked them as VLROP and VLREP, respectively. This knowledge enhances the true detection, reduces the missed and spurious detection of VLROPs and VLREPs at the begin and end points depending on the preceding and succeeding sound units.

Similarly, some of the methods are proposed in this direction for detecting end points of the speech utterance using *glottal activity detections* (GADs) obtained by using strength of excitation and periodicity from ZFFS [60]. GADs are detected accurately due to high energy in the voiced region. To improve the detection accuracy, Bessel expansion and AM-FM model is proposed to detect *glottal closure instants* (GCIs) and VOPs of the speech [61]. It is necessary to develop robust method to detect the speech end points accurately under clean, degraded speech (different levels of noise) and practical environmental conditions.

2.3 Different Methods for Speech Enhancement

Another direction is explored in the pre-processing stage before extracting features from the speech utterance. Similar to EPD technique, speech enhancement technique is used when the background noise is overlapping with speech regions. The speech enhancement techniques are used to improve the perceptual aspects of the speech such as the overall quality and intelligibility of speech regions. In practical scenarios, recorded speech is affected by different degradation like background noise, background speech, clipped data, reverberation, lip smacks, chirping noise, heavy breathing, mouth clicks and competing speech from other speakers, resulting in degraded speech. This section explores different techniques that are existing for reducing the effect of background noise.

This method considers the additive nature of background noise. Mathematically, speech degraded by the background noise $y(n)$ expressed as the sum of clean speech $s(n)$ and degraded by background noise $d(n)$ which is additive in nature. That is,

$$y(n) = s(n) + d(n) \tag{2.1}$$

In the frequency domain, it can be represented as

$$Y(k) = S(k) + D(k) \tag{2.2}$$

where, k is the index of frequency bin. $Y(k)$, $S(k)$ and $D(k)$ represents the degraded speech average *discrete Fourier transform* (DFT) magnitude spectrum, desired speech and degraded by background speech. To minimize the error, estimate $\hat{s}(n)$ which approximates the desired speech $s(n)$ represented by

$$e(n) = s(n) - \hat{s}(n) \tag{2.3}$$

where, $e(n)$ represents the error introduced between the desired $s(n)$ and the estimated speech $\hat{s}(n)$.

Speech enhancement gives benefit to enable smooth communication. Majority of these techniques are broadly categorized into spectral and temporal processing speech enhancement techniques.

2.3.1 Spectral Processing Methods

Spectral processing methods constitute most popular techniques for noise reduction. In this, spectral magnitude is estimated from the original speech. The estimated speech subtracted from the degraded speech magnitude spectrum on a frame by frame basis to obtain desired speech. These methods are *spectral subtraction*, *wavelet denoising*, *minimum mean square estimation* (MMSE) and *noise estimation methods* are explored for spectral enhancement.

2.3.1.1 Spectral Subtraction

Spectral subtraction is one of the early methods used for speech enhancement. This is obtained by subtracting the average magnitude spectrum which is subtracted from the overall speech on a frame by frame basis to estimate the enhanced speech spectrum [26]. This magnitude spectrum can be computed during the speech pauses to improve the quality of varying nature of background sources and degraded speech.

From Equation 2.2, the enhanced spectrum can be obtained by using spectral subtraction

$$\left|\hat{S}(k)\right| = |X(k)| - \left|\hat{D}(k)\right| \tag{2.4}$$

where, $|X(k)|$, $\left|\hat{D}(k)\right|$ and $\left|\hat{S}(k)\right|$ are the average magnitude spectrum of degraded speech, estimated speech from background noise and estimated clean speech.

Accurate estimation of desired speech may not always be possible. Overestimation of background noise are obtained due to dynamic nature of the speech utterance. The overestimation produce negative values in the estimated magnitude spectrum $\left|\hat{S}(k)\right|$ which results the *musical noise* [62, 63]. The musical noise in the enhanced speech is more severe than the background noise itself [63–65]. A method is proposed to reduce the musical noise in which the noise spectrum is over estimated and used to avoid the spectral components under certain minimum threshold values [66].

$$\left|\hat{S}(k)\right| = \begin{cases} |X(k)| - \varepsilon\left|\tilde{D}(k)\right|, |X(k)| - \varepsilon\left|\tilde{D}(k)\right| > \xi\left|\tilde{D}(k)\right| \\ \xi\left|\tilde{D}(k)\right|, otherwise \end{cases} \tag{2.5}$$

where, ε is used as an oversubtraction to measure degraded speech in terms of SNRs.

$$\varepsilon = \varepsilon_0 - \frac{3}{20} SNR, -5dB \leq SNR \leq 20dB \tag{2.6}$$

The ε value obtained by computing the musical noise and its signal distortion. In 2.6 ε_0 is the value of ε at SNR 0 dB. The ξ is the spectral components of enhanced speech which fall below a preset value. Higher values are subtracted at low frequencies with low SNR levels and a lesser values are subtracted at high frequencies with high SNR levels.

A multiband frequency based analysis using spectral subtraction method is used [64]. In this oversubtraction computed from the corresponding speech subbands. In some conditions, it is difficult to suppress the noise without introducing residual noise and degrading the quality of the speech. To reduce musical noise *spectral smoothing*, *formant enhancement* and *comb filter* are explored. However, under very low SNRs level masking property is applied to reduce the effect of residual noise.

2.3.1.2 MMSE Estimator

In spectral subtraction, no particular method is explored for the distribution of spectral components. In this the estimation of desired speech from degraded speech is perceptually mapped through *minimum mean square error* (MMSE) estimator. The enhancement is carried out using *short-time spectral amplitude* (STSA) of speech in its perception. The MMSE-STSA can be expressed as,

$$\widehat{M} = E\left\{(S_k - \widehat{S_k})^2\right\} \tag{2.7}$$

where, $\widehat{M}$ is MMSE-STSA estimator, S_k is clean speech and $\widehat{S_k}$ is estimated speech. The MMSE log-spectral amplitude (MMSE-LSA) estimator is an extension of MMSE-STSA, aims to minimize mean square error between the logarithm of STSA of the estimated speech. The MMSE-LSA expressed as,

$$\widehat{L} = E\left\{(logA_k - log\widehat{A_k})^2\right\} \tag{2.8}$$

where, $\widehat{L}$ is MMSE-LSA estimator, $logA_k$ is the logarithmic STSA of speech and $log\widehat{A_k}$ is the STSA of estimated speech. MMSE-LSA showed a better reduction of musical noise than MMSE-STSA method. The improved performance is achieved in the enhanced speech quality of MMSE-STSA as compared to MMSE-LSA. Some of the techniques are used to suppress noisy speech using *Gamma and Laplacian distributions*,*ideal binary masking* and *binary masking functions* methods.

2.3.1.3 Wavelet Denoising

Wavelet transform is a simple denoising method which uses thresholds to define a limit between the estimated speech and noise wavelet coefficients. The long time intervals are obtained at low frequency components and short time intervals at high frequency components and hence it may produce a more flexible time-frequency representation of speech. This technique uses *wavelet shrinkage* algorithm based on the thresholding between wavelet coefficients of the estimated and degraded speech [67]. However, thresholding may not always provide good results. Due to this, the perceptual quality is greatly affected [68]. The *wiener filtering* in wavelet domain and *wavelet* filterbank have been used [68–70] for degraded speech using wavelet denoising.

2.3.1.4 Noise Estimation Methods

The spectral processing approach depends on the stability and accuracy of the estimated magnitude spectrum. This method estimate the *a priori* and *a posteriori* SNR from MMSE algorithm. This method is also used as a *voice activity detection* (VADs) for achieving better enhancement without degrading speech quality [71]. The moving average STSA is obtained on a longer decay than the difference in speech. The noise and speech components are divided into subspaces in terms of mutually orthogonal low-rank linear model. The desired speech can be obtained by removing the noise and reconstructing the speech back into the subspace. The decomposition of degraded speech can be obtained by using *singular value decomposition* (SVD) [72] or *Karhuen-Loeve Transform* (KLT) [73].

2.3.2 Temporal Processing Methods

The basic principle of temporal processing is to identify the high SNR regions and derive a weight function that can emphasize the high SNR regions relative to low SNR regions. The degraded speech excitation source is multiplied with weight function and obtained the high SNR speech regions. The modified speech is used to excite the time-varying all-pole filter to generate the enhanced speech. This does not produce any type of musical noise type distortion.

2.3.2.1 LP Residual Enhancement

Some techniques interrupt the spectral strength of degraded speech, resulting in unpleasant distortions in the enhanced speech. Yegnanarayana *et al.* proposed a method for enhancement of noisy speech by exploiting the characteristics of excitation source information [74]. The principal approach for speech enhancement is to identify the high SNR regions in the noisy speech in the temporal and spectral domains. Then without affecting the distortion in the enhanced speech, the speech regions are detected by using high SNR speech regions relative to the low SNR speech regions. The temporal speech enhancement method is broadly categorized into three steps: 1) Identification and enhancement of high SNR regions of speech in the temporal domain at the gross level, 2) Identification and enhancement of high and low SNR regions in the noisy speech at the fine level and 3) Enhancement of spectral peaks over spectral valleys at the short-time spectrum level.

The speech regions corresponding to low and high SNR regions are identified from the characteristics of LP residual at the gross level. A weighting function for the residual signals is obtained based on the smoothed inverse flatness attributes degraded speech. At the fine level, the residual signal in voiced segments are given lower weight in low SNR compared to higher SNR segments. A weight function is derived at the fine level by considering the residual energy plot to deemphasize the short speech segments corresponding to the valleys relative to the speech segments corresponding to the peaks. However, under degraded conditions, the derivation of the weight function may not be reliable using energy of the short-time speech segments of residual signal. Hence, the Toeplitz prediction matrix is used to represent the short-time energy of the corresponding speech frame of the LP residual signal. In [74] *Hilbert envelope* (HE) of LP residual reconstruction is utilized. The instant of strong excitation at the HE has large amplitude and makes it a good indicator of *glottal closure instants* (GCIs). Therefore, the weighting function is obtained by using HE of LP residual which has effect on emphasizing the voiced speech. This leads to an enhanced LP residual signal. Similarly, a multi channel speech enhancement method using GCIs are more effective. These can be achieved by highlighting the excitation source information around the GCI events [75].

2.3.2.2 Signal Subspace Approach

The main principle of this approach is that each vector of noisy speech is composed of clean and noisy subspace. The noise subspace contains signal from noisy speech. The speech enhancement is achieved by removing the noise subspace and estimating the enhanced from the signal subspace. The decomposition of noisy speech into clean speech subspace can be done using either the SVD [72] or KLT [73, 76, 77]. The SVD decomposes the Eigen vectors and their corresponding Eigen values of the speech. The enhanced speech is reconstructed from the dominant Eigen values along with their corresponding Eigen vectors by neglecting small Eigen values. The smaller eigen values typically carry only noise information that can be discarded for further processing. Jensen *et al.* [78] proposed reduction of noise in speech on signal subspaces using the *quotient* SVD (QSVD). There are two signal estimators: 1) the *spectral domain constraint* (SDC) and 2) the *time domain constraint* (TDC). The decomposition of noisy signal to clean signal and noise subspace can be obtained by using KLT to the noisy signal. The KLT was modified by a gain function determined by the estimator while the remaining noise subspace was completely removed. The enhanced speech is obtained from the inverse KLT of the modified components.

Similarly, the masking subspace method is incorporated to produce better performance over conventional subspace methods [79]. In addition to all subspace based methods, there are several multi-microphone algorithms recommended for noisy speech enhancement. These methods can exploit the spatial information in the microphone signals when the speech and noise sources are detected under degraded speech regions. Hence, they are able to perform both spectral and spatial filtering on the degraded speech. The microphone based methods provide improved results as compared to single microphone speech enhancement based methods.

2.4 Different Features for Speaker Modeling

2.4.1 Spectral Features

From the speech perception point of view, the most widely used MFCCs are obtained by computing the log magnitude spectrum of the speech, while neglecting its phase spectrum. Basically, the log magnitude spectrum is passed through a non-linear Mel filterbank and finally DCT is applied on the Mel filtered coefficients to get the compact information. However, the phase spectrum of speech is equally powerful as the magnitude spectrum, and has found very important use in several speech processing applications [2]. The MFCC features are based on vocal tract system that does not always provide the necessary information. The Mel filter banks tuned for different applications might be more suitable to perform better SV system but not in all the conditions. To overcome this shortcoming, different features constituted in terms of their frequencies, amplitudes, phases and some complementary information are required to capture speaker-specific characteristics from the given speech utterance.

Several spectral features are available in the literature other than conventional MFCCs that capture speaker-specific characteristics. The complementary/ alternative features are *linear prediction cepstral coefficients* (LPCCs) which are well known and produces the cepstrum as a sequence of numbers that characterize a speech frame. Some of the popularly used feature representations of the LP coefficients are: LPCCs, line spectral frequencies, partial correlation coefficients (PARCORs), log area ratios (LARs) and perceptual linear prediction coefficients (PLPCs), formant frequencies and bandwidths [30, 31]. From the literature, the experimental studies show that on fusion of these developed systems by utilizing different spectral features along with conventional MFCCs helps to improve the overall performance of the system [80, 81]. Specifically, the experimental studies demonstrate that fusion of different spectral features augmented with MFCCs at the score level fusion significantly enhances the SV system performance.

2.4.2 Voice Source Features

The voice source features are unique for every speaker which encapsulate the glottal signal characteristics. Excitation source information is captured by implicit processing of LP residual in different ways to extract the speaker excitation source characteristics. The glottal flow vocal tract features are cepstral coefficients [82], LP residual phase [83] and wavelet analysis based features [84] are widely used in speech processing applications. The autoregressive model of the vocal tract is estimated and used to built *vocal tract mel-cepstrum coefficients* (VTCCs) to obtain voice source features. The VTCCs are subtracted from the MFCCs of speech frame. Similarly, the accurate detection of closed phase is a difficult task under degraded speech conditions. To overcome this problem, some of the methods are proposed in the same direction for source features such as *residual mel frequency cepstral coefficients* (RMFCCs), *mel power difference of spectrum in sub bands* (MPDSS) [85, 86]. Similarly, some recent exploration has been done by using the *integrated LP residual* (ILPR) which closely approximates the *glottal flow derivative* (GFD) from the speech signal. The obtained signal is processed in a pitch synchronous manner on which the DCT is applied and these obtained features are termed as DCT of ILPR (DCTILPR). The DCTILPR features showed improved performance in *speaker identification* (SI) task. The DCTILPR features represents its capability to carry definite speaker specific characteristics. Therefore, DCTILPR features have been found to be useful as a speaker-specific feature [87] and showed that MFCCs and DCTILPR techniques at score level combination is found to give a significant improvement over the MFCCs system performance alone [88].

2.4.3 AM-FM Analysis Based Features

From the *Amplitude Modulation - Frequency Modulation* (AM-FM) analysis of speech signal can be used to capture speaker specific information. The AM-FM analysis of speech signal aims to represent the sum of a finite number of narrowband signals with slowly varying degrees of amplitude and frequency modulation. The framework of *Multiband Demodulation Analysis* (MDA) was explored in this direction. The speech signal is passed through a fixed parallel filterbank of overlapping bandpass filters and generating different time-domain signals for feature extraction. These features are represented in terms of instantaneous amplitudes and frequencies which is estimated from the Hilbert Transform or *Teager Energy Operator* (TEO). The modern AM-FM based analysis methods use *empirical mode decomposition* (EMD) for processing speech signals. There are many modifications have been proposed to the EMD. However, the best performance have come by the motivation of noise to the speech signal. This was observed by combining speech with finite amplitude before providing it to the EMD process. The mode-mixing could be reduced significantly in EMD such a method is called as *Ensemble Empirical Mode Decomposition* (EEMD) [89]. Several methods are proposed such as Complementary Ensemble Empirical Mode Decomposition (CEEMD) [90] and the Improved Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (ICEEMDAN) was developed, which makes some precise and effective modifications to the CEEMDAN method [91]. These features have complementary speaker information in the vocal tract or voice source based features and their fusion at the feature or score level that helps in improving the system performance [92, 93].

2.5 Summary and Scope for Present Work

In **Chapter 2**, a brief review is presented addressing different approaches for implementing the field deployable systems under practical scenarios. The literature shows that the accuracy of the TDSV system falls significantly in the presence of varied noise conditions. For that several efforts have been made in the preprocessing stage. Among them, energy (Threshold) based EPD method is mostly used to detect begin and end points of the speech utterance. However, they are not suitable under degraded conditions. The long-term temporal information is used at different degrees of variability of speech and noise for the detection of begin and end points accurately.

The robust EPD is sufficient to separate speech and non-speech regions when speech is temporally non-overlapping. However, in some cases background noise temporally overlap with the speech regions and energy based EPD method alone may not be sufficient. Thus, it is essential to review speech enhancement as a principal direction in tandem with robust EPD method. The combination of speech enhancement and robust EPD method can be an effective solution in practical scenarios. In degraded speech conditions, by using the combination of speech enhancement and robust EPD method, the system performance may improve, but under clean and less affected speech conditions the performance may degrade. Therefore, many explorations are done for finding alternative features based on non-stationary data analysis have been explored for different speech conditions. Finally, organization of the works that are endeavoured in this thesis to comprehend the proposed framework will be discussed.

The **first objective** of this thesis, which is addressed in the **Chapter 3** explores the use of speech enhancement techniques for enhancing degraded speech. These enhanced speech regions are then passed through end point detection for the system development. This chapter investigates the ability of temporal and spectral enhancement performed on the varied degraded speech so that the enhanced speech assumed to be closer to clean speech. Using DTW technique, the combination of temporal and spectral enhancement along with energy based EPD methods contribute to the task by eliminating the degradation and obtained improved system accuracy for the TDSV system.

The **second objective** of this thesis revolves around the idea that the interfering noise sources do not affect the speech signal equally in varied data conditions. A few attempts have been made to develop effective system by neglecting more corrupted speech frames. These methods may not reduce environmental variability due to sound units during train and test phases. Therefore, without discarding speaker specific frames, the degradation effect and variability in sound units can be improved by capturing maximum speaker specific information of the varying nature of the speech. **Chapter 4** of this thesis is dedicated to show the adaptive filterbank nature, manifested in the *Hilbert Spectrum* (HS) constructed from its AM-FM components or *Intrinsic Mode Functions* (IMFs) obtained from *Modified Empirical Mode Decomposition* (MEMD) features which can complement the Mel filterbank for characterizing speakers. The features extracted from HS are tested individually and in augmentation with the MFCCs and also with *cepstral or energy-like* features for the TDSV task using DTW technique.

The **third objective** of this thesis is to demonstrate the utility of the robust end point detection, derived from the speech specific knowledge for the task of TDSV system. The energy based EPD method perform well for clean speech and fails significantly under degraded conditions as discussed in the **Chapter 3** and **Chapter 4**. Due to high energy in the background noise leads to incorrect detection. Therefore, a better system is developed using the speech-specific knowledge of *vowel-like regions* (VLRs) to detect accurate begin and end points of the speech utterance. With this viewpoint **Chapter 5** proposes a method using VLR *onset points* (VLROP) and VLR *end points* (VLREP) detection method obtained from both excitation source and vocal tract system information. To eliminate spurious VLRs in background noise *dominant resonant frequency* (DRF) and to remove spurious VLRs in the background speech region *foreground speech segmentation* (FSS) is used. To detect the sonorant consonants and missed VLRs, *glottal activity detection* (GAD) is applied. To include an unvoiced consonant, obstruent region detection is done at the begin of the first VLR and end of the last VLR. Finally, the combination of these obtained features improve the performance of the TDSV system consistently.

The **fourth objective** of this thesis is to demonstrate different types of speaker specific information at different stages for handling the field deployable systems. With this viewpoint, **Chapter 6** is combines different approaches explored in earlier chapters into a combined framework. The insights gained in **Chapter 5** by using speech specific knowledge for robust end point detection is used for noise compensation at signal level. The insights gained in **Chapter 4** at the feature level compensation by using robust features extracted from HS of the IMFs obtained from MEMD of the speech. The insights gained in **Chapter 3** used the combined temporal and spectral enhancement technique prior to the end point detection for enhancing speech regions embedded in noise. **Chapter 6** shows further improvement in the low SNR cases when the use of speech enhancement prior to signal and feature level compensation is used. Hence, these three directions are combined together. Finally, a common framework is developed for practical deployment of the system.

The final chapter, **Chapter 7** of this thesis, summarizes the prime contributions of this thesis towards development of a TDSV system under practical scenarios and outline future research directions of present work.

3

Processing Degraded Speech for Text-Dependent Speaker Verification

Contents

Objective

This work explores the use of speech enhancement for enhancing degraded speech which may be useful for TDSV system. The degradation may be due to noise or background speech. The TDSV system is based on the DTW method. Hence there is a necessity of the EPD. The EPD can be performed easily if the speech is clean. However the presence of degradation tends to give errors in the estimation of the end points and this error propagates into the overall accuracy of the SV system. Temporal and spectral enhancement is performed on the degraded speech to make the nature of enhanced speech closer to clean speech. Results show that the temporal and spectral processing methods do contribute to the task by eliminating the degradation and improved accuracy is obtained for the TDSV system using DTW technique.

3.1 Introduction

S*peaker recognition* (SR) is the task of recognizing speakers using their speech signal [94]. Depending on the mode of operation, SR can be either SI or SV. In case of SI, the most likely speaker of the test speech signal is identified by comparing among the enrolled speakers. SV involves identity claim and test speech signal. The objective of the SV is therefore to validate the identity claim. Accordingly, SV is one to one matching whereas SI is one to many matching task. Depending on the constraint on the lexicon used during enrollment and testing, SV is also classified into TISV and TDSV modes [6]. In case of TISV mode, there is no restriction on the lexicon used for enrollment and testing. Alternatively in case of TDSV mode, the lexicon of the test speech signal is a subset of the lexicon used during enrollment [16].

From the practical usability point of view, the user is free to provide test speech with no constraints on duration, quality, recording condition, channel and lexical content. Accordingly, the performance of the SV system that is influenced by many of these possible variabilities. Among these variabilities, lexical content and channel variations are the most detrimental. Compared to channel variability which is due to uncontrolled environmental factors, lexical variability can be manageable. The performance of TDSV system is expected to be better compared to that of the TISV case. Also matching of short duration phrases during training and testing with high accuracy makes it an attractive option for commercial speech based person authentication system. [38]. Speech is commonly used in biometric security technologies because of the user acceptance, ease of use, low cost, high accuracy and ease

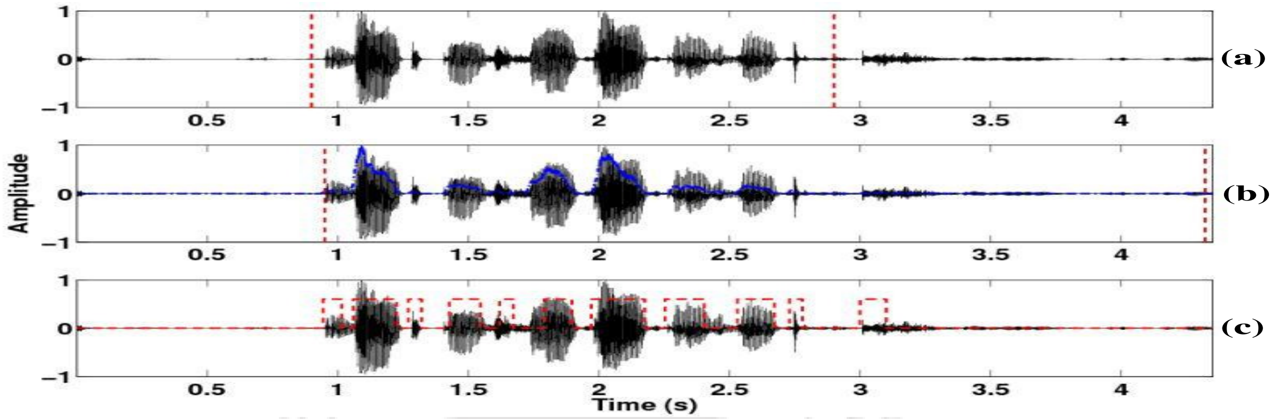


Figure 3.1: Illustration of end point detection on degraded speech (a) Degraded Speech Signal with the ground truth marked (b) Energy Based Method (c) VOP based method

of implementation. In the present day situation the most common natural biometric information source is telephone based remote access control applications [95–99]. In real time applications, the performance of the system may degrade due to various reasons such as background noise, sensor mismatches, reverberation, channel mismatch, background speech and many more. The speech is also affected by illness, amount of stress, aging and health conditions. Depending on the above mentioned conditions, the accuracy of the TDSV system tends to degrade. In order to overcome all the above mentioned conditions in TDSV system, detection of accurate begin and end points is necessary [49,50]. There are algorithms present which are based on the signal energy and have been proposed in [100,101]. Recently, a method using VOP was proposed to detect begin and end points accurately [49,50]. Energy based algorithms and the VOP based method perform well in clean speech conditions. However as seen in Figure 3.1(b), the energy based end point detection algorithm fails for the case when there is background noise and background speech. The VOP based method also fails in the presence of background noise and background speech case as shown in Figure 3.1(c). The background noise is present in the region around 3.2 s and the background speech is present in the region around 3.5 s in Figure 3.1.

Even though TISV system is widely explored, the goal of TISV system is different. TISV system involves general modeling of speaker and essentially captures the speaker-specific characteristics. However, in TDSV system there is no modeling of speaker characteristics. It basically involves matching of two templates and TDSV system is mostly used for the cooperating scenario. This can find applications in biometric authentication system which is why the TDSV system has been chosen over

TISV system. In addition to this, depending on the acoustic background and the presence of different kinds of degradation, poor SV performance may be encountered. A practical case of degraded speech is taken in this work and speech enhancement is applied on the degraded speech so as to improve the SV performance.

In this work, the degraded speech condition is assumed to have a foreground/ background model [102]. The reason is that for most cases of the speech taken for TDSV system, the speaker generally speaks close to the microphone. As a result, the speech which is recorded close to the microphone will be termed as foreground speech and the remaining other interfering signals captured by the sensor which are far from the microphone will be termed as background noise. Note that the background noise can be any kind of degradation which even includes background speech. It has been shown that the foreground speech enhancement performs well in [102] for any kind of degradation present in the background as long as the interfering noises are far away from the microphone. In this work the foreground speech enhancement will be used for TDSV. The exploration of the enhancement strategy will be of two types. The first type is after doing the energy based EPD, there may still be degradation present between the end points and hence foreground speech enhancement will be performed so as to obtain a good accuracy of the system. The second type consists of directly using foreground speech enhancement on the signal and the energy based EPD algorithm is applied on the enhanced speech which may improve the accuracy of the TDSV system. In this context the performance of the TDSV under degraded conditions are explored using temporal and spectral enhancement methods [28].

Temporal enhancement involves modifying the degraded LP residual by emphasizing the high SNR regions resulting in ideally a clean residual. The modification is based on deriving a total weight function which is used to weight the residual. This total weight function is based on the combination of the weight functions obtained from voiced/ unvoiced and high SNR region detection. This residual is then used to excite a time-varying vocal tract system which are the LP coefficients resulting in a temporally enhanced speech. Temporal enhancement focuses on the high SNR regions without considering the type of noise present [28].

Another kind of enhancement which has been widely explored is the spectral subtraction [26]. In this background noise is assumed to be stationary and uncorrelated. As a result the background noise can be modeled by calculating the average magnitude spectrum which is subtracted from the overall signal in the frame based approach to estimate the desired speech from the signal. The average magnitude spectrum is computed mostly during the speech pauses to obtain better noise modeling.

Detecting the end points after enhancement may be a better option since the goal of enhancement methods is to obtain an enhanced speech which ideally is free from the effect of noise or in other words it has the characteristics similar to the clean speech. This work combines the enhancement strategies along with the EPD algorithm so that improvements in the system can be obtained.

Even though the tools and techniques for TDSV system as well as enhancement methods proposed in this work are from existing work, however, combining the enhancement tools as well as the TDSV system gives a solution for the TDSV system which is robust for practical variabilities. These variabilities also involve degradation. Hence, combining speech enhancement with SV helps to overcome some of the issues encountered for practical SV like degradation.

The rest of the work is organized as follows. Section 3.2 describes the enhancement strategies. The TDSV system along with the end point detection is described in section 3.3. The experimental evaluation is given in section 3.4. Finally the Conclusion is given in section 3.5

3.2 Speech Enhancement

The degraded speech is passed through a certain level of enhancement. The enhancement strategies followed in this work involve the one which does not assume the noise characteristics which is the temporal enhancement and the other one which assumes the noise characteristics are stationary which is the spectral enhancement. Accordingly the details of the two types of enhancement are given below.

3.2.1 Temporal Enhancement

Temporal enhancement can be divided into two steps. One is the gross level processing which is similar to the VAD techniques. The other step is deriving the fine level processing which basically emphasizes the high SNR regions. The gross level processing involves deriving different speech production features. The details of this process can be found in [28]. An example plot of various speech production features and their role in the gross level processing can be seen in Figure 3.2.

The fine level processing involves emphasizing the high SNR regions and these regions mostly correspond to the location of the epochs. Hence a robust epoch extraction method needs to be utilized here and the epoch extraction method explored in this work is based on the ZFFS. This method has been shown to be robust to different noise types [58]. The method for obtaining the ZFFS can be found in detail in [58]. Based on the gross weight function and the fine weight function derived

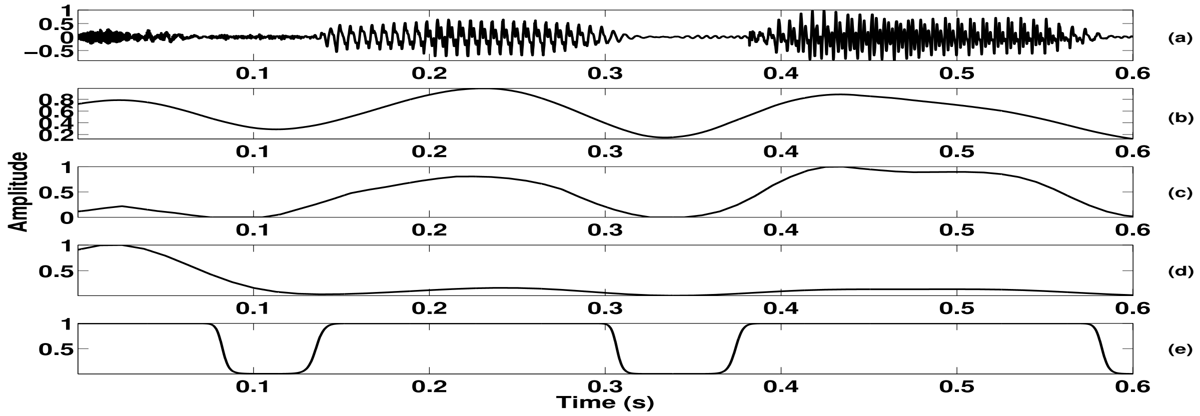


Figure 3.2: Illustration of Gross Weight Function Derivation (a) Speech degraded with background noise (b) HE of LP Residual (c) Sum of ten largest peaks of DFT spectrum (d) Modulation Spectrum Energy (e) Gross Weight Function

from the gross and fine level processing, a total weight function needs to be derived and this total weight function can be used to weight the LP residual for the temporal enhancement. An example of obtaining the total weight function can be seen in Figure 3.3.

The total weight function is multiplied by the LP residual signal shown in Figure 3.3(e) to obtain the weighted LP residual shown in Figure 3.3 (f). The temporally enhanced speech signal can be obtained by synthesizing as follows:

$$S_t(z) = \frac{R_w(z)}{1 + \sum_{k=1}^{p} a_k z^{-k}} \tag{3.1}$$

where, $S_t(z)$ is the temporally enhanced speech and $R_w(z)$ is the weighted LP residual and a_k are the LP filter Coefficients. The plot of the speech which has been temporally enhanced is shown in Figure 3.3 (g).

3.2.2 Spectral Enhancement

The conventional spectral processing methods have been explored here in which the short-term magnitude of the degraded speech are estimated. A spectral gain function corresponding to the MMSE-LSA estimator [103], is applied to the magnitude spectra of the degraded speech, to obtain the enhanced speech spectra. The spectral gain function of this estimator is given by [103]

$$H(k) = \frac{\zeta_k}{1 + \zeta_k} exp(\frac{1}{2} \int_{v_k}^{\infty} \frac{e^{-x}}{x} dx) \tag{3.2}$$

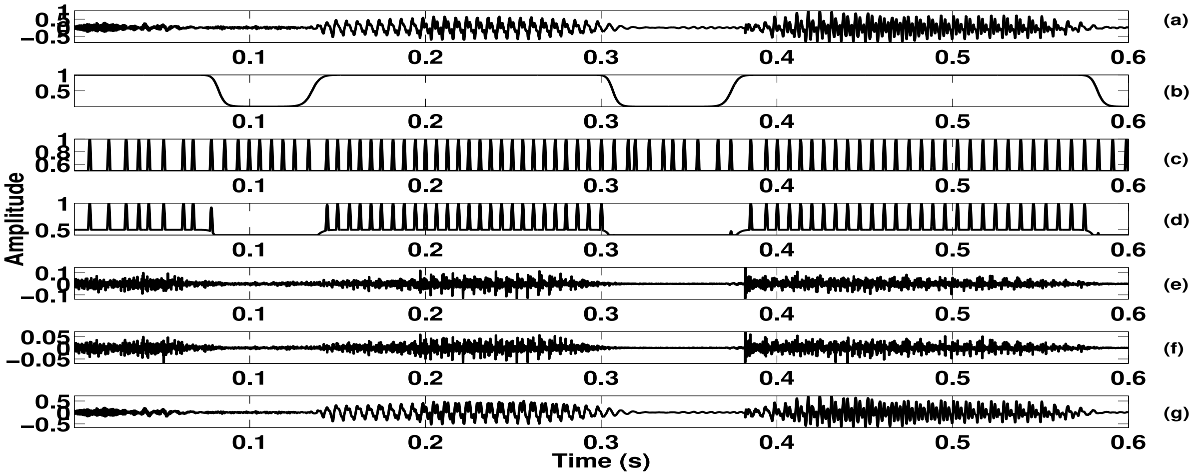


Figure 3.3: Illustration of Fine Weight Function Derivation (a) Speech degraded with background noise (b) Gross Weight Function (c) Epoch locations obtained from ZFF (d) Overall Weight Function (e) LP Residual from degraded Speech (f) Weighted LP Residual (g) Temporally Enhanced Speech

where,

$$v_k = \frac{\zeta_k}{1+\zeta_k}\gamma_k$$

ζ_k and γ_k are a priori SNR and a posteriori SNR,respectively.

The enhanced magnitude and degraded speech phase spectra are then combined to produce an estimate of clean speech and the overlap-add method is used for the re-synthesis in time domain.

3.3 Development of Text-Dependent Speaker Verification System

3.3.1 Energy Based End Point Detection

Robust EPD method is a crucial task for achieving good performance in SV systems since the error introduced in the EPD gets propagated into the overall performance of the TDSV system. Detecting the begin and end points of the speech signal under the clean case is comparatively easier when compared to the degraded condition cases. In most of the real time applications, the speech is mostly degraded since it may be recorded in different conditions. Hence there is a demand for EPD algorithms for the degraded conditions. The other advantages of EPD is that there will be a reduction in the computational cost and response time of the overall system. This is because only the useful speech frames can be passed to the system to do further processing. The most popular EPD is based on using the energy as a feature.

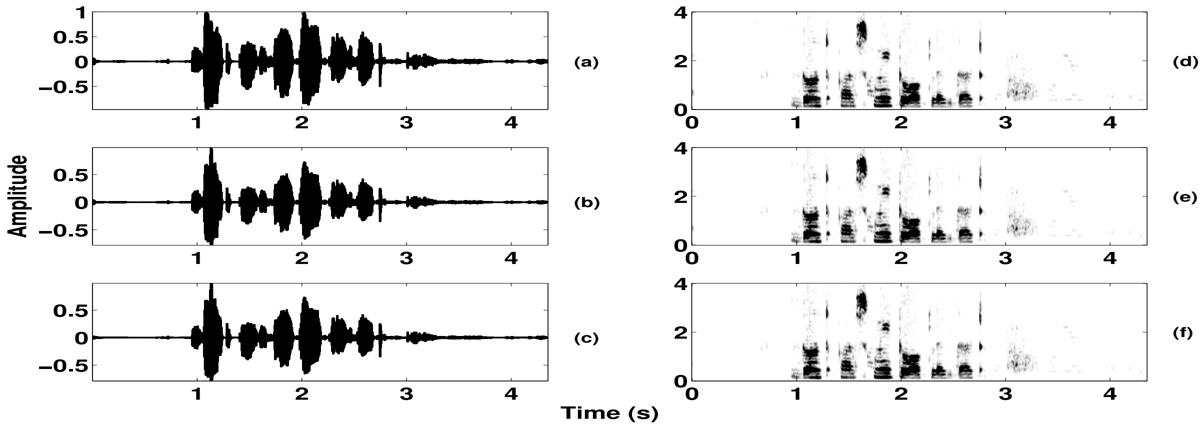


Figure 3.4: Illustration of different stages of Enhancement (a) Speech degraded with with background noise (b) Temporally Enhanced Speech (c) Spectrally Enhanced Speech (d) Spectrogram of (a) (e) Spectrogram of (b) (f) Spectrogram of (c)

The implementation of the energy based EPD method is to calculate the energy of all frames present in the speech signal and the average energy of total frames is computed. Next the energy of each frame is compared with the threshold taken as 6 percent of average energy. The energy threshold is decided based on several SV experiments (offline). Different thresholds are considered and the one which is equal to 6 percent of average energy gave the best performance.

The comparison starts with the starting frame from the left side of speech signal. The energy of the first three consecutive frames are compared to see if they are higher than the threshold which is equal to 6 percent of average energy. If this is true then the first frame is considered as the begin point. Similarly, starting with the last frame from the right side of the speech signal the energy of the three consecutive frames are compared with the 6 percent of average energy threshold. If this condition is true, the current frame is considered as the end point. If the condition is false, then the next frame is considered. The process is continued until the condition is satisfied and the end points are detected. Finally, the speech signal considered between these start and end points is used for further processing.

3.3.2 Dynamic Time Warping Based TDSV System

The details of DTW based TDSV system are described in the next section, where initially the feature extraction will be explained followed by the template matching.

3.3.3 Feature Extraction

The speech signal is processed with a frame size of 20 ms and a frame shift of 10 ms. Each 20 ms of the signal is Hamming windowed and MFCCs are computed using the 22 spaced logarithmic filters. The first 13 coefficients are considered as a feature vectors. The Delta (Δ) and delta-delta ($\Delta\Delta$) features are computed using two preceding and two succeeding feature vectors from the current feature vector [37]. The final feature vector dimension is 39, which includes the first 13 coefficients, along with the first and second order derivatives of the 13 dimensional coefficients. These feature vectors are normalized to have a zero mean and unit variance distribution using cepstral mean substraction (CMS) followed by cepstral variance normalization (CVN) [51]. The extracted features are used for further SV analysis purpose.

3.3.4 *i*-vector Based System

The motivation behind this work is to implement and investigate SV system by implicit modeling of speaker-specific information in the *i*-vectors. For performance comparison, first a baseline *i*-vector based SV system is developed using energy based EPD [14,38].The *i*-vector implementation, uses *total variability matrix* (T-matrix) modeling as introduced in [12]. The dimensionality of the GMM mean supervectors for a test speech is reduced by projecting it to a low rank subspace. The GMM mean supervector for a test speech supervectors are obtained by concatenating the mean vectors of the each user's model. The reduced dimension representation is called the *i*-vector. The implementation of the *i*-vector system described detailed in [14,38]. The GMM mean supervector can be represented as,

$$M_s = m + Tw_s \tag{3.3}$$

where, m is the user and channel independent supervector and M_s is the adapted GMM mean supervector and w_s is the linearly related to *i*-vector of the given test speech. The SV done by comparing the *i*-vectors corresponding to test speech and claimed user's training utterance using the cosine kernel score between these two *i*-vectors. To further reduce the dimensionality of the *i*-vectors LDA and WCCN is applied.

3.3.5 Template Matching

The system uses MFCCs and DTW based template matching for SV of a claimed identity. The main objective of the DTW was to exploit the linear time normalization alignment by implicitly assuming that the speaking rate variation is proportional to the time, speed and duration of the speech utterance. It is also independent of the speech pattern and mostly involves a sequence of short-time acoustic representations being spoken by the claimed identity. It is believed that the timing difference between train utterance and test speech utterances are minimized by using the warping in the time axis of one so that the best alignment is obtained with the test utterance. In other words, the test utterances are sometimes stretched and compressed so as to find the proper alignment that results in best possible match between the train utterance and test utterance on frame by frame basis. The distance between the MFCCs of the train and test utterances can be calculated by using the minimum Euclidean distance [104–106].

In time series analysis, these systems are usually based on template based matching in which the time axis of test template and reference models of registered speakers are initially aligned. Next the distances of aligned frames between the train and test templates are computed. Finally the distances computed between the aligned frames are accumulated from the beginning of utterance to the end of utterance [49, 50, 94]. TDSV systems performs well when the linguistic content of the utterance is known and also when the characteristics of speaker-specific information is contained in speech. From practical applications point of view, the DTW system should be designed in such a way that it should be language, sensor dependent and session independent.

The test feature vectors are compared with reference template of the claimed model using DTW. The DTW technique [105] calculates the accumulated distance score between reference model train template X and test template Y of different lengths by considering a warping path. The equation representing the DTW method is as follows,

$$DTW_{\phi}(X,Y) = \sum_{p=1}^{N} DTW(\phi_X(p), \phi_Y(p))m(p)/M_{\phi} \tag{3.4}$$

where, the $DTW_{\phi}(X,Y)$ is total accumulated cost distance of p, $DTW(\phi_X(p),\phi_Y(p))$ is the shortest time spectral distortion, $m(p)$ is a non-negative weighting co-efficient of the warping path and M_{ϕ} is the warping path normalizing factor. The claimed speaker matched to the claimed reference model using DTW to give the shortest distance score. The cohort speakers are selected randomly for each user from the enrolled users excluding the corresponding user and kept fixed for decision logic.The decision is based on the obtained DTW score with respect to a set of four cohort scores by comparing the distance score against the claimed model. This claimed user score compared to the scores obtained

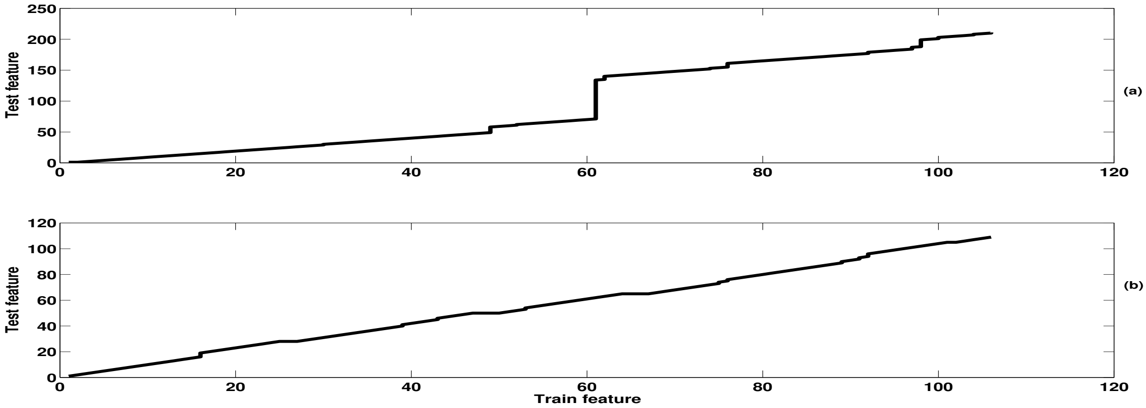


Figure 3.5: Evidences for degraded speech utterance for *Don't ask me to walk like that* (a) speech utterance showing the non-overlapping background noise and the warping path which is deviated away from the regression line before applying the enhancement and end point detection (b) showing the warping path nearer to the regression line after applying the enhancement followed by end point detection and follows the DTW constraints.

from the set of four cohort speakers of claimed user and the decision of claimed user is to accept/ reject a claim.

The accuracy of the system depends on the statistical match between the speaker model and test features [107]. DTW is applied on the features derived from the speech signal collected using the voice server. The performance of the system after applying the proposed EPD on this speech signal is evaluated. It can be observed in Fig. 3.5 (a) and (b) that the DTW warping path for an utterance before and after using the proposed EPD is different. The main observations are that without using the EPD, the warping path deviates from the regression line and using the EPD actually helps in bringing the DTW warping path closer to the regression line. From the practical usability point of view, the users are allowed to give test speech freely with no constraints on duration, recording condition, channel, lexical content and quality of speech. Compared to other variabilities, lexical variability can be manageable in the uncontrolled environmental conditions. The performance of TDSV systems are influenced by many of these other possible variabilities. Using the DTW, matching of short duration phrases in the training and testing utterances with high accuracy makes it an attractive option for speech based multilevel person authentication systems. The accuracy of a SV system depends on the matching between the speaker model and test features [50, 94]. TDSV system is developed using 39 dimensional MFCCs and DTW template matching technique [108, 109].

3.4 Experimental Results and Analysis

Two databases, the RSR2015 [14], and the IITG database [38, 46], which represents speakers from two different countries- Singapore and India. The RSR2015 database contains audio recordings, tablets and recorded on mobile phones, in a closed office environment, with age ranging from 17 to 42. The RSR2015 database is one of the most publicly available database, especially designed for the TDSV system based studies under different duration and lexical constraints, which contains 300 speakers (143 female and 157 male), each of the speakers have 9 sessions of each phrase, out of 9 sessions, three sessions are used for training, three sessions are used for calculation of speaker model and remaining three session for testing. Part I of the RSR2015 database is considered for the evaluation.

The RSR2015 database is collected in a closed room considered as clean database. In order to evaluate the performance under degraded conditions, the test speech files of RSR2015 database are corrupted with Babble noise from the NOISEX-92 database [110]. The energy level of noise is scaled such that the overall SNR of noise added to test speech signal is maintained at 20,15,10,5 and 0 dB, respectively. The performance of the RSR2015 database under clean and different noise degraded test speech conditions are evaluated using the i-vector and DTW algorithm.

The IITG database is collected on the IVR system and the user call flow through ISDN-PRI line which can handle telephone channel calls through CTI card. The database collected from the pre-defined mobile set calls of the students over the telephone channel. The IVR is facilitated on a voice-server which runs the Asterisk software, used for private branch exchange (PBX). PBX can be used to allow the server handling incoming and outgoing calls to and from PSTN or voice over internet protocol (VoIP) services. This call flow is made for execution of the database collected from the students of EEE department at IITG institute level. The users have to give their attendance by making a call from few pre-defined mobile handsets. In the initial stage an IVR system guides an user in enrollment and an enrolled user for testing. On the regular basis, the registered students are allowed to call to the toll-free number and mark their attendance in a practical setting. The overview of the database collection process over the telephone network which is developed for speech biometric based attendance system described in [38].

The RSR2015 database mainly focuses on TDSV task where the users pronounce same pass-phrases for verifying the claimed mode. To evaluate the performance of TDSV task, three sentences have been selected from the RSR2015 database among the 30 sentences selected from the TIMIT database, used for evaluation. The average duration per sentence varies from 2.73 to 3.65 s. These sentences are selected to evaluate the impact of different lexical content with a short duration by considering overall speakers and sessions. They are

"Only lawyers love millionaires"-**(TD-1)**,

"I know I did not meet her early enough" -**(TD-2)** and

"The Birthday party has cupcake and ice-cream"-**(TD-3)**.

In the RSR2015 database, for each speaker, 9 utterances for every sentence, out of 9 utterances 3 are used for training and remaining 6 for testing. The IITG database is collected as a part of the development of a student attendance system in the EEE department of IITG, Guwahati, India. The 325 students (276 male and 49 female), represent an adult age group between 22-30 years. In the training, 3 sentences (sessions) each for 3 speech sentences, are recorded for each speaker in clean environment and controlled conditions during the beginning of semester. In the testing phase, the speakers are allowed to move freely within, in and out of an open hall where the pre-defined mobile handset were kept. There are multiple users made simultaneous calls to the voice-server to mark their attendance. The database collected in the real and practical environment. The average duration of each sentence of the IITG database vary between 3 to 5 seconds over all the users and sessions. The three sentences are

"Don't ask me to walk like that"-**(TD-1),**

"Lovely picture can only be drawn"-**(TD-2),** and

"Get into the hole of tunnels"-**(TD-3).**

For the IITG database, the 3 sentences recorded in the enrollment phase are used for training. The number of sentences for a particular sentence, claimed by a particular user for testing, vary randomly. 30 sentences are available for every sentence on an average for every user, for testing.

The test speech data is collected in an uncontrolled environment at every other time. Due to the various factors like background noise, background speech, clipped data, sensor mismatches, reverberation, real environmental noise, blurred noise and some unwanted signals, degradation may be present in the test speech signal [38, 46]. A preliminary study is performed by using clean speech for training as well as for testing without the presence of background noise, background speech and any other real time environmental noise. Next the testings are made on a regular basis by the users which give the voice biometric based attendance system. As a result there are no constraints in the environment. This resulted in degradation of the speech signal due to various environmental conditions and poor performance in automatic speech processing tasks like SV and SR systems are obtained. These failures occur in TDSV system mainly because of the EPD problem, background noise, background speech and speech from the other competing speakers as well as impulses at the start and end of the speech which leads to degrading the overall performance of the system in terms of perceptual quality and intelligibility. The experimental evaluation is also taken under such degraded conditions to evaluate the performance of the system.

Initially the clean speech is used for training and clean test speech is taken recorded in controlled environment to evaluate the performance of the TDSV system. In the TDSV system, three selected predefined text dependent prompts used during training and testing are used. The scores are generated for each speaker test trial using DTW algorithm. The score level analysis is done on the three different basis (stringent, moderate and less moderate). The verification of claimed identity is performed during the testing phase using the sequence of the extracted feature vectors from test speech and the scores are obtained using the DTW approach against the 4 cohort speakers for each of the enrolled speakers. The controlled database results are used as the baseline performance using the cohort based method.

Next the database collected in the practical and uncontrolled environment which is mostly degraded speech. This is used to evaluate the performance of TDSV system. Using various methods for the speech recorded in uncontrolled environment, different sets of experiments are conducted on the voice biometric based attendance for the TDSV system in order to find the differences between the SV performances for the clean speech database condition and the degraded speech condition. A test speech is taken under such degraded conditions to evaluate the performance.

3.4.1 Evaluation of Text-Dependent Speaker Verification System

The database collected from the IITG voice server, is practical and under the influence of real time environmental conditions. This database contains the degraded speech. The text-dependent module of SV system is based on the three predefined prompts used during the training and testing speech utterances. Using DTW algorithm, the optimal warping path that provides the best matching under two speech segments is obtained. By carefully observing the score level analysis, it reveals that among the accumulated distances computed, the genuine claimed identity gives minimum distance compared to the cohort set scores.

The scores are generated for each test utterance using DTW algorithm. The algorithm calculates the distance of test utterance with each of the three training models and four fixed cohort set utterances for a particular speaker. The decision of a claimed test utterance is either accepted or rejected based on the cohort set distance scores. For each of the enrolled speakers, there exist a fixed pre-defined cohort set. The DTW distance score is compared to the scores generated from the claimed model and four cohort set for arriving the scores at the decision level. The test utterance is tested against the claimed model (which consists of three generated claimed scores and twelve cohort set scores obtained from the four cohort set of the claimed speakers). Hence, for a particular test utterance, 15 scores are generated which are sorted in the ascending order of their distance scores. The claimed identity is accepted if one out of the three distance scores (1/3), (2/3) and (3/3) generated for the claimed speakers, occupies a place in the top three position in the array of the 15 ascending distance scores. The TDSV system is evaluated and the performances are reported in EER and mDCF.

3.4.2 Results Using Energy Based Method on Clean Speech

The performance of the RSR2015 database under clean condition is evaluated on the RSR2015 Part I out of the three different cases for the challenging task. At the first level, optimal value of energy threshold for the baseline TDSV system is selected on the RSR2015 database and the performance of the TDSV system under clean speech condition is evaluated using the energy based EPD. The database is collected under controlled conditions and considered as the clean speech. The experiments on the clean speech are performed and the performance is measured in terms of EER and mDCF on the development data set Table 3.1 and evaluation data set. As can be seen from the Table 3.1

Table 3.1: The performance of the *i-vector* system on the RSR2015 database using development set of part I shows the evaluation in terms of equal error rate (EER) and minimum DCF for different test trials

Data	Male		Female	
Method	*i-vector*		*i-vector*	
Metrics	EER	DCF	EER	DCF
Development	3.41	0.35	6.94	0.47

all the genuine utterances are giving good performance in the development data set as compared to the imposter claim. Note that the performance for the clean condition is expected to be better than the degraded speech conditions which will be evaluated later and this validates the TDSV system

Table 3.2: The performance of the *i-vector* system on the RSR2015 database using evaluation set of part I shown in terms of equal error rate (EER) and minimum DCF for different test trials

Data	Male		Female	
Method	*i-vector*		*i-vector*	
Metrics	EER	DCF	EER	DCF
Evaluation	3.4	0.32	3.88	0.36

performance. By comparing the male and female data sets of the RSR2015 database, male speakers are giving better performance as compared to the female speakers in both development data set and evaluation data set.

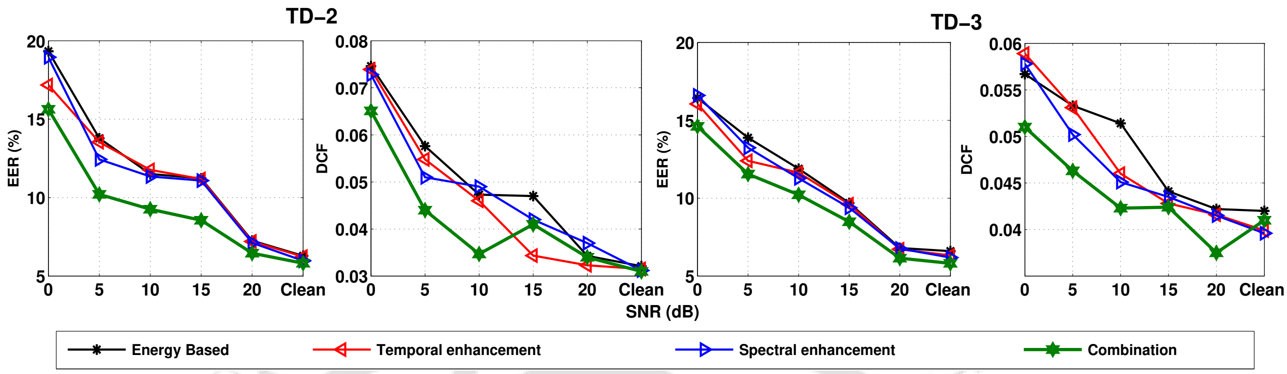
3.4.3 Results Using Energy Based Method on Degraded Speech

For all the experimental conditions considered using DTW algorithm SV systems provide better performance in clean condition and as the level of noise increases, the TDSV system performance reduces for each system decreases in RSR2015 database as shown in the 3rd and 4th columns of the (Table 3.3). For instance, in the RSR2015 database, the DTW algorithm based SV system EER increases from 7.59% to 20.29% as the noise varies from clean speech to degraded speech in terms of SNR ranging from 20-0 dB in steps of 5 dB. The performance improvement in the proposed system can be observed for the IITG database using energy based EPD method for the TDSV system shown in Table 3.4. The Table 3.4 and Table 3.3 are for the TD-1 case. The performance for the TD-2 and TD-3 case is shown in figure 3.6. The trend in the results are similar to the TD-1 case.

Table 3.3: Performance of TDSV systems in terms of EER and DCF using RSR2015 database using different methods.

Method –>	TD-1	Energy Based		Temporal enhancement		Spectral enhancement		Combination	
Database	SNR	EER (%)	DCF	EER (%)	DCF	EER (%)	DCF	EER (%)	DCF
	Clean	7.59	0.0388	7.4	0.0385	7.37	0.0382	6.62	0.0379
RSR2015	20 dB	8.1	0.0394	8.06	0.0393	7.77	0.039	6.95	0.0382
	15 dB	9.26	0.0404	8.65	0.0398	8.54	0.0395	7.58	0.0389
	10 dB	11.23	0.042	11.13	0.0418	10.48	0.0413	10.27	0.0409
	5 dB	12.51	0.0431	12.47	0.0429	12.37	0.0427	11.91	0.0425
	0 dB	20.29	0.0523	20.76	0.0496	19.47	0.0485	18.95	0.0481

Figure 3.6: Summary of TD-2 and TD-3 test trials, DTW based TDSV systems performance in terms of the EER and DCF for different experimental setup on RSR2015 database.



3.4.4 Results Using Energy Based End Point Detection Followed by Speech Enhancement

In this section the performance of the TDSV system when EPD is first performed followed by enhancement is analyzed. The performance can be found in the Table 3.3 in terms of EER and DCF on RSR2015 database using different methods. Initially the temporal enhancement is performed on the files after EPD and the performance of the TDSV system after temporal enhancement is found to be better compared to the case when only energy based EPD is performed. The reason being that the temporal enhancement is able to suppress the noise components in between the end points. The enhanced file may have characteristics of the clean speech and hence better performances are achieved. Similarly, the spectral enhancement which is basically spectral subtraction is performed after EPD and this process gives a better result than the temporal case. The results for this case are given in the Table 3.4 in terms of EER and DCF on IITG database using different methods.

3.4.5 Results Using Speech Enhancement Followed by Energy Based End Point Detection

This section presents the results of the same enhancement strategies but they are applied before the EPD for the TDSV system. The results are given in Table 3.3 and Table 3.4. It can be seen that enhancement followed by EPD shows better performance than the case earlier. The trend in the results of Table 3.3 and Table 3.4 showed that the combined temporal and spectral enhancement performed better compared to the temporal and spectral enhancement cases. The combined results are given in the last two column of the Table 3.3 and Table 3.4. The EER and mDCF showed improved performance

Table 3.4: Performance of the TDSV systems in terms of the EER and DCF using IITG database using different methods.

Method ->		Energy Based		Temporal enhancement		Spectral enhancement		Combination	
Database	SNR	EER (%)	DCF	EER (%)	DCF	EER (%)	DCF	EER (%)	DCF
IITG	TD-1	13.19	0.0422	12.57	0.0405	11.49	0.0418	11.11	0.0408
	TD-2	11.37	0.0404	10.86	0.0385	10.71	0.038	10.33	0.0366
	TD-3	18.42	0.0526	17.86	0.0509	17.46	0.0496	17.13	0.0486

when compared to the baseline energy based EPD method. The contribution of enhancement abilities of temporal and spectral case are combined and thus giving a better noise suppression. The best performance is obtained with the combination of temporal and spectral enhancement followed by EPD, respectively. The reason for the better performances of the enhancement followed by the EPD is again attributed to the noise suppression. First of all the TDSV system depends on the EPD and a good accuracy EPD system will give good results for the TDSV system. In the earlier case where the EPD is first performed prior to the enhancement, the end points obtained may not be accurate due to the presence of noise. However in case when the EPD is performed after enhancement, the enhancement modules suppress the noise and hence better EPD can be done on the enhanced signal, since now the signal will be having the quality nearer to the clean speech case. As a result the EPD using energy will be better when performed on the enhanced case compared to directly performing on the degraded case. Also the enhancement helps in noise suppression throughout the signal. The overall improvements of the TDSV system can be attributed to the better end points detected as well as the enhanced signal.

3.5 Conclusion and Future Work

In this work the performance of the TDSV system on the degraded speech conditions is analyzed. The temporal and spectral enhancement strategies are performed on the signal either before or after the energy based EPD. The enhancement strategies tend to improve the performance of the TDSV system where the case when enhancement is performed after EPD showed the best performance.

The future work can employ other forms of enhancement for improving the TDSV system. These enhancement strategies can be in terms of the source, vocal tract system and suprasegmental characteristics of speech. Additionally better versions of the EPD algorithms can be used in place of the energy based EPD to improve the accuracy of the overall TDSV system.

4

Analysis of the Hilbert Spectrum for Speaker Verification

Contents

Objective

This work explores the Hilbert spectrum (HS) of speech signal, constructed from its AM-FM components, or IMFs in characterizing speakers for the task of TDSV. The IMFs of speech signal are obtained using a non-linear and non-stationary data analysis technique called MEMD. The HS, which is a representation of the instantaneous frequencies and instantaneous energies of the IMFs is processed in short time-segments to generate features, which are then experimented for TDSV task. Two databases - RSR2015 and IITG - are utilized in this work, for validating the experimental findings. The performances of the TDSV system are evaluated for the individual features, and their combinations with the 39 dimensional MFCCs. To assess the practical utility of features, they are tested not only for clean speech, but also for speech corrupted by low-frequency (Babble) noise, and environmental noise. The experiments reveal that the features obtained from the HS, in combination with the MFCCs, enhances the performance of TDSV system. Further, the features extracted are effective at very low dimensions. Moreover, the features extracted from the HS are found to be consistently more effective than cepstral/ energy feature obtained from the raw IMFs, under noisy conditions.

4.1 Introduction

Speech is the principal mode of communication for human beings. It is a signal which contains a multitude of information such as the speakers age, height, emotion, accent, health and physiological disorders, identity, etc. [3,4]. It is also a signal which can be easily acquired, and at a low cost, which makes it a very attractive signal for the scientific community for various human-machine applications [3,4]. Hence, the forensic processing of the speech signal, i.e., the machine validation of the identity of the speaker from his/her speech signal, is one of the leading areas in speech processing today [3,4]. The *Mel filterbank*, designed to imitate the band pass filterbank characteristics of the human ear, remains the cornerstone for a majority of speech processing applications, including *Speaker Recognition* [3,4]. The cepstral or energy features obtained from the Mel filterbank, called the MFCCs, are considered the baseline features for most applications. However, it is fair to argue that the MFCCs may not represent the optimum features for all machine applications [111,112]. Hence, there have been many explorations for the purpose of finding alternative features for different machine applications [113–117]. In this direction, different methodologies based on non-stationary data analysis techniques, such as *Evolutionary Algorithms* [118–120] and *Wavelet Transform* [121,122], have been found to be useful in different speech processing applications. In the case of SR, it has been observed that the performance of the system, especially when using MFCCs as the features, degrades starkly under practical and challenging testing conditions [123]. Therefore, as in the case of other speech applications, alternative ways for SR are always being explored [112,124–133] In a bid to explore new features for characterizing speakers, we have investigated, in our recent work [42], the efficacy of *energy* or *cepstral* features obtained from a non-linear and non-stationary data analysis technique - EMD [134], and its variant - MEMD [135]. EMD is an adaptive AM-FM analysis technique which can decompose any real-world signal, without using any *a priori* basis, into its constituent AM-FM components, or *oscillatory functions*. These AM-FM components, called the IMFs, have been found to manifest important hidden characteristics of the signal [134]. Henceforth, EMD has gained widespread recognition for processing real-world signals for different real-world applications including that of speech processing [133, 134, 136–147]. Amongst many speech applications, EMD has been particularly useful in two important tasks - *Formants tracking* [135, 148], and *Pitch tracking* [149, 150]. This shows that the IMFs of the speech signal manifest the *vocal tract structure*, and the *glottal source* information, which are critical signatures of the speaker. Thus, EMD manifests the characteristics of the speech production system in its IMFs. This is contrary to the characteristics of the Mel filterbank, which is designed to represent the speech perception mechanism. Therefore, there might be benefits in using EMD for the task of TDSV - it may capture speaker specific cues in a different manner than the Mel filterbank. Based on this argument, our previous work [42] explored energy features derived from the IMFs of EMD and MEMD for the task of TISV. The objective in [42] was to investigate whether the *concise but adaptive filterbank nature* of EMD/MEMD could complement the Mel filterbank in the TISV task. The present work is a continuation of the study of the utility of the IMFs of the speech signal for extracting speaker-specific information. Thus, the motivation behind the TISV and current manuscript, is discussed below :

- **Hilbert Spectrum** : The IMFs were processed directly to obtain energy features from them. However, there is another representation of the IMFs, called the HS [111, 134]. The HS is the representation of the *instantaneous energies* of the IMFs in the *time-instantaneous frequency* plane. It is an image, which has found use in speech processing tasks like *Emotion Classification* [151]. More importantly, the HS has been utilized both for estimating the *pitch* or *fundamental frequency* (F_0) [150] of the speech signal, and in estimating the *vocal tract resonances* or the *formants* of the speech signal. As mentioned earlier, both pitch and formants are important speaker characteristics. This motivates us to explore the HS for extracting features which can manifest speaker signatures, which is the prime objective of this work.

- **TDSV** : Because of the uncontrolled variabilities that are present in human speech uttered at any particular instant, a TISV system is *data hungry* [3, 4]. By limiting the system to a fixed set of *phrases/ sentences/ lexicons*, the randomness of speech could be limited to a large extent. Such a system, called a TDSV system [3, 4], circumvents the requirement of large volumes of speaker-data that is utilized by any state-of-the-art TISV system, both for modeling the speaker space, and testing the system reliably. In our exploration of energy features for TISV [42], it was observed that when short duration utterances ($\leq$ 10 s) were used for testing against the pre-built speaker models, the relative improvement in performance was better (though the scope for improvement was also more) than that in the case of normal duration utterances. However, it must be noted that sufficient amount of speaker data was still available to model the feature space of the speaker. Hence, we ponder - *"Will the features derived from EMD/ MEMD still be effective if the data available for training is also limited ?"*. The aforementioned arguments motivate us to implement a TDSV system, in this work, where data of only a few seconds duration are available as reference or training templates of the speaker.

- **Interference** : The exploration of the energy features for TISV [42] was done on normal and short duration utterances, and for different speaking styles. However, all speech utterances used for training and testing were recorded under clean conditions. This work focuses on investigating the performance of a TDSV system, when the speech utterances are subjected to external interference. For this purpose, two databases are considered - RSR2015 database [14] and IITG database [38, 46]. In both databases, the speech utterances are recorded using a fixed set of mobile handsets/ tablets. The RSR2015 database is recorded in a clean environment, without any background noise. To simulate a crowded environment, the utterances used for testing are corrupted with varying degrees of Babble noise [110]. The IITG database is recorded in a natural environment, in a somewhat open hall where students could roam around, and walk in and out, while using the mobile handsets to give their attendances, with possibly their friends chattering around. Further, the students had to call a voice server using the mobile devices, thus incorporating telephone channel effects in the utterances. By using these two databases, we aim to evaluate whether proposed features will be robust to practical challenges.

- **MEMD** : In our exploration of energy features for TISV [42], it was observed that the features obtained from EMD, and MEMD, in combinations with the MFCCs, performed almost identically. However, it was observed that for different speaking styles, the energy features obtained from MEMD always provided optimum performances. For the EMD features, the optimum performances were observed at different dimensions for different speaking styles. This indicates that IMFs of MEMD provides more stable and consistent characteristics than EMD. Again, when used in a standalone fashion (without combining with the MFCCs), the energy features of MEMD exhibited superior performances compared to that of EMD. The principal reason for both of these observations may be attributed to reduced *mode-mixing* - an undesirable but unavoidable phenomenon that is a characteristic of EMD [111, 134]. Reduction in mode-mixing makes the IMFs more narrowband, and produces a more reliable HS [111], as is discussed elaborately in Section 4.2 of this manuscript. As such, in this work, only MEMD is used for experimentatio.

Thus, the objective of this work is to investigate the utility of the HS of the speech signal, obtained using MEMD, for the purpose of augmenting the performance of a TDSV system, under practical conditions. More specifically, in this work, features are extracted from the instantaneous frequencies and energies of IMFs of speech signal, which constitute its HS, for the task of TDSV. The features are derived for every 20 ms frame or segment of speech signal, with a frameshift of 10 ms. The features are tested individually, and in conjunction with the 39 dimensional MFCCs for the task. Just as in [42], the subset of IMFs that are useful for extracting the features are experimented in this work. This allows us to limit the number of IMFs extracted, and find the set useful for complementing the MFCCs in extracting speaker-specific information. The proposed features are also compared with the *refined Sum Log Squared Amplitude* feature explored in [42], to evaluate whether the HS representation is indeed useful in the TDSV setup. To complete the analysis, the features extracted from the HS are also tested on the NIST 2003 corpus [152] for TISV, using sufficient data to build the system. This allows us to observe the effect of data limitations and lexical variability in the performance of the SV system. The standard technique of DTW [3, 4, 94] is used to implement the TDSV system, whereas the state-of-the-art *i-vector* [12] is used to implement the TISV system.

The rest of the Chapter is organized as follows : Section 4.2 discusses the EMD and MEMD algorithms, and the HS. Section 4.3 describes the features extracted from the HS of the speech signal. Section 4.4 describes the experimental setup. Section 4.5 presents the experimental results and analysis. Finally, Section 4.6 summarizes and concludes this work.

4.2 EMD, MEMD and HS

EMD is an adaptive AM-FM analysis technique which decomposes any time-series dataset into a small finite set of AM-FM signals called its IMFs. These IMFs are obtained without using any prior knowledge about the characteristics of the signal, i.e, without using any suitable *a priori* basis, unlike in the case of techniques such as the Wavelet Transform [111]. This makes it a truly adaptive and independent technique for analyzing real world signals - every signal has its unique decomposition.

4.2.1 EMD, mode-mixing and MEMD

The EMD algorithm progressively decomposes the signal into its IMFs, and a *final residue* which represents the trend of the signal. Using a non-linear mechanism, called *sifting*, the signal is first decomposed into its high frequency component (IMF), and its low frequency component (*residue*). This residue is further operated upon by the *sifting* mechanism to obtain another IMF and another residue. This process is repeated, and finally a residue is obtained which cannot be further split - this is the final residue. For, a digital speech signal, $s(n)$, the EMD decomposition may be represented as,

$$s(n) = r_M(n) + \sum_{k=1}^{M} h_k(n) = \sum_{k=1}^{M+1} h_k(n) \tag{4.1}$$

where, $h_k(n)$ represents an IMF, $r_M(n)$ the final residue, and M the total number of IMFs extracted.

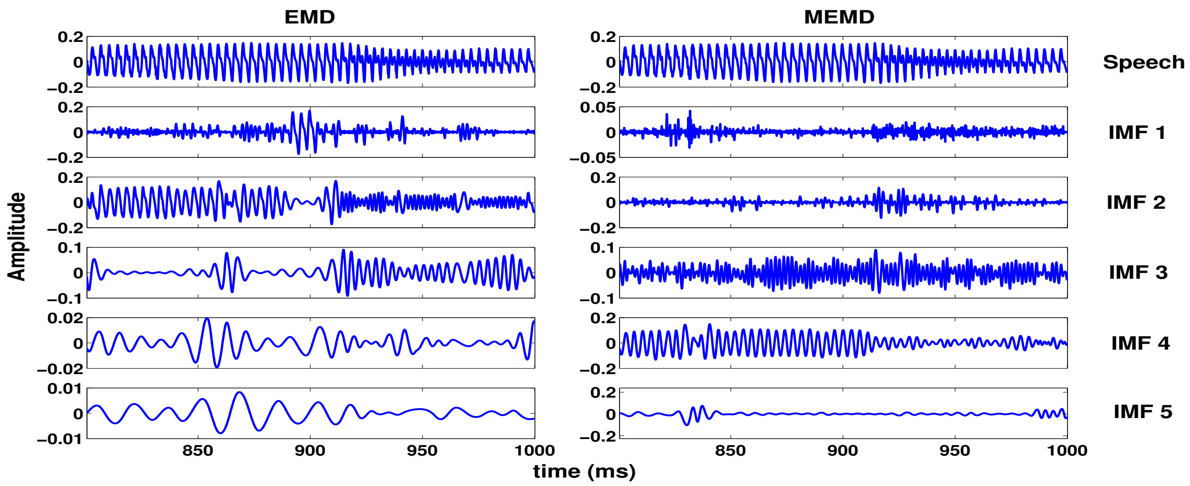


Figure 4.1: IMFs 1-5 generated from a speech signal, $s(n)$, using EMD (left column) , and MEMD (right column).

The left column of Figure 4.1 shows the first five IMFs of a digital speech signal, $s(n)$, taken from the RSR2015 corpus, obtained using EMD. One may easily observe that both high frequency oscillations and low frequency oscillations are present at different times within the same IMF, particularly in the case of IMFs 1-3. This phenomenon is called mode-mixing. Mode-mixing represents the distribution of a narrowband signal component across different IMFs at different points of time, and conversely the manifestation of oscillatory functions of disparate frequencies in the same IMF. Thus, mode-mixing makes the IMFs less narrowband, in a global sense, leading to less accurate estimations of their instantaneous frequencies and amplitude envelopes. Evidently, mode-mixing makes it difficult to segregate or characterize a certain subset of the IMFs as being useful for analysis, for a particular task. Hence, in this work, we utilize a recently proposed variant of EMD, called MEMD, which reduces mode-mixing in the IMFs of the signal [135]. The MEMD algorithm is structurally same as the EMD algorithm but employs a *modified sifting process*, which enables detection of the higher frequency content of the speech signal more effectively. As the end result, the IMFs of MEMD have been shown to provide a better segregation of speech spectrum amongst them, thus providing a better and more precise manifestation of the formants [135]. The right column of Figure 4.1 shows the first five IMFs obtained using MEMD on the same digital speech signal, $s(n)$, which is decomposed by EMD. As is observed from the figure, the IMFs of MEMD exhibit far lesser mode-mixing than that of EMD. Also, the IMFs represent a more gradual shift from higher to lower frequencies in the case of MEMD.

Theoretically, an IMF is defined as a locally symmetric signal, such that the number of extrema in the IMF is equal to its number of zero crossings [134]. This ensures that the IMFs produce a reliable HS, which is discussed in the next subsection. To ascertain that the IMFs exhibit their theoretical properties as closely as possible, a number of *sifting criteria* have been proposed [153]. And these characteristics are well adhered to when the number of *sifting iterations*, N, in the sifting mechanism, is kept around ten [153, 154]. Henceforth, in this work, a fixed number of sifting iterations, $N = 10$, is used in every sifting process in the EMD/ MEMD algorithm. Again, as discussed earlier, the decomposition stops when the residue takes the form of a trend. This means that the number of extrema in $r_M(t)$ is two or less [134]. Practically, however, the decomposition is stopped when a user-defined maximum number of IMFs, M, has been extracted. This is done in order to avoid unnecessary generation and processing of higher-order IMFs, which represent very low frequency trend like signals. Hence, in this work, the decomposition is curtailed to a maximum of ten components ($M = 9$), for the EMD/ MEMD algorithm.

$$s(n) = r_9(n) + \sum_{k=1}^{9} h_k(n) = \sum_{k=1}^{10} h_k(n) \tag{4.2}$$

4.2.2 The Hilbert Spectrum

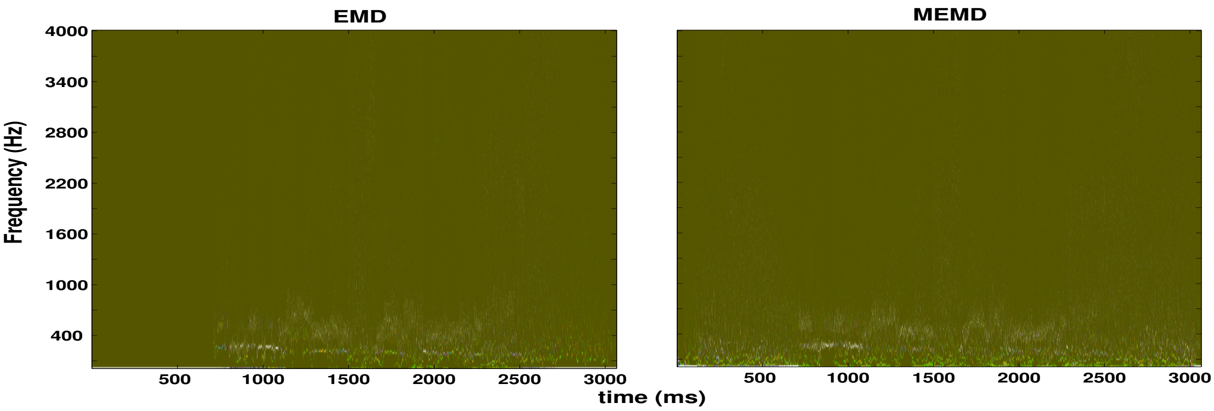


Figure 4.2: Hilbert spectrum of the speech signal used in Figure 4.1, constructed from the IMFs of EMD (left), and MEMD (right) respectively.

As is evident from Figure 4.1, the IMFs represent AM-FM signals whose center frequencies decrease as the order of the IMF increases. Having derived an IMF, the Hilbert Transform may be used to estimate its dynamically changing frequency and amplitude envelope. The Hilbert Transform is a reliable estimator of the frequency and amplitude envelope functions of an AM-FM signal, provided certain conditions are met [111, 112]. These conditions are :

Assuming these conditions are satisfied, the Hilbert Transform, $H[x(t)]$, of an analog signal, $x(t)$, is derived from its Fourier Transform, as,

$$x(t) \leftrightarrow X(\omega) \ ,$$

$$\frac{1}{\pi t} \leftrightarrow -j \ sgn(\omega) = \begin{cases} -j \ , \ \omega > 0 \\ j \ , \ \omega < 0 \end{cases} ,$$

$$H[x(t)] = x(t) * \frac{1}{\pi t} \ ,$$

$$H[x(t)] \leftrightarrow -j \ sgn(\omega) X(\omega) = \begin{cases} -jX(\omega) \ , \ \omega > 0 \\ jX(\omega) \ , \ \omega < 0 \end{cases}$$

(i) : The frequency variation should not be large, i.e., the signal should be narrowband.

(ii) : The amplitude variation should not be large.

(iii) : The rate of frequency and amplitude variation should not be large.

The instantaneous frequency function, $f(t)$, and amplitude envelope function, $a(t)$, is derived from the analytical signal, $z(t)$, which is devoid of any negative frequency Fourier components.

$$z(t) = x(t) + jH[x(t)] = a(t)e^{j\phi(t)} \ , \tag{4.3}$$

$$z(t) \leftrightarrow Z(\omega) = \begin{cases} 2X(\omega) \ , \ \omega > 0 \\ \ \ 0 \ , \ \omega < 0 \end{cases} ,$$

$$a(t) = |z(t)| \ , \tag{4.4}$$

$$\phi(t) = \arctan \frac{\Im\{z(t)\}}{\Re\{z(t)\}} \ , \ f(t) = \frac{1}{2\pi}\frac{d\phi(t)}{dt} \tag{4.5}$$

Correspondingly, the *Discrete Fourier Transform* (DFT) is utilized for estimating the instantaneous frequency and amplitude envelope of any discrete time signal, $x(n)$. Hence, if $a_k(n)$ and $f_k(n)$ represent the instantaneous amplitude envelope and frequency of $h_k(n)$, the HS may be represented as,

$$H(f, n) = \left\{a_k^2(n) \mid f_k(n), n\right\} \ , \ k = 1, ..., K \leq M + 1 \tag{4.6}$$

Thus, the HS is the distribution of the *instantaneous energies* of the IMFs, in the *time-instantaneous frequency* plane. Figure 4.2 shows the Hilbert spectra of the speech signal used in Figure 4.1, obtained using EMD and MEMD, where all its components ($K = M + 1 = 10$) are used. Readers are advised to refer to the softcopy of this manuscript, rather than hard print, for better visualization of the figure. In the figure, one can observe more dots (particularly between 0-500 Hz) in the HS of MEMD compared to that of EMD. Again, there are more dots in the higher frequencies (above 500 Hz) of the HS of MEMD compared to that of EMD. This illustrates the ability of MEMD to provide a better representation of the higher frequency content of the speech signal [135]. This is expected, as the IMFs obtained from MEMD (as observed in Figure 4.1) meet the requirements of the Hilbert Transform more closely. Of course, even the IMFs of MEMD only partially meet the three requirements - so, there are bound to be inaccuracies in the estimates of the instantaneous frequencies and amplitude envelopes. Hence, one could argue that extracting a multitude of AM-FM signals from the speech signal, using narrowband filters, could provide a better HS. Unfortunately, that would result in a number of spurious AM-FM components, which may in fact obscure the meaningful information. EMD (hence its variants) has a unique property - it acts as a *concise and adaptive* filterbank on the signal, thereby generating a small number of *meaningful* components [111]. This property gives a useful trade-off between the stringent requirements of the Hilbert transform and the number of meaningful AM-FM components, resulting in unprecedented benefits [42]. Therefore, in this work, we restrict our experiments to the HS constructed from the IMFs of MEMD only, and investigate its utility in capturing speaker-specific information.

4.3 Utilizing the Constituents of the HS for Characterizing Speakers

As has been thoroughly discussed in [42], EMD/ MEMD acts as a concise but adaptive implicit filterbank on the speech signal, extracting its IMFs. While the Mel filterbank is a fixed structure, the implicit filterbank structure of EMD/ MEMD is adaptive to the nature of speech signal at every instant. Therefore, even though energy or cepstral features were extracted from the IMFs, analogous to the MFCCs, they captured information that was different to the MFCCs and helped in enhancing the overall performance of the TISV system [42]. In this work, though the IMFs remain the same, we extract features which are quite different from the cepstral features explored in our previous work. Instead of processing the IMFs of speech signal directly, we process their instantaneous frequencies and energies - the constituents of HS of speech signal. As such, we should expect the features extracted to capture different sort of information, or the same information in a different manner, which hopefully should be useful in characterizing the speaker.

4.3.1 HS as a Method of Masking mode-mixing

In [42], we observed the capability of time-domain IMFs of speech signal in capturing speaker-specific information. Hence, it is quite natural to expect the HS derived from the IMFs to also manifest speaker-specific cues. It is our hypothesis that the HS, in the form of an image, captures patterns specific to the speaker. In order to analyze the HS further, we split it into its constituents - the instantaneous frequencies and instantaneous energies of the IMFs. Figure 4.3 shows the instantaneous frequencies and instantaneous energies of the first four IMFs (derived using MEMD), corresponding to a 20 ms segment of the speech signal from which the HS in Figure 4.1 has been constructed. The instantaneous frequencies and instantaneous energies of rest of the IMFs are not plotted for visual clarity. Thus, at each time instant, there are four instantaneous frequency values, and four corresponding instantaneous energy values. The instantaneous energy values are normalized at each time instant, and the instantaneous frequencies are represented in kHz to reduce the dynamic range.

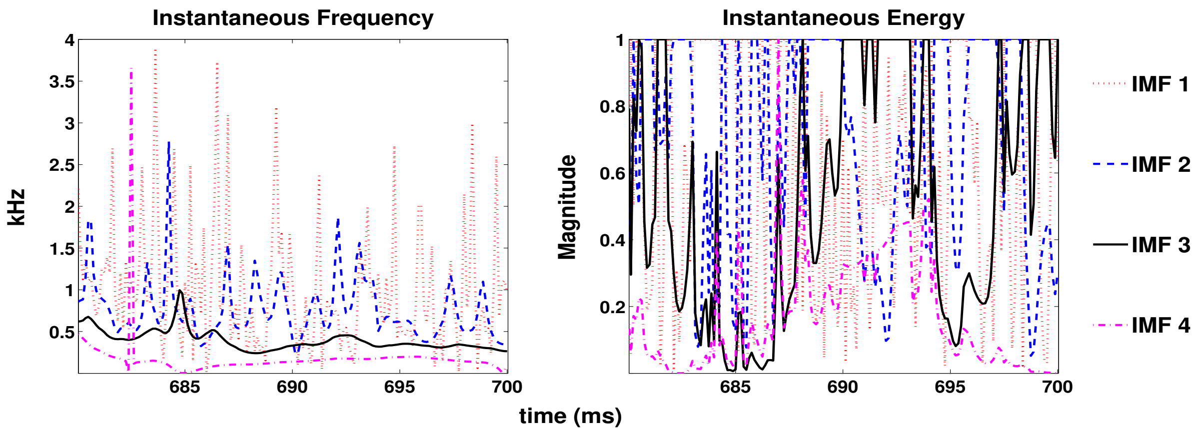


Figure 4.3: Instantaneous frequencies (kHz) and normalized instantaneous energies of the first four IMFs (derived using MEMD) corresponding to the HS shown in Figure 4.2.

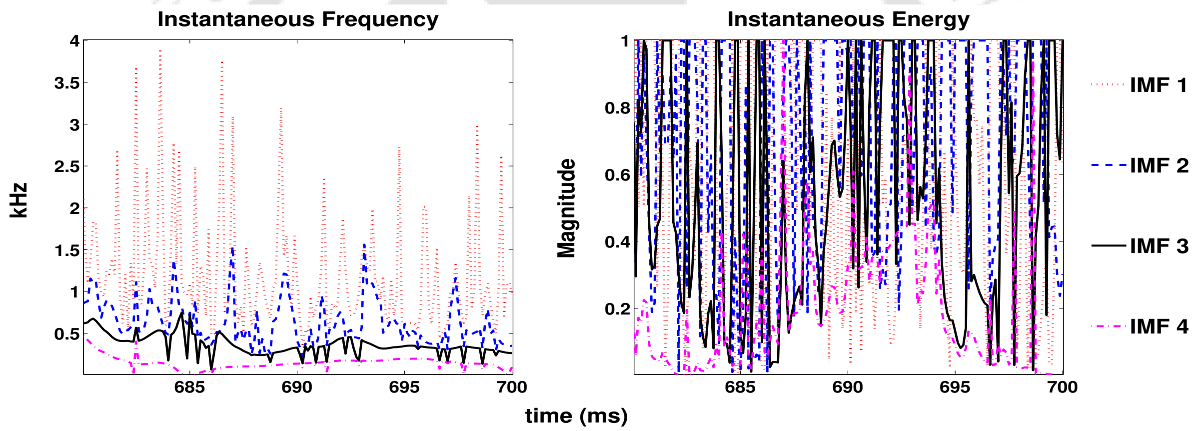


Figure 4.4: Instantaneous frequencies (kHz), *sorted in decreasing order*, and the corresponding rearranged normalized instantaneous energies of the first four IMFs (derived using MEMD) corresponding to Figure 4.3.

As is evident from the plots of instantaneous frequencies, the IMFs exhibit some degree of mixing or overlap of information. This may be more or less severe for another segment of speech signal. Even though mode-mixing is reduced in MEMD, it is not completely eliminated. Because of the inevitable existence of mode-mixing, there is always a certain degree of inaccuracy in the estimated instantaneous frequencies and instantaneous energies. Hence, it is beneficial to extract features from them after some degree of averaging over the time segment. This, of course, also reduces the feature space, and enables the features derived to be concatenated with the MFCCs, which are obtained for every frame or time-segment, after dividing the entire speech signal/ utterance into overlapping frames.

The representation of instantaneous frequencies and instantaneous energies in the form of HS in a way hides the underlying phenomenon of mode-mixing that is manifested in the IMFs. The information regarding source IMF corresponding to a certain point in the image is lost. Hence, features may be derived by processing this image itself in blocks, i.e, by segregating the image into pre-defined frequency bands and time-segments or frames, as was explored in [151]. This, however, raises the question regarding the number and structure of the subbands. Also, one may argue that predefining the subbands may result in unfruitful processing, as there may be certain subbands which have little representation in the HS. Therefore, in this work, we investigate the two constituents of the HS separately. At every time instant there are ten instantaneous frequency values, corresponding to $M + 1 = 10$ IMFs. Correspondingly, there are ten instantaneous energy values at every time instant. By using this split representation, we circumvent the requirement of artificially partitioning the frequency (or energy) space.

However, to conceal the effect of mode-mixing (as in the case of HS), before further processing, the instantaneous frequencies and energies need to be redistributed. If there were no mode-mixing, then, in Figure 4.3, at every time instant, the first IMF, $h_1(n)$, would provide the highest instantaneous frequency. The second highest frequency would be contributed by $h_2(n)$, the third highest by $h_3(n)$, and the fourth highest by $h_4(n)$. Based on this idea, the four instantaneous frequencies at every time instant are sorted in decreasing order of frequency. Based on this sorting of the instantaneous frequencies, the instantaneous energies are then rearranged. Figure 4.4 shows the plots of the rearranged instantaneous frequencies and energies, for the IMFs shown in Figure 4.3. The legends in Figure 4.4 do not correspond to the actual IMFs, but the IMFs if there were no mode-mixing at all.

Let us consider IMF_1 first. In Figure 4.3, IMF_1 has instantaneous frequencies ranging approximately between 0 - 4 kHz. In Figure 4.4, however, IMF_1 does not have any instantaneous frequency below $\sim$ 0.5 kHz. Similarly, for IMF_2, in the unsorted case, the instantaneous frequencies vary approximately between 0.5 - 2.5 kHz, which are restrained to 0.5 - 1.5 kHz after sorting. Again, in the case of IMF_3, the unsorted instantaneous frequencies range approximately between 0.25 - 1 kHz, whereas the sorted instantaneous frequencies vary approximately between 0.1 - 0.75 kHz. The instantaneous frequencies of IMF_4 remain more or less the same (approximately between 0 - 0.5 kHz) even after sorting. Thus, the sorting or rearrangement process attempts to mask the large variations in the instantaneous frequencies of the IMFs. It might seem in Figure 4.4 that discontinuities (abrupt variations in instantaneous frequencies) have been introduced in IMFs 2 and 3 by the rearrangement process. However, on closer inspection, one would realize that considering the overall time segment

(680 - 700 ms), their variations have been curtailed. In other words, the sorted instantaneous frequencies represent IMFs which adhere consistently to the information they manifest. Again, comparing Figures 4.3 and 4.4, one may feel that the rearrangement process results in a random distribution of the instantaneous energies. However, as we will observe in Section 4.4 (Results and Analysis), the features obtained from the instantaneous energies are useful in discriminating speakers. Therefore, even though, visually, the instantaneous energies seem to lack any pattern, they, in fact, contain speaker specific information, which we investigate in this work.

The rearrangement process is further clarified in the next subsection using a simple example. It is our hypothesis that the sorted instantaneous frequencies and instantaneous energies contain patterns, specific to the speaker, which may be utilized for extracting features for the task of TDSV.

4.3.2 Feature Extraction From the Instantaneous Frequencies and Energies

Based on the preceding discussion, we now describe various features that are extracted from the IMFs of speech signal, for the purpose of TDSV. As represented by equation (4.2), ten IMFs are obtained from the speech signal, $s(n)$. Let $K \leq M+1 = 10$ be the number of IMFs from which the instantaneous frequencies, $\{f_k(n)\ ,\ k = 1, ..., K\}$, and the instantaneous energies, $\{|a_k(n)|^2\ ,\ k = 1, ..., K\}$, are extracted. Let $s^i(n)$ represent the i^{th} frame of $s(n)$, given a frame size of 20 ms, and a frame shift of 10 ms. Then, $h^i_k(n)$ represents i^{th} frame of the k^{th} IMF of $s(n)$. Correspondingly, $f^i_k(n)$ and $|a^i_k(n)|^2$ represent the instantaneous frequencies and energies, respectively, of $h^i_k(n)$, extracted using Hilbert Transform. Let N_f represent the number of samples in a 20 ms frame. Then, the instantaneous frequencies and energies are rearranged.

To clarify the rearrangement process, we take a hypothetical example of i^{th} speech frame having $N_f = 5$ samples. We consider that only the first four IMFs are used to extract the instantaneous frequencies and energies. Their values are given below.

$$\begin{bmatrix} & f^i_1(n) & f^i_2(n) & f^i_3(n) & f^i_4(n) \\ n=0 & 1.93 & 0.60 & 0.52 & 0.14 \\ n=1 & 0.46 & 0.59 & 0.51 & 0.14 \\ n=2 & 0.18 & 0.70 & 0.49 & 0.13 \\ n=3 & 0.77 & 1.57 & 0.48 & 0.11 \\ n=4 & 1.38 & 2.78 & 0.51 & 0.09 \end{bmatrix} \begin{bmatrix} & |a^i_1(n)|^2 & |a^i_2(n)|^2 & |a^i_3(n)|^2 & |a^i_4(n)|^2 \\ n=0 & 1.00 & 0.24 & 0.08 & 0.02 \\ n=1 & 1.00 & 0.61 & 0.24 & 0.09 \\ n=2 & 1.00 & 0.14 & 0.10 & 0.05 \\ n=3 & 1.00 & 0.25 & 0.66 & 0.43 \\ n=4 & 1.00 & 0.20 & 0.10 & 0.08 \end{bmatrix}$$

After rearrangement, the instantaneous frequencies and energies may be represented by the following matrices :

$$\left[\begin{array}{ccccc} & f_1^i(n) & f_2^i(n) & f_3^i(n) & f_4^i(n) \\ n=0 & 1.93 & 0.60 & 0.52 & 0.14 \\ n=1 & 0.59 & 0.51 & 0.46 & 0.14 \\ n=2 & 0.70 & 0.49 & 0.18 & 0.13 \\ n=3 & 1.57 & 0.77 & 0.48 & 0.11 \\ n=4 & 2.78 & 1.38 & 0.51 & 0.09 \end{array}\right] , \left[\begin{array}{ccccc} & |a_1^i(n)|^2 & |a_2^i(n)|^2 & |a_3^i(n)|^2 & |a_4^i(n)|^2 \\ n=0 & 1.00 & 0.24 & 0.08 & 0.02 \\ n=1 & 0.61 & 0.24 & 1.00 & 0.09 \\ n=2 & 0.14 & 0.10 & 1.00 & 0.05 \\ n=3 & 0.25 & 1.00 & 0.66 & 0.43 \\ n=4 & 0.20 & 1.00 & 0.10 & 0.08 \end{array}\right]$$

Having obtained the rearranged instantaneous frequencies and energies, the following features are then extracted corresponding to each speech frame.

Mean Instantaneous Frequency : It is obtained as,

$$F_k^i = \frac{1}{N_f} \sum_{n=0}^{N_f-1} f_k^i(n)/1000 \ , \ k = 1, ..., K \leq M+1 \tag{4.7}$$

$$F_K^i := [F_1^i, F_2^i, ..., F_K^i]^T \ , \ F_K^i := \Gamma\{F_K^i\} \tag{4.8}$$

$$F_K = \{F_K^i, \forall i \in \mathbb{N}\} \tag{4.9}$$

In equation 4.7, $f_k^i(n)$ is divided by 1000 to represent it in kHz, thus reducing the dynamic range. In this work, Γ denotes the operation of *Discrete Cosine Transform* (DCT) [3, 4], used to refine the raw feature vector, obtained from each frame. The DCT of any K-dimensional feature vector extracted from the i^{th} speech frame, $V_K^i = [v_1^i, v_2^i, ..., v_K^i]^T$, is hence denoted by $\Gamma\{V_K^i\}$, and obtained as,

$$cv_p^i = \begin{cases} \frac{1}{\sqrt{K}} \sum_{k=1}^{K} v_k^i \cos\left\{\frac{\pi(2k-1)(p-1)}{2K}\right\} \ , \ p = 1 \\ \frac{1}{\sqrt{2K}} \sum_{k=1}^{K} v_k^i \cos\left\{\frac{\pi(2k-1)(p-1)}{2K}\right\} \ , \ p = 2, ..., K \end{cases} , \tag{4.10}$$

$$\Gamma\{V_K^i\} = [cv_1^i, cv_2^i, ..., cv_K^i]^T \tag{4.11}$$

Absolute Deviation of Instantaneous Frequency : It is obtained as,

$$\Delta F_k^i = \frac{1}{N_f} \sum_{n=0}^{N_f-1} |\frac{f_k^i(n)}{1000} - F_k^i| \ , \ k = 1, ..., K \leq M+1 \tag{4.12}$$

$$\Delta F_K^i := [\Delta F_1^i, \Delta F_2^i, ..., \Delta F_K^i]^T \ , \ \Delta F_K^i := \Gamma\{\Delta F_K^i\} \tag{4.13}$$

$$\Delta F_K = \{\Delta F_K^i, \forall i \in \mathbb{N}\} \tag{4.14}$$

Correlation between Instantaneous Frequencies : It is obtained as,

$$\sigma F_k^i = \sum_{n=0}^{N_f-1} [\ \{\frac{f_k^i(n)}{1000} - F_k^i\}\{\frac{f_{k+1}^i(n)}{1000} - F_{k+1}^i\}\ ]\ ,\ k = 1, ..., K-1 \leq M \tag{4.15}$$

$$\sigma F_{K-1}^i \coloneqq [\sigma F_1^i, \sigma F_2^i, ..., \sigma F_{K-1}^i]^T\ ,\ \sigma F_{K-1}^i \coloneqq \Gamma\{\sigma F_{K-1}^i\} \tag{4.16}$$

$$\sigma F_{K-1} = \{\sigma F_{K-1}^i, \forall i \in \mathbb{N}\} \tag{4.17}$$

Mean Instantaneous Energy : It is obtained as,

$$|a_k^i(n)|^2 \coloneqq \frac{|a_k^i(n)|^2}{\max\limits_k |a_k^i(n)|^2}\ ,\ k = 1, ..., K \leq M+1 \tag{4.18}$$

$$E_k^i = \frac{1}{N_f} \sum_{n=0}^{N_f-1} |a_k^i(n)|^2\ ,\ k = 1, ..., K \leq M+1 \tag{4.19}$$

$$E_K^i \coloneqq [E_1^i, E_2^i, ..., E_K^i]^T\ ,\ E_K^i \coloneqq \Gamma\{E_K^i\} \tag{4.20}$$

$$E_K = \{E_K^i, \forall i \in \mathbb{N}\} \tag{4.21}$$

Absolute Deviation of Instantaneous Energy : It is obtained as,

$$|a_k^i(n)|^2 \coloneqq \frac{|a_k^i(n)|^2}{\max\limits_k |a_k^i(n)|^2}\ ,\ k = 1, ..., K \leq M+1 \tag{4.22}$$

$$\Delta E_k^i = \frac{1}{N_f} \sum_{n=0}^{N_f-1} |\ \{|a_k^i(n)|^2 - E_k^i\}\ |\ ,\ k = 1, ..., K \leq M+1 \tag{4.23}$$

$$\Delta E_K^i \coloneqq [\Delta E_1^i, \Delta E_2^i, ..., \Delta E_K^i]^T\ ,\ \Delta E_K^i \coloneqq \Gamma\{\Delta E_K^i\} \tag{4.24}$$

$$\Delta E_K = \{\Delta E_K^i, \forall i \in \mathbb{N}\} \tag{4.25}$$

Correlation between Instantaneous Energies: It is obtained as,

$$|a_k^i(n)|^2 \coloneqq \frac{|a_k^i(n)|^2}{\max\limits_k |a_k^i(n)|^2}\ ,\ k = 1, ..., K \leq M+1 \tag{4.26}$$

$$\sigma E_k^i = \sum_{n=0}^{N_f-1} [\ \{|a_k^i(n)|^2 - E_k^i\}\{|a_{k+1}^i(n)|^2 - E_{k+1}^i\}\ ]\ ,\ k = 1, ..., K-1 \leq M \tag{4.27}$$

$$\sigma E_{K-1}^i \coloneqq [\sigma E_1^i, \sigma E_2^i, ..., \sigma E_{K-1}^i]^T\ ,\ \sigma E_{K-1}^i \coloneqq \Gamma\{\sigma E_{K-1}^i\} \tag{4.28}$$

$$\sigma E_{K-1} = \{\sigma E_{K-1}^i, \forall i \in \mathbb{N}\} \tag{4.29}$$

Instantaneous Energy weighted Instantaneous Frequency: It is obtained as,

$$|a_k^i(n)|^2 := \frac{|a_k^i(n)|^2}{\max\limits_k |a_k^i(n)|^2} \ , \ k = 1, ..., K \leq M + 1 \tag{4.30}$$

$$\Upsilon_k^i = \sum_{n=0}^{N_f - 1} \frac{|a_k^i(n)|^2}{\sum_{k=0}^{K} |a_k^i(n)|^2} f_k^i(n) \ , \ k = 1, ..., K \leq M + 1 \tag{4.31}$$

$$\Upsilon_K^i := [\Upsilon_1^i, \Upsilon_2^i, ..., \Upsilon_K^i]^T \ , \ \Upsilon_K^i := \Gamma\{\Upsilon_K^i\} \tag{4.32}$$

$$\Upsilon_K = \{\Upsilon_K^i, \forall i \in \mathbb{N}\} \tag{4.33}$$

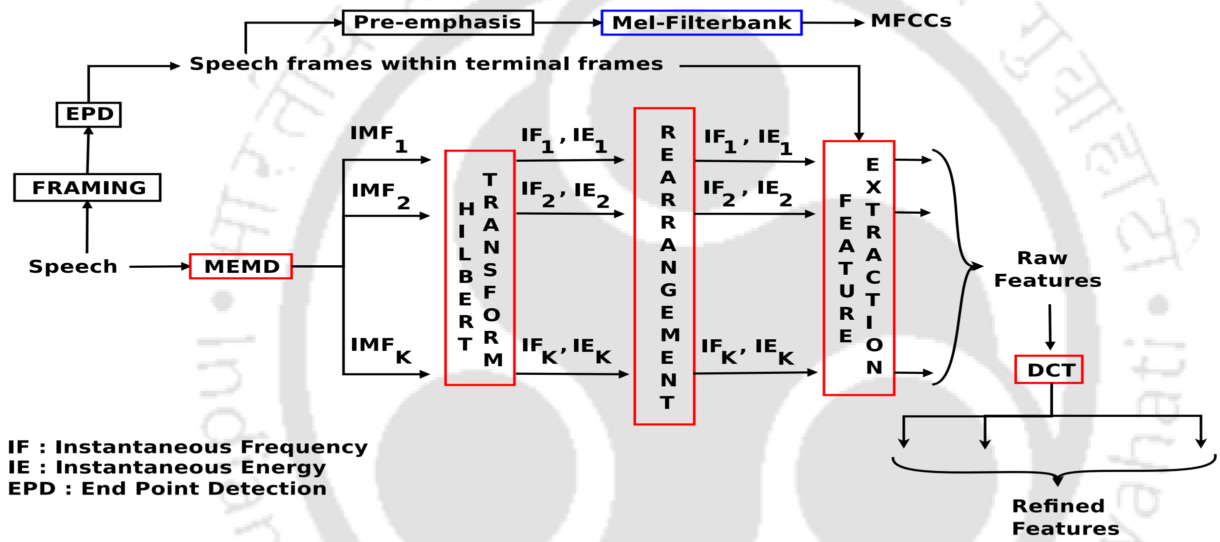


Figure 4.5: Feature Extraction Process for the TDSV setup.

Figure 4.5 shows the entire feature extraction process in a schematic diagram. Apart from the features derived from the instantaneous frequencies and energies of the IMFs, 39 dimensional MFCCs are also extracted, with the expectation that their combinations would significantly enhance the capability of TDSV system in discriminating speakers. The *refined Sum Log Squared Amplitude* feature, cG_K, which performed well in the TISV task [42], is also extracted from the raw IMFs for the sake of comparison.

4.3.3 Motivation Behind the Features

It has been shown in [111, 135] that the first few IMFs of speech signal manifest the *vocal tract resonances* of speech signal. In fact, the HS constructed from the IMFs has been utilized to estimate the vocal tract resonances. The resonances are produced by the different cavities that are formed when the speaker utters a speech signal [3, 4]. The cavities formed are not only dependent on the sound produced, but are also influenced by the physical structure of the vocal tract, which is never the same for two different speakers. Similarly, during the production of continuous speech, the fine and coarse changes in the cavities are not only dependent on the sounds produced, but are also speaker-specific. Hence, the first few IMFs (which manifest the vocal tract resonances) may be expected to carry important speaker signatures. Therefore, the HS may also be expected to manifest important speaker signatures.

Just like the vocal tract resonances, the rate of vibration of the vocal folds in the glottis during the production of *voiced speech* (vowel-like sounds) is considered an important speaker signature. For any speaker, the rate of vibration of his/ her vocal folds, known as his/ her pitch or fundamental frequency, depends on the anatomy of his/ her glottis [3, 4]. Hence, the pitch frequencies of adult males, adult females and children are quite different from one another. Moreover, speakers involuntarily also exhibit variations in their pitch frequency during normal communication. The range of such variations are also speaker-specific [3, 4]. It has been observed that certain IMFs of speech signal also manifest pitch frequency of the speaker or its harmonics [111, 155]. In fact, in [150], the lower-frequency portion of the HS was processed to identify high energy bands from which the pitch frequency of the speaker was tracked. Therefore, we may expect the HS to capture important speaker signatures corresponding to his/ her glottis.

Based on the above discussion, in this work, we explore the HS for extracting speaker-specific features. Depending on the intensity or strength of the instantaneous energies, certain instantaneous frequencies are more discernible in the HS, whereas others are not visible though present. More specifically, certain bands are vivid in the HS, whereas others are subdued. It is expected that the different bands in the HS would represent the anatomical characteristics of the vocal tract or the glottis. By using the constituents of the HS separately, we have, in fact, segregated out the different frequency bands and their corresponding energy bands that may represent useful speaker characteristics. The proposed seven features aim to utilize these frequency and energy bands to compute speaker signatures that can complement the MFCCs.

The motivation behind the F_K feature is to capture the dominant frequencies of the different frequency bands that constitute the HS. F_K represents the center frequencies of the first K frequency bands (in decreasing order of frequency) present in the HS. If the frequency bands constituting the HS are influenced by the anatomy of the speech production system, the F_K feature could act as an important speaker-specific cue.

The motivation behind the ΔF_K feature is to capture the fluctuations, spreads or widths of the first K frequency bands (in decreasing order of frequency) constituting the HS. For a speech signal, some frequency bands of the HS might show large fluctuations, whereas other frequency bands might be relatively steady. If the range of such fluctuations are a manifestation of the anatomical movements of the speech production system, the ΔF_K feature could be useful in characterizing the speaker.

The σF_{K-1} feature attempts to capture the dependence or correlation between successive frequency bands of the HS. For the first K frequency bands (in decreasing order of frequency), there are $K-1$ successive frequency bands, and hence $K-1$ correlation values are obtained. For a given speech utterance, it is possible that while a certain frequency band of its HS vary in a particular manner, its nearest frequency band might not be affected at all. On the other hand, a different pair of adjacent frequency bands might be closely related. If the degree of such interband relations are speaker-specific, σF_{K-1} could serve as an useful speaker specific feature.

The E_K feature represents the average energy content of the different frequency bands of the HS. For the first K frequency bands (in decreasing order of frequency), there are corresponding K energy bands. E_K represents the mean value of these K energy bands. We hypothesize that different speakers may radiate different levels of energies at different frequencies, and hence the E_K feature could be useful in discriminating them.

The ΔE_K feature represents the energy variations in different frequency bands of HS. Corresponding to the first K frequency bands (in decreasing order of frequency) in the HS, there are corresponding K energy bands. Assume that a particular speaker can emit certain frequencies at varying strengths, whereas he/ she can only emit certain other frequencies at only specific strengths. Then, we may observe large energy fluctuations in certain frequency bands (energy bands) of HS, whereas in other frequency bands (energy bands), the instantaneous energies may be relatively steady. Then, the ΔE_K feature may serve as an important speaker specific cue.

The σE_{K-1} feature has the same motivation as the σF_{K-1} feature, but attempts to capture the relation between the successive frequency bands in terms of their energy variations. For the first K frequency bands (in decreasing order of frequency), there are corresponding K energy bands. It is possible that the increase in energy in a particular energy band affects its succeeding energy band. The degree of such dependencies may be speaker-specific, and hence σE_{K-1} may serve as an important feature of the speaker.

The Υ_K feature merges the information contained in instantaneous frequencies and instantaneous energies of the HS. Assume that, for any particular speaker, the pattern and degree of fluctuations in a frequency band is correlated with that of the corresponding energy band. For different speakers such correlations may vary. The Υ_K feature captures the correlations between the first K frequency and energy bands, hoping to capture speaker-specific cues superior to the features obtained solely from either the instantaneous frequencies or the instantaneous energies.

4.4 Experimental Setup

Two databases, RSR2015 [14], and IITG [38,46], which represent speakers from two different countries - Singapore, and India - in South-East Asia, are used in this work for the task of TDSV. RSR2015 contains audio recordings, recorded on mobile phones and tablets, in a closed office environment, from 300 speakers (143 female and 157 male). 30 sentences were recorded from the speakers in nine sessions (utterances) each. The speakers were selected to be representative of the ethnic distribution of Singaporean population, with age ranging from 17 to 42. Out of the 30 sentences, three sentences have been randomly selected in this work for the TDSV task. For the three sentences, the duration of utterances vary between 3-5 s, over all the speakers and sessions. The sentences are :

TD-1: *"Only lawyers love millionaires"*.

TD-2: *"I know I did not meet her early enough"*.

TD-3: *"The Birthday party has cupcake and ice-cream"*.

The IITG database consists of speech utterances recorded as part of development of a students attendance system in the EEE department of IITG, Guwahati, India. For this work, a total of 325 speakers (286 male and 39 female) have been used. The speakers represent an adult age group between 25-35 years. For recording the utterances, the speakers were required to call a voice-server using a fixed set of mobile phones. In the training phase, three utterances (sessions), each for three speech sentences, were recorded for each speaker in isolation, using one of the handsets. During the testing phase, the speakers uttered one of the sentences, randomly picked by the voice-server, for a period of six months. On an average, 30 utterances were recorded for every sentence and speaker, during the testing phase. In the testing phase, the speakers could move freely within, and in and out of an open hall where the mobile handsets were kept, as multiple speakers made simultaneous calls to the voice-server for marking their attendances. The duration of three sentences of IITG database vary between 3-5 s over all the speakers and sessions. The three sentences are :

TD-1: *"Don't ask me to walk like that"*.

TD-2: *"Lovely pictures can only be drawn"*.

TD-3: *"Get into the hole of tunnels"*.

As this work is a continuum of our previous work on speaker verification [42], it may be interesting to test our proposed features on the same database that was used in [42]. In other words, the proposed HS features may be tested on a TISV system, for a wider and fairer comparison with the cepstral features already explored in [42]. Therefore, the NIST 2003 corpus [152] is also used in this work. The NIST corpus consists of 356 speakers, for which 356 speech utterances, each of ~2 mins duration are provided for training. Out of the 356 speakers, 144 are male, and 122 female. For testing, 2559 random speech utterances whose durations vary between ~14 s to ~44 s, are provided.

4.4.1 TDSV and TISV System Setup

While sophisticated statistical techniques such as the *i-vector* [12] and HiLAM [14] exist for implementing speaker verification systems, they require large quantities of data, which in the case of low-cost and practical implementation, is not a luxury. DTW has the advantage that it does not require large volumes of data to build the system. Hence, DTW continues to be used in many speech processing applications including speaker recognition, speech recognition and emotion recognition [156–161]. As this work is focussed on speaker verification under limited data condition, we utilize DTW in this work to implement the TDSV systems on the RSR2015 and IITG databases.

For the NIST 2003 database, the state-of-the-art *i-vector* system is used to build the TISV system, as in [42]. Though this work focuses on the TDSV task, it is essential to test the proposed HS features on the same set-up that has been used in [42]. Thus, a TISV system is built on the NIST 2003 corpus under sufficient data conditions.

4.4.1.1 DTW Based TDSV Setup for the RSR2015 and IITG Databases

The RSR2015 database documentation [14] divides the database into three sets of speakers : Set-I or Background set (50 males, 47 females), Set-II or Development set (50 males, 47 females), and Set-III or Evaluation set (57 males, 49 females). In [14], the standard *i-vector* system [12] has been implemented separately on Set-II and Set-III for speaker verification. For building the *i-vector* system, many tens of hours of speech data from the NIST 2004, 2005 and 2006 databases, the Switchboard corpus, and all utterances corresponding to the 30 sentences of Set-I were used. The performances were evaluated separately on Set-II and Set-III, by splitting each of them into Train and Test data sets. Similarly, the HiLAM system has also been implemented separately on Set-II and Set-III of the database [14]. As discussed in the Introduction section, such a luxury of data may not be available or economically viable to all (for example in developing countries), and hence not considered in this work. Also, one may argue that using such enormous volumes of data defies the principal objective of a TDSV system, which is to simplify speaker recognition by reducing the lexical/ sentence variability and data requirement. Further, it would be almost computationally impossible to perform extensive sets of experiments using such large databases.

Unlike sophisticated techniques, DTW is not *data hungry*, and can be implemented quickly, which makes it a suitable choice for performing large sets of experiments. The DTW algorithm finds an *optimal warping path* between two feature sets, based on *Euclidean distance* between feature vectors. The final DTW distance or score is the accumulation of Euclidean distances over the *optimal warping path*. The DTW score obtained by matching two feature sets corresponding to two different speech signals/ utterances reflects the similarity between the two signals. A lesser score indicates higher similarity between the two utterances, and vice versa. Thus, for a fixed phrase/ sentence, if the two utterances belong to the same speaker, a low DTW score is expected, whereas a higher score is expected when the utterances belong to different speakers.

For the RSR2015 database, as prescribed in [14], for each sentence in question (TD-1/TD-2/TD-3), out of the nine utterances/ sessions (for every speaker), three are used for training. Sessions 1, 4 and 7 are the training sessions. The remaining six utterances/ sessions are used for testing. While the *i-vector* and HiLAM frameworks were implemented on only a subset (Set-II/ Set-III) of the RSR2015 database, DTW scores are evaluated on all the 300 speakers of the data set.

For the IITG database, the three utterances (for every speaker) recorded in the training phase are used for training. For testing, the number of utterances available for a particular sentence/ phrase, uttered by a particular speaker, vary randomly. On an average, 30 utterances are available for every sentence, for every speaker, for testing. All 325 speakers are used to evaluate the performance of the system.

Testing process : Using DTW, every train utterance of a given speaker, Q, is matched with all testing utterances of 15 speakers. Only utterances of the same gender are matched against one another. Out of 15 speakers, one is the same speaker, Q, providing the true/ genuine claims. The other 14 speakers are imposters, and provide the false/ imposter claims. Out of these 14 speakers, four speakers are the closest to Q in the feature space defined by the 39 dimensional MFCCs. The other 10 speakers are randomly chosen. We refer the speakers which are similar to Q as its *adjacent speakers* [51]. The four *adjacent speakers* are included to test the capability of system to discriminate speakers with similar speech characteristics. To obtain the *adjacent speakers* for a particular speaker, in this work, the following procedure is used :

Step 1 : Extract 39 dimensional MFCCs from each of the three training utterances of all the speakers. The process of extracting the MFCCs is discussed in the next section. Thus, three feature sets of 39 dimensional MFCCs are obtained for every speaker.

Step 2 : Match the three feature sets of every speaker against one another using DTW. This results in three DTW scores for every speaker. Compute the average of three scores for every speaker. Thus, every speaker has a single average DTW score obtained by intra-matching his/ her training utterances.

Step 3 : For every speaker, find four average DTW scores which are closest to its average DTW score. The corresponding four speakers are considered its *adjacent speakers*.

The motivation behind the above testing approach is based on observations made during the creation of IITG database in building the students attendance system. It was observed that mismatches in the DTW based attendance system occurred generally amongst a specific set of speakers. These speakers, inevitably, produced similar sounding utterances of similar duration. Now, let us assume two random speakers, A and B, uttering the same sentence. Let $s_1^A(n)$ and $s_2^A(n)$ be two utterances of A. Similarly, let $s_1^B(n)$ and $s_2^B(n)$ be two utterances of B. If A and B are really close in the MFCCs feature space, matching $s_1^A(n)$ with $s_1^B(n)$ or $s_2^B(n)$ could result in a DTW score which is similar to the DTW score obtained by matching $s_1^A(n)$ with $s_2^A(n)$. Similarly, matching $s_1^B(n)$ with $s_1^A(n)$ or $s_2^A(n)$ could result in a DTW score which is similar to the DTW score obtained by matching $s_1^B(n)$ with $s_2^B(n)$. Therefore, the DTW score obtained by matching $s_1^A(n)$ with $s_2^A(n)$, and that obtained by matching $s_1^B(n)$ with $s_2^B(n)$, could be similar if A and B are closely placed in the feature space. In fact, during the creation of IITG database, it was observed that the speakers which were generally mismatched by the system produced similar DTW scores when intra-matching their utterances.

The observations made during the creation of IITG database show that including the *adjacent speakers* of every speaker during testing is required to properly evaluate the capability of system to discriminate speakers. Matching the utterances of a given speaker with utterances of hundreds of random speakers may not actually reflect the true capability of the system. However, the method we have utilized to determine the *adjacent speakers* may not be the only way or the best way. It is purely based on observations made during the creation of IITG database. Nevertheless, using the aforementioned testing process, in this work, the DTW scores obtained are used to evaluate the performance of TDSV system.

4.4.1.2 *i*-vector Based TISV Setup for the NIST 2003 Database

The implementation of *i-vector* system generally requires four data sets - *Universal Background Model* (UBM) data set, *Development* data set, *Train* data set, and *Test* data set. From the *Development* data set, the *total variability matrix*, T_v, is learned, which is a representation of all possible session and channel variabilities that may be encountered in the system. The UBM data set is used to build a GMM model of an universal or average speaker. Let S_{UBM} represent the mean supervector of the UBM data set, obtained by building a GMM on the combined feature set of all its utterances. Let S_u represent the *mean supervector* of any given utterance, obtained from GMM of its feature set. Then, in the *i-vector* framework, the two supervectors are related by a low-dimensional vector, called the *i-vector* [12], which is representative of the speaker corresponding to the utterance.

$$S_u = S_{\text{UBM}} + T_v i_u \ , \tag{4.34}$$

The *i-vector* so generated from a given utterance does contain session and channel variabilities [12,162]. To curb the effects of the data conditions, the *i-vector* is processed by *Linear Discriminant Analysis* (LDA), followed by *Within Class Covariance Normalization* (WCCN) [163]. The LDA and WCCN matrices are learned from the *i-vectors* extracted from the utterances of the Development data set. Let $\hat{i}_{tr}$ and $\hat{i}_{ts}$ represent two processed *i-vectors*, corresponding to two utterances (one from the Train data set, other from the Test data set). The verification is, then, performed based on the *cosine kernel score*, given by,

$$i_s = \frac{< \hat{i}_{tr}, \hat{i}_{ts} >}{\sqrt{||\hat{i}_{tr}||_2 \ ||\hat{i}_{ts}||_2}} \tag{4.35}$$

For the NIST 2003 corpus, the same set-up used in [42] is utilized in this work. In this case, we are not concerned about the scarcity of training data or the practicality of implementation. Our objective is to merely evaluate the performances of the proposed HS features in comparison to the cG_K feature, in the TISV task. Hence, the Switchboard Corpus II data set [164], amounting to about 10 hours of speech data, is used as the Development data set. A subset of the Development data set is taken as the UBM data set. A 512 mixture GMM is used to model the UBM data set. The 356 training utterances of NIST 2003 corpus are used as the Train data set, and the 2559 testing utterances are used as the Test data set. The Test data set is also corrupted with varying degrees of Babble noise [110] as in the case of the RSR2015 database to evaluate the robustness of features.

The experimental setup for the *i-vector* framework may be summarized as,

- A gender independent GMM of 512 mixtures built on the UBM data set.
- A T_v matrix of 400 columns obtained from the Development data set.
- 200 dimensional LDA matrix and 200 dimensional WCCN matrix using the Development data set *i-vectors*.

Based on the *cosine kernel scores* obtained by matching training utterances with testing utterances (according to the key-list specified in [152]) the performance of TISV system is evaluated.

4.4.2 Feature Extraction

In this work, any feature set extracted from a speech utterance/ signal is composed of feature vectors extracted from frames of 20 ms duration. Adjacent frames have a 50 % overlap between them, i.e., the frame shift is 10 ms. In the case of TDSV setup, energy based EPD is performed prior to constructing the feature set. The *silence frames* at the beginning and end of the speech utterance are eliminated by EPD. In the case of TISV setup, instead of EPD, energy based *Voice Activity Detection* (VAD) is done on the speech signal. VAD eliminates all the *silence* frames from the analysis, not only those at the terminating ends of the utterance. As DTW matches both the features and temporal sequences of features, EPD is used for TDSV. As all the *i-vectors* have a fixed length, they do not represent the temporal information of speech utterance, and hence VAD is used for TISV. After the selection of frames by EPD/ VAD, the feature vectors corresponding to them are stacked. This forms the feature set for the speech signal/ utterance. Finally, for every feature dimension of the feature set, *mean subtraction and variance normalization* is performed. Figure 4.5 shows the entire feature extraction process for the TDSV setup.

In this work, the seven different feature sets, obtained from the HS of a given speech utterance, are used individually, and in combination with the feature set corresponding to 39 dimensional MFCCs. The dimensions of each of the seven feature sets are varied, by varying the number of IMFs from $K = 10 - 2$, in steps of 2. This is done in order to find out the set of IMFs which is useful for the TDSV task. For extracting the MFCCs from a speech frame, the speech frame is first preemphasized, and then a 22 filter Mel filterbank is applied on its DFT spectrum. Then, a standard 39 dimensional MFCCs feature vector, and an Extended 51 dimensional MFCCs (Ext. MFCCs) feature vector, are obtained. The standard 39 dimensional MFCCs feature vector consists of the first 13 cepstral (excluding the 0^{th} coefficient), the first 13 Δ cepstral, and the first 13 $\Delta\Delta$ cepstral coefficients. The Ext. MFCCs feature vector consists of the first 17 cepstral (excluding the 0^{th} coefficient), the first 17 Δ cepstral, and the first 17 $\Delta\Delta$ cepstral coefficients. The MFCCs and the Ext. MFCCs are extracted using the VOICEBOX toolbox [165]. Comparing the performances of the Ext. MFCCs with that of 39 dimensional MFCCs enables us to observe how useful the higher dimensions of the MFCCs are in the TDSV system. Also, it enables us to ascertain whether the combinations of seven HS features with the 39 dimensional MFCCs are truly useful.

Apart from the aforementioned features, the cG_K feature set is also obtained from the speech signal, using its raw IMFs, as described in [42]. The two other features explored in [42] are not extracted, as they provide similar performances as the cG_K feature. Using the cG_K feature in our experiments enables us to observe how effective the proposed HS features are in comparison with the energy or cepstral features derived from the raw IMFs. In other words, we will be able to investigate whether the HS is really useful for characterizing speakers, or the raw IMFs are better off.

In the case of RSR2015 database, the testing utterances are also corrupted with varying degrees of Babble noise [110], prior to extracting features. This enables us to evaluate the robustness of features in the TDSV setup. The Test data set of NIST 2003 corpus is also corrupted by Babble noise, to evaluate the robustness of features in the TISV setup. The testing utterances of the IITG database already have telephone channel effects, environmental noise, and noise from nearby speakers. Hence, no artificial noise is incorporated in this case for the assessment of robustness of features.

4.4.3 Performance Metrics

The standard metrics - EER and mDCF) - are used to measure the performance of TDSV and TISV systems [42]. Let S$c = \{sc_1, sc_2, ..., sc_S\}$ represent the entire set of DTW scores/cosine kernel scores obtained by verifying the claims of the test utterances (in the Test data set) against the train utterances (in the Train data set). The set Sc is normalized so that $\{sc_i \in [0,1] \mid i = 1,2,...,S\}$. Let $\xi^s \in [0,1]$ be the decision threshold, below (for DTW) or above (for *i-vector*) which the claim of test utterance against the train utterance is considered valid, i.e, both the utterances are considered to belong to the same speaker. Let S_G denote the number of genuine/ true claims that have been accepted. Let S_I denote the number of imposter/ false claims that have been rejected. Let S_M denote the number of genuine/ true claims that have been rejected. Let S_F denote the number of imposter/ false claims that have been accepted. We have, $S = S_G + S_I + S_M + S_F$. For any given threshold, ξ^s, the *Probability of Miss* is given as $P_M = \frac{S_M}{S_G+S_M}$. The *Probability of False Alarm* is given as $P_F = \frac{S_F}{S_I+S_F}$. We may now define EER and DCF as follows :

Equal Error Rate (EER) : At a particular threshold, $\xi^s = \xi^s_0$, $P_M = P_F$. Then, EER $= P_M \times 100$ % $= P_F \times 100$ %.

minimum Detection Cost Function (*m*DCF) : Two parameters, C_M and C_F, assign costs to the event of a *miss* (a genuine claim rejected) and that of a *false alarm* (an imposter claim is accepted), respectively. Also, an *a priori* probability, P_T, is assigned, which assumes that out of all the test claims, only a fraction P_T are genuine claims. For any given threshold, ξ^s, a cost parameter is then evaluated, given by,

$$C_{\xi^s} = C_M \times P_M \times P_T + C_F \times P_F \times (1 - P_T) \tag{4.36}$$

The DCF is then given by,

$$\text{DCF} = \min_{\xi^s \in [0,1]} C_{\xi^s} \tag{4.37}$$

In this work, $C_M = 10$, $C_F = 1$, and $P_T = 0.01$ are considered, for all the three databases.

4.5 Experimental Results and Analysis

In this section, the experimental results and their analysis, for the TDSV systems implemented separately on RSR2015 and IITG databases, and the TISV system implemented on the NIST 2003 corpus. The performances of seven experimental HS features, the MFCCs, and their combinations, are evaluated. The performances of Ext. MFCCs and cG_K features are also evaluated.

Table 4.1: Performances of 39 dimensional MFCCs, 51 dimensional Ext. MFCCs, cG_K feature and the proposed seven features using all the IMFs ($K = 10$). The performance metrics are evaluated for the TD-1, TD-2 and TD-3 sentences of RSR2015 database.

Technique →		Mel filterbank		MEMD filterbank	MEMD - Hilbert Spectrum						
Feature →		MFCCs (39)	Ext. MFCCs (51)	cG_{10}	F_{10}	E_{10}	σF_9	σE_9	ΔF_{10}	ΔE_{10}	Υ_{10}
Sentence	Metric										
TD-1	EER (%)	7.59	6.22	27.71	15.48	12.57	11.65	13.14	22.61	19.33	16.22
	DCF	0.0388	0.0378	0.0555	0.0454	0.0428	0.0421	0.0433	0.0513	0.0484	0.046
TD-2	EER (%)	6.29	6.4	23.16	13.33	16.75	9.02	8.85	12.63	13.43	11.84
	DCF	0.0421	0.0423	0.0546	0.0474	0.0497	0.044	0.0439	0.0469	0.0472	0.0463
TD-3	EER (%)	6.6	6.25	17.5	11.4	8.18	15.17	15.93	7.83	9.9	10.19
	DCF	0.0422	0.0418	0.058	0.0492	0.0444	0.0544	0.0555	0.0441	0.0469	0.0475

4.5.1 Performances of the TDSV System Implemented on the RSR2015 Database

Table 4.1 presents the performance metrics for each of the seven experimental HS features, derived using all ten IMFs ($K = 10$). For comparison, the performance metrics for the MFCCs, Ext. MFCCs and cG_K are also shown. For the TD-1 sentence, the 39 dimensional MFCCs provide an EER of 7.59 %. The 51 dimensional Ext. MFCCs show improvement, with a EER of 6.22 %. For the TD-2 and TD-3 sentences, the EERs of the MFCCs and Ext. MFCCs are much closer. For the MFCCs, the average EER over the three sentences is 6.83 %. For the Ext. MFCCs, the average EER is 6.22 %. In the RSR2015 database documentation [14], using the *i-vector* framework, an EER of $\sim$ 5.32 % (4.03 % for male, 6.61 % for female) is reported, for all 30 sentences combined, on Set-III of the database. For the HiLAM system, even better performances are reported. Therefore, one may be critical in using a simple process such as DTW [3, 4, 94] for our task. However, it must be noted that the performances in this work have been reported on the whole database, and not a much smaller subset. It is to be expected that increasing the number of speakers would bring the performance of *i-vector* and HiLAM systems much closer to that of DTW based system. More importantly, as mentioned earlier, DTW is independent of the requirements of large amounts of data that are associated with statistical methods, which is aligned with the objective of our work.

Ideally, one would assume that, in the feature space, a speaker would be confined to a small region (low intra-speaker variability). As, in a TDSV system, the sentence/ phrase is fixed, speech variability is significantly curbed, and hence, such an assumption should be more valid. Also, one would expect that if the speaker information is adequately captured by the features, the performance of TDSV systems for different sentences would be similar or comparable. If there are a large number of speakers in the feature space, the average separation between any two speakers is expected to be low, and the likelihood of the system confusing a particular speaker with one or more of its *adjacent speakers* is more. As such, having more speakers for testing the system should be more challenging, rather than utilizing a large number of sentences for evaluation. That is why, in this work, we have utilized only three sentences of RSR2015 database, but used all its 300 speakers for evaluation. Because the performances of *i-vector* and HiLAM systems are reported for only 106 speakers, we believe their performances may be worse if all 300 speakers are used for evaluation.

Again, as we have discussed earlier, for any given speaker, there is generally a limited set of *adjacent speakers* which can cause mismatch in the system. Therefore, if a large number of imposters are used to match against a given speaker, the P_F (*Probability of False Alarm*) will be quite low. On the contrary, since any given speaker has only six testing utterances for true/ genuine claims, a single mismatch significantly increases the value of P_M (*Probability of Miss*). That is why, unlike in the case of RSR2015 documentation, where for any given speaker (there are 106 speakers), 105 imposters are considered, in this work, we have considered only 14 imposters for any given speaker. As mentioned earlier, out of the 14 imposters, four are *adjacent speakers*. This setup, in our opinion, provides a more balanced evaluation, apart from making the process fast. Even in the case of performance evaluation on the NIST 2003 database, according to the key list provided, every speaker is tested against only a small set of imposters and not all the other speakers of database [152]. It may, however, be noted that there is nothing wrong in utilizing a large set of imposters, as prescribed in the RSR2015 documentation [14]. Adhering strictly to whatever protocol one has chosen is more critical for analysis.

As is observed from Table 4.1, for all the three sentences, as standalone features the MFCCs and Ext. MFCCs are far better than the cG_K and HS features. Thus, whether it be the raw IMFs, or the HS, on their own they are limited in characterizing speakers. As such, we proceed towards combining them with the MFCCs, and evaluating the effect of the combinations on the TDSV system. Table 4.2 presents the performances of feature combinations for the TD-1 sentence. For the seven HS features and cG_K, the dimensions are reduced by changing K from 10 to 2, in steps of 2. In other words, the number of IMFs from which the features are extracted are varied from the first ten (all the IMFs), to the first two only. As can be observed from the table, in general, the best performances of combinations are obtained for $K \leq 4$. With respect to the MFCCs, absolute improvements in

Table 4.2: Performances of 39 dimensional MFCCs, 51 dimensional Ext. MFCCs, and combinations of 39 dimensional MFCCs with the cG_K feature and each of seven experimental features. The performance metrics are evaluated for TD-1 sentence of RSR2015 database. The dimensions of MEMD features are varied by changing K from 10 to 2.

Technique →		Mel filterbank		MEMD filterbank	MEMD - Hilbert Spectrum						
Feature →		MFCCs (39)	Ext. MFCCs (51)	MFCCs + cG_K	MFCCs + F_K	MFCCs + E_K	MFCCs + σF_{K-1}	MFCCs + σE_{K-1}	MFCCs + ΔF_K	MFCCs + ΔE_K	MFCCs + Υ_K
K	Metric										
10	EER (%)	7.59	6.22	8.65	6.48	8.17	9.36	8.82	8.41	8.4	7.11
	DCF	0.0388	0.0378	0.0398	0.038	0.0392	0.0402	0.0397	0.0396	0.0394	0.0385
8	EER (%)	–	–	7.34	6.54	7.07	9.7	8.54	7.58	7.39	7.41
	DCF	–	–	0.0387	0.038	0.0383	0.0404	0.0395	0.0389	0.0385	0.0388
6	EER (%)	–	–	6.96	**6.54**	7.56	8.27	7.02	6.95	7.04	7.21
	DCF	–	–	0.0384	**0.038**	0.0387	0.0393	0.0382	0.0384	0.0382	0.0386
4	EER (%)	–	–	**5.03**	6.57	**6.62**	6.91	**6.39**	7.28	**6.71**	**7.2**
	DCF	–	–	**0.0368**	0.0381	**0.0379**	0.0381	**0.0377**	0.0386	**0.038**	**0.0386**
2	EER (%)	–	–	6.66	6.96	6.93	**6.85**	7.7	**6.75**	7.42	7.27
	DCF	–	–	0.0381	0.0384	0.0382	**0.0381**	0.0388	**0.0382**	0.0386	0.0386

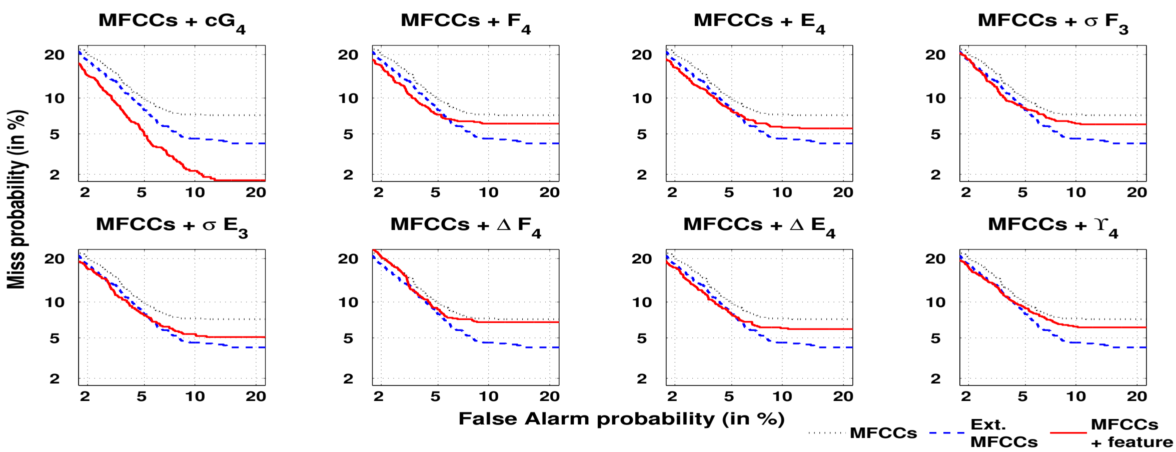


Figure 4.6: DET curves of the 39-dimensional MFCCs, 51-dimensional Ext. MFCCs, and the combinations of the 39-dimensional MFCCs with the cG_K feature and each of the seven experimental features. The DET curves are plotted for the TD-1 sentence of the RSR2015 database. The dimensions of the cG_K and seven HS features are fixed at $K = 4$.

EER of around $0.5 - 1.0$ % are obtained for the different HS features. The cG_K feature performs remarkably with an improvement in EER of ~ 1.5 % at $K = 4$. The additional 12 dimensions of Ext. MFCCs also significantly improve the EER. Similar trends are observed for DCF. Figure 4.6 plots the *Detection Error Trade off* (DET) curves for different feature combinations, considering $K = 4$. As can be seen from the figure, the MFCCs + cG_4 combination outperforms the MFCCs and Ext. MFCCs. The MFCCs + HS combinations outperform the MFCCs for the entire range. However, they are not always competitive with the Ext. MFCCs.

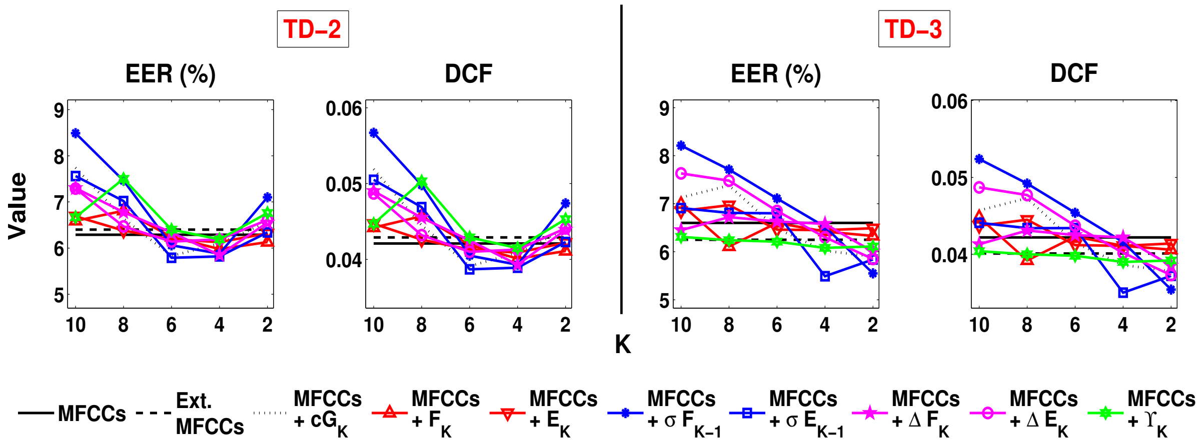


Figure 4.7: Performances of the 39-dimensional MFCCs, 51-dimensional Ext. MFCCs, and the combinations of the 39-dimensional MFCCs with the cG_K feature and each of the seven experimental features. The performance metrics are evaluated for the TD-2 and TD-3 sentences of the RSR2015 database. The dimensions of the cG_K and seven HS features are varied by changing K from 10 to 2.

Figure 4.7 presents the performance metrics of the feature combinations for TD-2 and TD-3 sentences. In the case of TD-2 sentence, the Ext. MFCCs *marginally degrades* the performance, whereas in the case of TD-3 sentence, it marginally improves the performance. The MFCCs + cG_K and MFCCs + HS features, again, obtain their best performances around $K = 4$.

The above observations clearly indicate that the HS features and cG_K can *consistently complement* the MFCCs, *under clean conditions*. The additional 12 dimensions of the Ext. MFCCs do not always improve the performance of the system. The improved performances of the feature combinations under low dimensions clearly indicate that only the first few IMFs are useful in characterizing speakers, as was observed in [42].

4.5.1.1 Canonical Correlation Analysis of the Features

Canonical Correlation Analysis (CCA) [166] is a useful tool used for comparison of features of different dimensions. Consider $X_{\mathrm{D1}}^{N_f} = [X_1^{N_f} X_2^{N_f} \ldots X_{\mathrm{D1}}^{N_f}]$ to be a feature set (matrix) of $N_f \times \mathrm{D1}$ size. The rows correspond to feature vectors of dimension D1, obtained from N_f speech frames. Assume $Y_{\mathrm{D2}} = [Y_1^{N_f} Y_2^{N_f} \ldots Y_{\mathrm{D2}}^{N_f}]$ to be another feature set, but of dimension D2, obtained from the same speech frames. It is a matrix of $N_f \times \mathrm{D2}$ size. The *first Canonical variate pair*, (U_1, V_1), for the two features, is given by,

$$U_1 = \sum_{d=1}^{\mathrm{D1}} u_d X_d^{N_f} \ , \ V_1 = \sum_{d=1}^{\mathrm{D2}} v_d Y_d^{N_f} \tag{4.38}$$

Thus, U_1 and V_1 are two vectors of length N_f. The variables $\{u_d,\ d = 1, 2, ...\mathrm{D1}\}$ and $\{v_d,\ d = 1, 2, ...\mathrm{D2}\}$ are optimized so as to maximize the *correlation coefficient* between U_1 and V_1. Optimal values of u_d and v_d result in the canonical variate pair. The correlation coefficient between U_1 and V_1 is called the *Canonical correlation coefficient* of the first canonical variate pair [166].

Table 4.3: Canonical correlation coefficients of various feature-pairs. 39-dimensional MFCCs, cG_K and the seven HS features using $K = 4$ are considered. The coefficients are evaluated for the TD-1, TD-2 and TD-3 sentences, separately, of the RSR2015 database. The testing utterances are used for evaluating the coefficients.

Sentence	Feature	MFCCs (39)	cG_4	F_4	E_4	σF_3	σE_3	ΔF_4	ΔE_4	Υ_4
TD-1	MFCCs (39)	1	0.9	0.85	0.83	**0.38**	0.7	0.79	0.78	0.87
	cG_4	0.9	1	0.81	0.92	**0.38**	0.75	0.82	0.84	0.85
	σF_3	0.38	0.38	0.61	0.42	1	**0.21**	0.45	0.37	0.43
TD-2	MFCCs (39)	1	0.85	0.83	0.84	**0.47**	0.66	0.79	0.81	0.86
	cG_4	0.85	1	0.75	0.93	**0.44**	0.68	0.8	0.88	0.8
	σF_3	0.47	0.44	0.59	0.44	1	**0.28**	0.42	0.43	0.5
TD-3	MFCCs (39)	1	0.85	0.84	0.84	**0.46**	0.72	0.78	0.8	0.87
	cG_4	0.85	1	0.78	0.94	**0.43**	0.78	0.77	0.87	0.83
	σF_3	0.46	0.43	0.58	0.43	1	**0.23**	0.45	0.4	0.48

In order to quantify how different the proposed HS features are with respect to the MFCCs and cG_K feature, we perform CCA for the various feature pairs. The CCA is performed for the features at their best performing dimensions ($K = 4$), using the testing utterances. Table 4.3 presents the Canonical correlations for the first canonical variate pairs corresponding to the feature pairs. As the MFCCs and Ext. MFCCs differ only in the number of dimensions, only one of them (MFCCs) has been used for the analysis.

Considering the TD-1 sentence, one may observe (first row) that there is a very strong similarity between the MFCCs and the cG_4 feature. In spite of this, the MFCCs + cG_4 combination enhances the performance of the system credibly, as already discussed. Both the features are cepstral features, yet cG_4 captures speaker-specific information in a different manner than the MFCCs, and hence complements it [42]. Compared to cG_4, the HS features (particularly σF_3 and σE_3) have less correlation with the MFCCs. However, though their combinations with the MFCCs result in enhanced performance, their combinations are still inferior to the MFCCs + cG_4 combination. This shows that while CCA is an effective tool for investigating complementary features, it does not guarantee the usefulness of combinations of features. Correlation between two random feature sets may be low, yet their combination may not be useful, and vice versa. CCA is an aid, not a complete solution. Considering the second row of the table, one may observe that the cG_4 feature has lesser similarity with the HS features (barring E_4) than with the MFCCs. Again, considering the third row, one may observe that σF_3 has little similarity with all the other features.

The observations for TD-2 and TD-3 sentences are consistent with the observations for TD-1 sentence. This suggests that the MFCCs, cG_4, and the seven HS features manifest different kinds of information, or they present the same information in a different manner. Therefore, there should be benefits in combining the MFCCs, the cepstral features extracted from the raw IMFs, and the HS features. It may be possible to eliminate the higher dimensions of MFCCs considerably, and yet achieve greater performance by using an optimum combination of various low-dimension features. How to achieve such a combination lies in the realm of feature combination and machine learning methodologies. This, however, is beyond the realm of this already exhaustive work, and may be taken up in the future. Such a work may utilize more sophisticated HiLAM and *i-vector* frameworks as well.

4.5.1.2 Performances under Babble Noise

Table 4.4: Performances of the 39 dimensional MFCCs, 51 dimensional Ext. MFCCs, and the combinations of the 39 dimensional MFCCs with the cG_K feature and each of seven experimental features. The performance metrics are evaluated for the TD-1 sentence of the RSR2015 database. The testing utterances are corrupted by Babble noise, with SNR varying from 20-0 dB. The dimensions of MEMD features are kept constant, with $K = 4$.

Technique →		Mel filterbank		MEMD filterbank	MEMD - Hilbert Spectrum						
Feature →		MFCCs	Ext. MFCCs	MFCCs	MFCCs	MFCCs	MFCCs	MFCCs	MFCCs	MFCCs	MFCCs
SNR	Metric	(39)	(51)	+ cG_4	+ F_4	+ E_4	+ σF_3	+ σE_3	+ ΔF_4	+ ΔE_4	+ Υ_4
20 dB	EER (%)	8.1	11.91	6.55	6.71	6.67	6.98	**6.48**	7.34	7.4	7.84
	DCF	0.0394	0.0425	0.038	0.0382	0.0379	0.0382	**0.0378**	0.0387	0.0385	0.0391
15 dB	EER (%)	9.26	12.47	11.13	7.77	**6.95**	7.64	7	7.37	7.56	8.06
	DCF	0.0404	0.0429	0.0418	0.039	**0.0382**	0.0387	0.0382	0.0387	0.0387	0.0393
10 dB	EER (%)	11.23	12.9	13.22	13.23	10.31	10.27	10.29	**9.89**	10.34	10.48
	DCF	0.042	0.0433	0.0436	0.0436	0.0409	0.0409	0.0409	**0.0408**	0.041	0.0413
5 dB	EER (%)	12.51	20.93	14.18	14.35	12.6	**12.37**	12.59	13.31	13.51	13.14
	DCF	0.0431	0.0499	0.0443	0.0445	0.0428	**0.0427**	0.0428	0.0436	0.0436	0.0435
0 dB	EER (%)	20.29	23.64	21.04	21.6	20.42	20.76	**20.1**	22.17	20.68	22.03
	DCF	0.0493	0.0522	0.05	0.0505	0.0493	0.0496	**0.049**	0.051	0.0495	0.0508

Having evaluated the performances of features and their combinations under clean conditions,

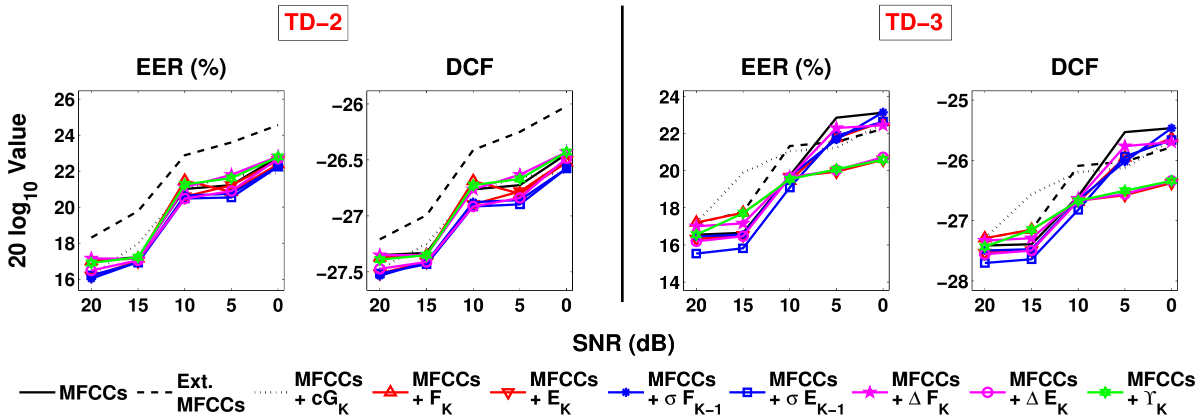


Figure 4.8: Performances of 39 dimensional MFCCs, 51 dimensional Ext. MFCCs, and the combinations of the 39 dimensional MFCCs with the cG_K feature and each of the seven experimental features. The performance metrics are evaluated for the TD-2 and TD-3 sentences of RSR2015 database. The testing utterances are corrupted by Babble noise, with SNR varying from 20-0 dB. The dimensions of the cG_K seven features are kept constant, with $K = 4$.

we now wish to evaluate how robust the features are to external interference. For this purpose, the testing utterances are corrupted with Babble noise prior to extracting feature sets from them. Table 4.4 presents the performances of feature combinations for TD-1 sentence. For clean speech, as the best performances were observed around $K = 4$, the dimensions of cG_K feature and HS features are fixed by setting $K = 4$. As can be seen in the table, the performance of TDSV system degrades starkly with the increase in noise levels. The Ext. MFCCs degrade the performance of system with respect to the MFCCs. The MFCCs + cG_4 also degrade the performance of the system, except in the case of SNR = 20 dB. The MFCCs + HS features, however, consistently improve the performance of system. Except the MFCCs + F_4, all the other MFCCs + HS features improve the performance of system credibly for SNR $\geq$ 10 dB. At extreme noise levels, the MFCCs + σF_3 and MFCCs + σE_3 are still able to obtain marginal improvements, while all other combinations fail.

Figure 4.8 presents the performance metrics for the feature combinations, for the TD-2 and TD-3 sentences. The performances have been plotted in the log-scale for better visualization. As can be observed in the figure, similar to the observations made in Table 4.4, the MFCCs + HS features provide significant improvement with respect to the MFCCs. In fact, all the MFCCs + HS features, for these two sentences, provide improvement over the MFCCs. On the other hand, the Ext. MFCCs and MFCCs + cG_4 combination perform inconsistently.

4.5.1.2 Effect of the Rearrangement Process

Table 4.5: Performances of 39 dimensional MFCCs, 51 dimensional Ext. MFCCs, and combinations of 39 dimensional MFCCs with the cG_K feature and each of the seven experimental features. *The seven HS features are obtained without rearrangement.* The dimensions of cG_K and HS features are kept constant, with $K = 4$. The performance metrics are evaluated for the TD-1 sentence of RSR2015 database under clean and noisy testing conditions. *The cases where the HS features obtained without rearrangement outperform those obtained with rearrangement (Tables 4.2 and 4.4) are italicized.*

Technique →		Mel filterbank		MEMD filterbank	MEMD - Hilbert Spectrum						
Feature →		MFCCs (39)	Ext. MFCCs (51)	MFCCs + cG_4	MFCCs + F_4	MFCCs + E_4	MFCCs + σF_3	MFCCs + σE_3	MFCCs + ΔF_4	MFCCs + ΔE_4	MFCCs + Υ_4
SNR	Metric										
Clean	EER (%)	7.59	6.22	5.03	6.71	6.67	***6.88***	6.83	7.34	7.4	***6.84***
	DCF	0.0388	0.0378	0.0368	0.0382	0.0379	0.0381	0.0381	0.0387	0.0385	***0.0383***
20 dB	EER (%)	8.1	11.91	6.55	6.87	7.34	9.76	9.84	8.95	7.91	***7.57***
	DCF	0.0394	0.0425	0.038	0.0383	0.0385	0.0405	0.0406	0.04	0.039	***0.0389***
15 dB	EER (%)	9.26	12.47	11.13	11.11	11.81	13.74	12.45	11.49	11.16	11.97
	DCF	0.0404	0.0429	0.0418	0.0418	0.0422	0.0438	0.0427	0.0421	0.0417	0.0425
10 dB	EER (%)	11.23	12.9	13.22	13.57	12.57	13.84	13.33	13.92	11.79	13.47
	DCF	0.042	0.0433	0.0436	0.0438	0.0428	0.0439	0.0434	0.0441	0.0422	0.0438
5 dB	EER (%)	11.51	20.93	14.18	***14.15***	13.44	15.13	15.88	14.76	13.93	14.05
	DCF	0.0422	0.0499	0.0443	***0.0443***	0.0435	0.0449	0.0456	0.0448	0.0439	0.0442
0 dB	EER (%)	20.29	23.64	21.04	***19.18***	***18.95***	22.92	21.45	***21.55***	***19.47***	***21.35***
	DCF	0.0493	0.0522	0.05	***0.0485***	***0.0481***	0.0514	0.0502	***0.0504***	***0.0485***	***0.0503***

As discussed in Section 4.3, a rearrangement process is employed during the extraction of HS features, with the hope of masking mode-mixing. Therefore, it may be interesting to evaluate the benefit of using the rearrangement process. For this purpose, the same seven HS features are extracted without rearranging the instantaneous frequencies and energies of the IMFs, and then subjected to the TDSV task. Table 4.5 presents the performances of MFCCs + HS features, obtained without rearrangement, for the TD-1 sentence. The cases where the MFCCs + HS features (without rearrangement) outperform the MFCCs + HS features (with rearrangement) are italicized.

Comparing Table 4.5 with Tables 4.2 and 4.4, one can clearly identify that the features obtained with rearrangement work better than the features obtained without rearrangement. In general, the features obtained with rearrangement show marginal improvements over the features obtained without rearrangement. However, at SNR = 15 dB, there is a significant gap in the performances of the two sets of features. In a few cases (italicized), the performances of features obtained without rearrangement are better. However, they are marginal improvements. Similar trends are observed for the TD-2 and TD-3 sentences, and hence not presented here. Thus, in general, the rearrangement process does not inhibit the performance of features even though it does not contribute to large scale benefits. In other words, the process of rearranging the instantaneous frequencies and energies may be considered beneficial, even though the gains are not extravagant.

The limited gains may be attributed to the significant reduction in mode-mixing by MEMD with respect to EMD. Had the IMFs been extracted using EMD, the rearrangement process may have been more beneficial. In fact, when we started the initial experiments for this work, we had utilized the IMFs obtained from EMD. In our preliminary experiments, using limited data sets and DTW, we had observed that the rearrangement process was consistently useful in the case of EMD. The framework developed for EMD was simply applied on MEMD in search of better results. As MEMD provided better results than EMD in our preliminary experiments, and since there is no merit in performing the same task using two similar methods (plus it would not have been computationally feasible), MEMD was chosen for the extensive set of experiments. Nevertheless, as we have discussed in Section 4.3, the rearrangement process is a sensible approach, and hence we continue to use it for extracting the HS features in this work.

4.5.2 Performances of the TDSV System Implemented on the IITG Database

Having observed the performances of the different feature combinations on clean speech, and speech corrupted by artificially inserted Babble noise, we now wish to test the features using speech utterances affected by telephone channel conditions, environmental noise, and interference from background speakers. For this purpose, the IITG database is considered.

Table 4.6: Performances of the 39-dimensional MFCCs, 51-dimensional Ext. MFCCs, and the combinations of the 39-dimensional MFCCs with the cG_K feature and each of the seven experimental features. The performance metrics are evaluated for the TD-1, TD-2 and TD-3 sentences of the IITG database. The dimensions of the MEMD features are kept constant, with $K = 4$.

Technique →		Mel filterbank		MEMD filterbank	MEMD - Hilbert Spectrum						
Feature →		MFCCs	Ext. MFCCs	MFCCs	MFCCs	MFCCs	MFCCs	MFCCs	MFCCs	MFCCs	MFCCs
Sentence	Metric	(39)	(51)	+ cG_4	+ F_4	+ E_4	+ σF_3	+ σE_3	+ ΔF_4	+ ΔE_4	+ Υ_4
TD-1	EER (%)	13.19	12.84	10.91	11.46	10.62	10.41	**10.25**	11.83	10.68	11.48
	DCF	0.0422	0.0412	0.035	0.0368	0.0339	0.0332	**0.0327**	0.0379	0.0341	0.0368
TD-2	EER (%)	11.37	11.65	11.56	11.3	10.33	10.71	**9.86**	11.1	10.86	10.38
	DCF	0.0404	0.0415	0.0412	0.0403	0.0366	0.038	**0.0349**	0.0395	0.0385	0.037
TD-3	EER (%)	18.42	17.93	17.96	19.26	17.86	17.48	17.35	**17.3**	17.9	17.93
	DCF	0.0526	0.0513	0.0514	0.0551	0.0509	0.0498	**0.0494**	0.0495	0.051	0.0513

Table 4.6 presents the performances of the MFCCs + HS features and the MFCCs + cG_K combination for each of the three sentences. Taking cue from the observations made for the RSR2015

database, the dimensions of HS features and cG_K are kept fixed by setting $K = 4$. In the table, one can observe that the Ext. MFCCs and MFCCs + cG_4 combination improve the performance of system (with respect to the MFCCs) for the TD-1 and TD-3 sentences, but degrade for the TD-2 sentence. The MFCCs + HS features, excepting the MFCCs + F_4 combination, however, consistently improve the performance of system, for all the three sentences. The consistent performances of MFCCs + HS features indicate that the HS may be a more reliable representation than the raw IMFs. Again, as the best performances are spread across different feature combinations, a fusion of the various low-dimensional features could be useful, as indicated earlier using CCA.

4.5.3 Performances of the TISV System Implemented on the NIST 2003 Corpus

Table 4.7: Performances of the 39 dimensional MFCCs, 51 dimensional Ext. MFCCs, and the combinations of the 39 dimensional MFCCs with the cG_K feature and each of the seven experimental features. The performance metrics are evaluated for TISV system implemented on the the NIST 2003 database. The testing utterances are corrupted by Babble noise, with SNR varying from 20-0 dB. The dimensions of the MEMD features are kept constant, with $K = 4$.

Technique →		Mel filterbank		MEMD filterbank	MEMD - Hilbert Spectrum						
Feature →		MFCCs	Ext. MFCCs	MFCCs	MFCCs	MFCCs	MFCCs	MFCCs	MFCCs	MFCCs	MFCCs
SNR	Metric	(39)	(51)	+ cG_4	+ F_4	+ E_4	+ σF_3	+ σE_3	+ ΔF_4	+ ΔE_4	+ Υ_4
clean	EER (%)	3.03	3.03	**2.48**	3.57	3.03	2.89	2.89	3.66	2.62	3.34
	DCF	0.0119	0.0137	0.0120	0.0163	0.0125	0.0117	**0.0115**	0.0174	0.0123	0.0140
20 dB	EER (%)	3.48	3.30	**3.03**	3.93	3.39	3.07	3.43	4.29	3.39	3.93
	DCF	**0.0144**	0.0163	0.0149	0.0180	0.0150	0.0145	0.0149	0.0194	0.0145	0.0178
15 dB	EER (%)	4.25	4.16	4.11	5.15	4.02	**3.97**	4.20	5.42	4.20	4.83
	DCF	0.0183	0.0205	0.0192	0.0224	0.0181	**0.0179**	0.0183	0.0235	0.0172	0.0223
10 dB	EER (%)	8.58	8.40	8.76	10.30	**7.68**	7.90	8.45	10.61	7.99	9.94
	DCF	0.0399	0.0399	0.0408	0.0458	**0.0373**	0.0378	0.0384	0.0482	0.0373	0.0443
5 dB	EER (%)	15.94	14.81	15.72	18.07	**14.18**	14.72	15.09	18.65	15.09	17.52
	DCF	0.0711	0.0719	0.0681	0.0744	0.0655	0.0666	0.0671	0.0787	**0.0645**	0.0753
0 dB	EER (%)	23.26	**22.81**	23.58	25.38	23.22	23.58	23.35	27.37	23.22	25.61
	DCF	0.0959	0.0958	0.0917	0.0956	0.0905	0.0914	0.0920	0.0988	**0.0880**	0.0951

In [42], we explored the utility of raw IMFs for TISV. In this work, until now, we have explored the utility of features extracted from the HS for TDSV, and compared it with that of cG_K feature extracted from the raw IMFs. To complete the analysis, we must subject the HS features to TISV task as well, using the same setup as was used in [42]. Hence, as mentioned in Section 4.4, the MFCCs + HS features are tested on the NIST 2003 corpus for TISV. Table 4.7 shows the performances of MFCCs, Ext. MFCCs, and the various feature combinations. $K = 4$ is used for the feature combinations. Under clean testing conditions, the performances of the MFCCs and Ext. MFCCs are not exactly the same as reported in [42]. There are minor differences. This is because, in this work the MFCCs and Ext. MFCCs are extracted using the VOICEBOX toolbox, whereas in [42] they were extracted using the HTK toolkit [167].

As may be observed in the table, all the feature combinations are less effective in the TISV task than they were in the TDSV task. The amount of data used to build the system is a principal reason. The MFCCs are more effective under sufficient data conditions. The significant reduction in lexical variability (session variability) in the TDSV task is another reason why the feature combinations are more effective in the TDSV task. When sufficient amount of data is available to build the system, the higher dimensions of the Ext. MFCCs become more useful, and hence the Ext. MFCCs provide consistent improvement (under both clean and noisy conditions). As is evident from the table, the MFCCs + cG_4 combination is less effective under noisy testing conditions. In multiple noisy cases, they show degradation in both EER and DCF.

Out of all the features, the MFCCs + E_4, MFCCs + σF_3, MFCCs + σE_3 and MFCCs + ΔE_4 combinations are the most consistent performers in terms of both EER and DCF. In terms of EER, the MFCCs + E_4 and MFCCs + ΔE_4 show improvements for all the cases. Similarly, the MFCCs + E_4 and MFCCs + ΔE_4 improve the EER for all cases except at SNR = 0 dB. In terms of DCF too, these four combinations improve the performance for almost all the cases. Only at SNR = 20 dB, they exhibit minute degradations in DCF. These observations, again, highlight the utility of HS in capturing speaker-specific information. Once again, different feature combinations provide the best performances for different conditions. Hence, as mentioned earlier, the fusion of various low-dimensional features might be more effective, with more sophisticated techniques, and may be considered as future work.

4.6 Summary and Conclusion

The objective of this work was to utilize the constituents of the HS, obtained from the IMFs of speech signal, for the task of TDSV, under practical scenarios. For this purpose, the constituents of HS - the instantaneous frequencies and energies of the IMFs - were processed in short time segments to obtain seven different features from them. These features were then tested for the TDSV task, considering three different sentences each of the RSR2015 and IITG databases. The DTW technique was used to implement the TDSV system, as it did not require large volumes of data. The features were tested on the RSR2015 database, not only under clean conditions, but also under Babble noise of varying strength. The testing conditions for the IITG database also provided high degree of mismatch from the training conditions.

The seven features when tested in a standalone fashion, were unable to compete with the MFCCs. However, in combination with the MFCCs, they were able to improve the performance of the TDSV system consistently, particularly at low dimensions, and under mismatched and noisy conditions. It was observed, using CCA, that the seven HS features had low correlation with the MFCCs, and even amongst them, which indicates that they capture different speaker characteristics. The MFCCs + HS features were also tested on the NIST 2003 corpus, for TISV, using sufficient data to build the system using the *i-vetor* framework. It was observed that the performance improvement in TISV was lesser than TDSV. Overall, it was observed that the HS features provide more consistent performances compared to the cepstral feature, cG_K, under noisy and mismatched conditions. In most cases, the best performances are spread across different feature combinations. Hence, as an extension of this work, the fusion of the various low dimensional features may be attempted using a more sophisticated framework than DTW, with the aim of reducing and replacing the higher dimensions of MFCCs.

5

Significance of Speech Specific Knowledge for End Point Detection

Contents

Objective

This chapter proposes a method using speech specific knowledge to detect the begin and end points of speech under degraded condition. The method is based on VLR detection and uses both excitation source and vocal tract system information. Existing method for VLR detection uses excitation source information. Vocal tract system information from DRF is used to eliminate spurious VLRs in background noise. FSS using excitation and vocal tract system information is carried out to remove spurious VLRs in the background speech region. Better localization of the end points is done using more detailed information about excitation source in terms of GAD to detect the sonorant consonants and missed VLRs. To include an unvoiced consonant, obstruent region detection is done at the beginning of first VLR and at the end of last VLR. Detected begin and end points are evaluated by comparing with manually marked end points as well as by conducting the TDSV experiments. The proposed method performs better than some of the existing techniques.

5.1 Introduction

The state-of-the-art SV systems provide good performance when the speech utterance is of high quality and free from any mismatch [107]. Such a speech utterance is treated as *clean speech* in the present work. However, in most practical operating conditions, the speech signal is affected by different degradation like background noise, background speech, sensor mismatch, channel mismatch, lip smacks, chirping noise, heavy breathing, and other environmental conditions, resulting in *degraded speech* that degrades the TDSV system performance significantly [168, 169]. There are numerous techniques available for dealing with training and testing speech conditions due to degradation [24,51]. To deal with such degraded speech signals, the compensation is done at the signal level, feature level, model level, score level or all of them together. In this chapter, we aimed at the signal level compensation methods used for removing the effect of noise from the degraded speech utterances. The main goal is to develop a robust and accurate begin and end points detector which can improve the performance of the SV system. The accurate begin and end points detection gives faster response and better matching, as useful frames are passed for further processing.

Accurate begin and end points detection is very crucial in TDSV task [49, 50, 94, 170]. In such systems, first begin and end points are detected and then feature vectors belonging to the speech region specified by the end points are considered for preparing the templates for training and testing. The testing template is compared with the reference templates using pattern matching technique like DTW [94]. The most widely used EPD method is based on the *voice activity detection* (VAD) [19, 20, 171], where voicing decision is made by considering the average energy as a threshold. Most of the existing end point detection methods uses energy based VAD [100, 101, 172]. Energy based techniques perform well for clean speech conditions and fail badly in degraded conditions [51]. Fig. 5.1 shows speech signal for the utterance "Get into the hole of tunnels" with non-overlapping background noise (in (a)) and its short term energy contour (in (b)). The background noise include both speech and non-speech background noise. The high energy in the background noise leads to incorrect detection of begin and end points. In [21, 173], pitch information is used for end point detection. However, pitch extraction is also a challenging task in degraded speech condition. In [22], a method for robust end point feature selection is proposed based on a self-adaptive approach. The periodicity based VAD perform better compared to the conventional energy based VAD [21]. The statistical model-based VAD was effectively developed based on the decision directed parameter estimation for the likelihood ratio test which is robust to detect the speech regions under degraded conditions [22]. A better SV system was developed under degraded speech conditions by using speech, only from the less affected speech regions [172]. This robustness is achieved by exploiting the knowledge of VLRs which are relatively high SNR regions [51]. Another method using VOP was proposed to detect begin and end points in [50]. Although this VOP based method is found to be better than the energy based methods, there exists several drawbacks in the method. This EPD is based on the VOP based method developed in [49]. After detecting the VOPs, begin and end points were detected using the knowledge of first and last VOPs. In some cases, there may be a consonant at the beginning of sentence and instead of detecting this consonant, begin point is marked 200 ms before the first VOP. Similarly, end point is marked 200 ms after the last VOP without estimating the length of the vowel. Therefore, the begin and end points detected in that way may not be correct and some explicit consonant region detection may be more useful. Moreover, instead of using the last VOP, *vowel end points* (VEPs) can be used. An algorithm for VLRs detection by detecting VLROPs and VLREPs was proposed in [51] and the method was further improved in [51] and [24]. VLRs include vowels, diphthongs and semivowels. Since, objective is to detect speech end points, instead of using VOP, VLR can be used as it will provide both VLROP and VLREP. Fig. 5.1 (a) shows the manually marked VLROPs and VLREPs along with the speech signal. Although the begin point is near to the first VLROP, the end point is

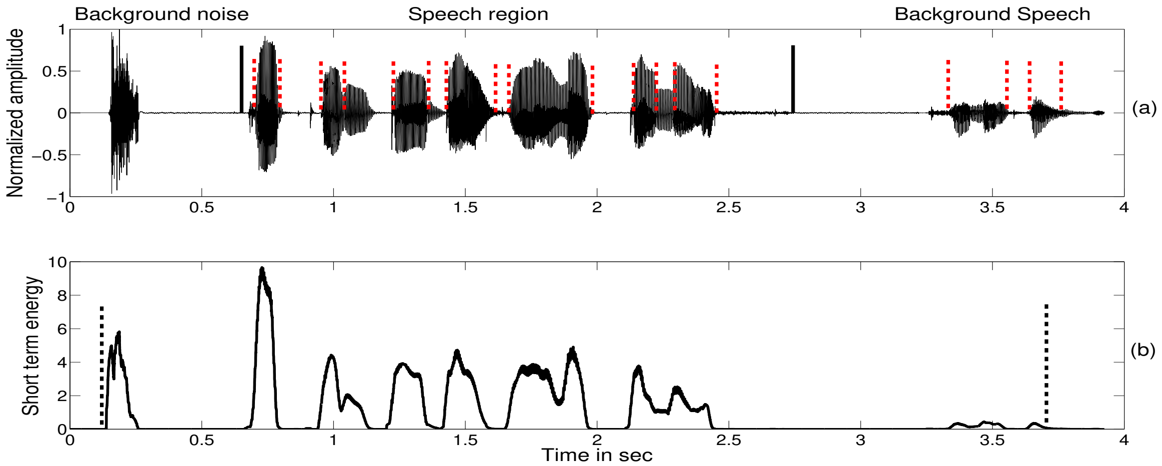


Figure 5.1: (a) Speech signal containing non-overlapping background noise and background speech. Dark vertical lines show the ground truth begin and end points. Dotted lines are manually marked vowel-like region onset and offset points in the speech region. (b) Short term energy of the signal in (a). Dotted lines show the end points detected by the energy based end points detection.

much deviated from the last VLREP of the speech region. The obstruent region at the end of sentence must be detected to correctly detect the end point.

Another disadvantage of VOP based method is the use of only excitation source information. Excitation source information normally captures the impulse like excitation of the source. Sometimes impulse like noise present in the signal is also captured by the instantaneous source information and detects them as spurious VLRs [174]. Therefore, vocal tract system information is required for eliminating the spurious detection. Moreover, VOPs are detected well when there is a consonant attached to it, for example, in case of an obstruent-vowel unit. It is because the signal strength changes suddenly at the onset of the vowel when there is a stop consonant or a fricative. In case of semivowel-vowel units, the VOP detection performance degrades [55,175]. VEPs are also detected well when there is a consonant at the end. In the absence of a consonant, the vowel energy decreases slowly and it becomes difficult to detect the VEP. Such vowel regions with very low strength can be detected by exploring more detailed information of the excitation source in terms of glottal activity. GAD will detect some other sonorant consonants along with the VLRs. But in this work, those sonorants are desirable as the final goal is to detect the end points of speech.

In Fig. 5.1, it can be seen that the speech is degraded with background speech and, VLROPs and VLREPs are detected in the background speech regions as well. In such a scenario, a FSS module may be required to properly detect the begin and end points. It is also observed that proposed method doesn't work where non-speech and background speech regions are overlapped with speech regions.

Taking all these issues into consideration, this work proposes a method to detect the begin and end points using some speech specific knowledge. The proposed method uses both excitation source and vocal tract system information to reduce the spurious detection. Following is the step-by-step process of the proposed method.

- The vowel-like regions (VLRs) are detected using excitation source information.
- The vocal tract system information obtained from dominant resonant frequency (DRF) is used to reduce spurious detection in background noise.
- Glottal activity detection is used to detect the sonorant consonants present at the beginning and end of the speech signal, which cannot be detected by VLRs detection.
- Obtruent consonants are detected by using a sub fundamental frequency filtering method.
- To eliminate background speech, a foreground speech segmentation module is used.
- Finally, *speech duration knowledge* (SDK) is used to further refine the output of the begin and end points detection.

Depending on the background speech and the presence of different kinds of degradation, a poor system performance may be encountered. To improve such system, speech specific knowledge is applied on the degraded speech so as to improve the performance in a better way. Even though most of these are existing techniques, our contribution is in terms of exploiting the advantages of different individual techniques and combining them in a suitable manner for robust end point detection in a TDSV system.

Rest of the Chapter is organized as follows: Section 5.2 describes significance of different speech specific knowledge for the detection of speech end points. Section 5.3 illustrates the procedure for robust EPD. Experimental evaluation of the proposed EPD method is performed in section 5.4. Use of the proposed EPD in TDSV system is described in section 5.5. Section 5.6 evaluates the proposed TDSV system and section 5.7 summarizes the work and concludes.

5.2 End Point Detection Using Speech Specific Knowledge

The task of EPD involves marking the starting and end points of a speech utterance. Presence of background noise makes it difficult to locate the end points. In practical environment, there are different types of background noises present. Each background noise have their own type of characteristics. Therefore, looking the EPD problem from speech knowledge point of view is more appreciable in practical environment. Some of the existing methods indirectly use speech specific knowledge. However, those methods fail because the speech knowledge used in those systems is not sufficient. In this section, speech specific knowledge required for eliminating the limitations of energy based and VOP based methods are described in detail.

5.2.1 Vowel-like Region Onset and End Point Information

Vowels, semi-vowels and diphthongs are considered as the VLRs. VLRs are high SNR regions and easier to detect in degraded condition [51]. There are many methods in the literature for the VLR detection [51,176]. Most of these methods are based on detection of VLROP and VLREP. Knowledge of VLREPs can be used for detecting the speech end points, instead of using the last VOP. Vowels are of variable duration and many languages have long and short version of the same vowel. Detecting the last VOP and marking the end point 200 ms after the VOP is not a good idea in such cases, rather, use of VLREP may be more accurate. The method described in [51] is used for VLRs detection in this work. Although VLROP and VLREP detection uses excitation source information only, it uses two different evidences to explore the information unlike the previous method [50], where VOPs are detected using only a single evidence. One VLROP evidence is obtained using the HE of LP residual by processing the speech signal through the following steps : The speech frame is processed in blocks of 20 ms frame size with a frame shift of 10 ms. For each 20 ms frame size, 10^{th} order LP analysis is performed to estimate the LPCs. The speech frame is passed through the inverse filter to extract the LP residual. This excitation source characteristics are further enhanced by determining the HE of LP residual [56]. The detailed method used to detect the VLROP evidences using HE of LP residual is described below.

Method 1 : VLROP Evidence Using HE of LP Residual

→ Let $l(n)$ is the LP residual signal

→ Its analytic signal is $l_a(n)$ is given by

$l_a(n) = l(n) + jl_h(n)$

where, $l_h(n)$ is the Hilbert transform of $l(n)$

→ Let $h_l(n)$ be HE, which is the magnitude of $l_a(n)$

$h_l(n) = |l_a(n)|$

$h_l(n) = \sqrt{l^2(n) + l_h^2(n)}$

→ This HE of LP residual enhances information about glottal closure instants (GCIs)

→ The evidence is further enhanced by taking maximum value of the HE of LP residual for every 5 ms block with one sample shift

→ The smooth contour is convolved with a first order Gaussian differentiator (FOGD) window of length 100 ms and standard deviation of one sixth of window.

→ The output of the convolution is the VLROP evidence using HE of LP residual.

Low frequency components are accommodated in the HE ($h_l(n)$) and relatively high frequency components are accommodated in phase ($\Phi(n)$) of the analytic signal. Hence, the HE of LP residual enhances speaker information about glottal closure instants (GCIs) [24, 55].

Following steps are used for computing another evidence from the ZFFS. The output of ZFF consists of excitation information which is mainly impulses due to excitation. The rate of change of excitation strength is the main event at VLROP [24].

Final VLROP evidence using the excitation source information is obtained by adding the two evidences and normalizing by the maximum value of sum. The location of peaks between two successive positive to negative zero crossings of the combined evidence represent the hypothesized VLROPs. To find the VLREP evidences, both the smoothed HE of LP residual and second order difference of ZFFS are convolved with FOGD from right to left as opposed to left to right in case of VLROPs and all other steps are kept same [59]. The locations of peaks between two successive positive to negative zero crossings of the combined VLREP evidence represent the hypothesized VLREPs.

Method 2 : VLROP Evidence Using ZFFS

The algorithmic steps for computing VLROP evidence from ZFFS are as follows:

→ Difference the speech signal $s(n)$

$d(n) = s(n) - s(n-1)$

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

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

where, $a_1 = 4$, $a_2 = -6$, $a_3 = 4$, $a_4 = -1$ and $d(n)$ is the differenced speech signal

→ Remove the trend i.e.,

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

where, $\bar{y}(n) = \frac{1}{(2N+1)} \sum_{m=-N}^{N} y(n+m)$ and $2N+1$ corresponds to the average pitch period computed over a longer segment of speech.

→ The signal ($\hat{y}(n)$) obtained after removing the trend is the ZFFS

→ The strength of excitation at the epochs are extracted by computing the first order difference of ZFFS.

→ The second order difference of ZFFS contains change in the strength of excitation .

→ This change is detected by convolving with a FOGD.

→ The convolved output is called the VLROP evidence using ZFFS.

The number of hypothesized VLROPs and VLREPs will be different due to independent detection, with misses and spurious detection occurring for both events. To get the most optimum sets of VLROP and VLREP, an algorithm is proposed using some speech specific knowledge [51]. The algorithm forced the detection of missing cases in one evidence, if other evidence is sufficiently strong. The algorithm also reduces spurious detection of one event using knowledge of the other. Refined sets of VLROP and VLREP are used to obtain the VLRs.

Fig. 5.2 (a) shows the same speech signal as shown in Fig. 5.1 (a) containing speech and non-speech background noise. Fig. 5.2 (b), (c) and (d) show the VLROP and VLREP evidences and detected VLRs, respectively. The non-speech background noise present in between 0 - 0.5 sec may also have impulse-like characteristic due to which the region is detected as VLR. Some portion of speech background present in between 3 - 4 sec are also detected as VLRs. These spurious detection will lead to wrong detection of the begin and end points. Therefore, some vocal tract system information should be explored to reduce the spurious detection.

5.2.2 Dominant Resonant Frequency Information

To remove spurious VLRs, DRF can be used as vocal tract system information. DRF is the frequency which is resonating the most by the vocal tract [177]. DRF is computed from the *Hilbert*

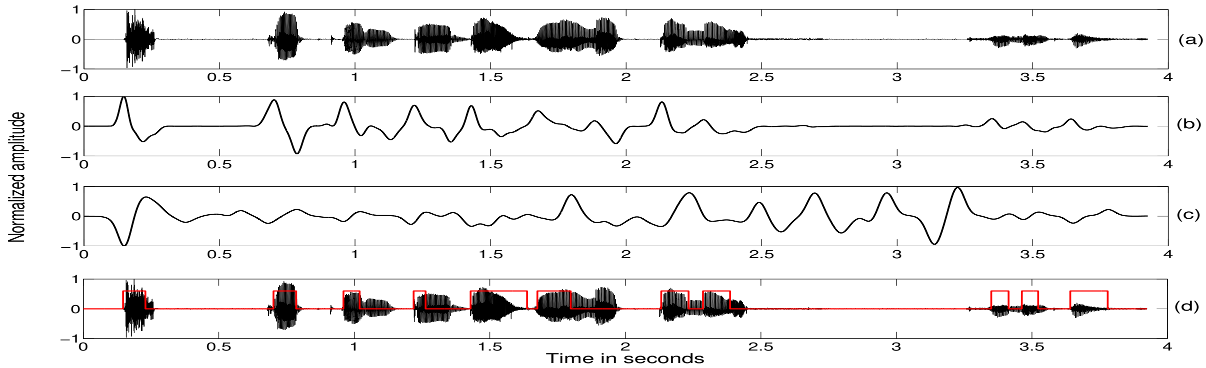


Figure 5.2: (a) Speech signal containing non-overlapping background noise (b) VLROP and (c) VLREP evidences obtained using excitation source information. (d) Detected VLRs marked using dark lines.

envelope of numerator group delay (HNGD) spectrum of zero time windowed speech [178,179]. DRF is less than 1 kHz for vowel like regions. Non-speech noise mostly contains high frequency components. Hence, DRF is more than 1 kHz for non-speech noise. This knowledge can be used for identifying and removing the spurious VLRs.

Fig. 5.3 (a) shows speech with background noise along with the detected VLRs. The DRF contour is shown in Fig. (b). The DRF values are low for VLRs and high for background noise. The VLRs where DRF value is more than 1 kHz are considered as non-speech region. In Fig. 5.2 (b), the refined VLRs are shown using dotted lines. It is seen that spurious VLRs in the non-speech background noise region in between 0 - 0.5 sec is removed. Thus, DRF can be used for removing some of the spurious VLRs.

The procedure for detection of non-speech background noise using DRF is described below :

Method 3 : Detection of Non-speech Background Noise Using DRF

→ Differenced speech signal $s[n]$ is multiplied two times with a zero time window ($w_1[n]$)

$$w_1[n] = \begin{cases} 0, & n = 0 \\ \frac{1}{4sin^2(\pi n/2N)}, & n = 1, 2, ..., N-1 \end{cases}$$

where, N is window length. In this work N is considered to be samples equivalent to 5 ms.

→ Zero time windowing operation is performed to get better time resolution.

→ The truncation effect at the end of the window is reduced by multiplying with a tapering window function $w_2[n]$ given by

$w_2[n] = 4cos^2(\frac{\pi n}{2N}), n = 0, 1, ...N-1$

→ Numerator group delay function ($g[k]$) of the the resultant signal is obtained to nullify the loss in frequency resolution.

$g[k] = X_R[k]Y_R[K] + X_I[k]Y_I[K], k = 0, 1..., N-1$

where, $X[k] = X_r[k] + jX_I[k]$ is the N-point discrete Fourier transform (DFT) of the windowed signal $x[n]$ and $Y[k] = Y_r[k] + jY_I[k]$ is the N-point DFT of the sequence $y[n] = nx[n]$

→ Final spectrum is obtained by taking Hilbert envelope of the differenced numerator group delay function.

→ The frequency corresponding to the maximum value in the spectrum is considered as the DRF.

→ The VLRs where DRF value is more than 1 kHz are declared as non-speech region.

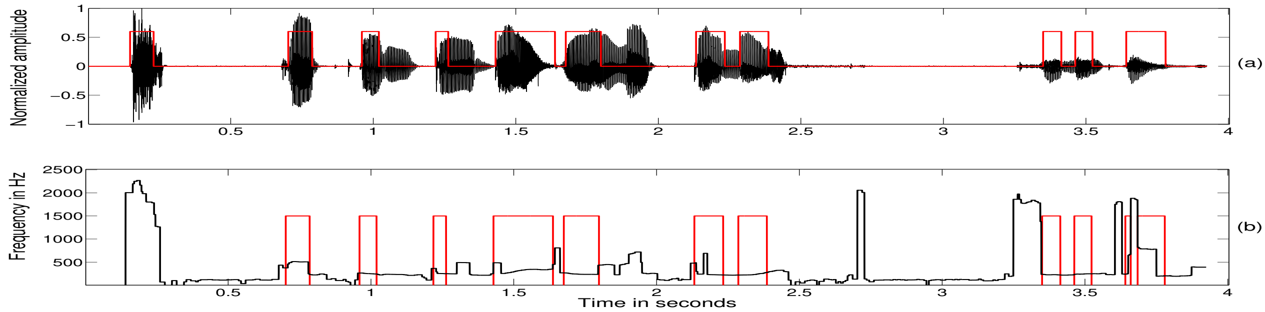


Figure 5.3: (a) Speech signal with non-overlapping background noise. Hypothesized VLRs are shown using dark lines. (b) DRF contour of the speech along with refined VLRs. Spurious VLRs in non-speech background region are removed after using the DRF information

5.2.3 Foreground Speech Segmentation

The speech of speaker who is close to the microphone sensor and directly communicates to a speech application is known as the foreground speech. The rest of the acoustic background captured by the sensor is known as the background degradation [180]. Background degradation contains both speech and non-speech noise. Non-speech noise can be removed by using DRF as described in the previous subsection. To remove the background speech, a FSS is necessary. A FSS algorithm was proposed in [180], which was further modified in [102]. The modified version of the FSS method is used in this work. The method uses excitation source and vocal tract system information. Excitation source information is extracted using ZFFS analysis and vocal tract auditory features are extracted from the modulation spectrum energy.

Method 4 : Detection of Foreground Speech Segmentation

The foreground speech segmentation algorithm is described as follows:

→ Excitation source information is extracted using ZFF analysis

→ The ZFF signal is processed in blocks of 50 ms with shift of one sample

→ Normalized first order auto-correlation coefficients and strength of excitation are extracted

→ The features show high value in the foreground speech regions compared to background speech region

→ Vocal tract articulatory features are extracted from the modulation spectrum energy using compressive Gamma chirp auditory filter

→ The linear combination of these features are further subjected to an end point detection algorithm to segment the foreground speech regions over a longer segment of speech.

Fig. 5.4 (a) shows speech signal with background speech in the 3 - 4 sec region. Fig. 5.4 (b) shows

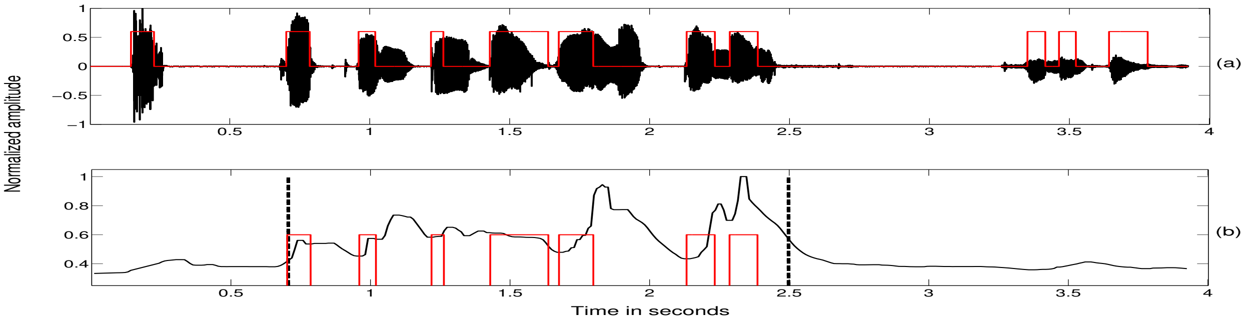


Figure 5.4: (a) Speech signal with non-overlapping background noise. Detected VLRs are shown using dark lines. (b) Final evidence for foreground speech segmentation. Vertical dotted lines show the segmentation boundaries. Spurious VLRs in speech background can be removed using foreground speech segmentation

the final evidence obtained from the foreground segmentation algorithm along with the segmented boundaries (in dotted lines). With the help of foreground segmentation, the spurious VLRs in the background speech region can be removed and the VLRs present in the foreground speech region can be retained.

5.2.4 Glottal Activity Information

After detecting the VLRs, glottal activity information can be explored to add a sonorant consonant at the begin or at the end of speech utterance. GAD detects the VLRs as well as other sonorant consonant regions, such as voice bars, nasals etc. Therefore, better localization of voiced region can be done by using GAD. This GAD method also helps to minimize the VLRs miss detection rate.

A procedure for detection of voiced region using ZFF method is described in [181]. Detection of voiced speech involves detection of significant glottal activity in speech. Rest of the regions include unvoiced speech, silence as well as background noise. The method exploits the nature of ZFFS that addition of a small amount of noise to the speech signal does not affect the zero crossings of ZFFS in the voiced region, whereas it leads to zero crossings at random locations in unvoiced and silence regions. The reason behind this is the glottal closure that takes part in producing the most significant impulse-like excitation in voiced speech. The epochs due to such impulse-like excitation are robust to noise and can be located with high accuracy in presence of degradation. On the other hand, unvoiced and silence region do not contain any significant excitation and results in zero crossings located at random instants. The drift in the epoch location after addition of a little amount of noise is used to segment the voiced regions from rest of the regions. Additional spurious epochs are eliminated using

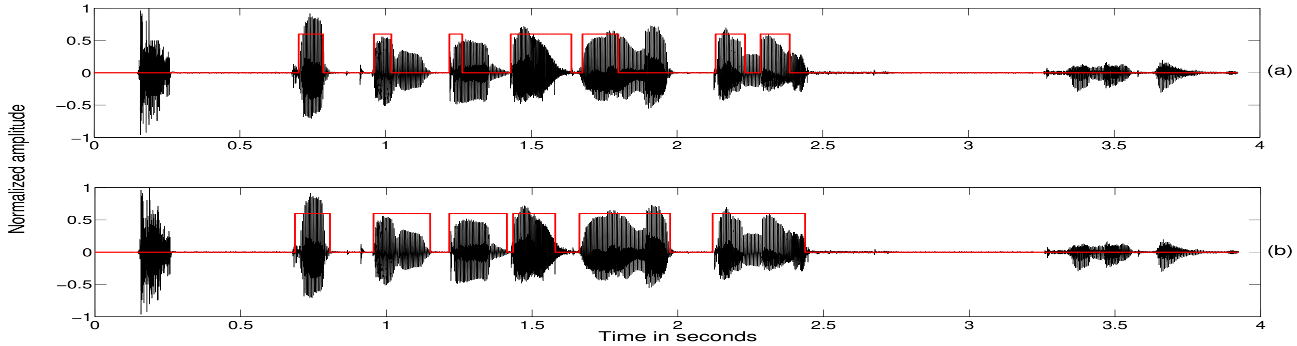


Figure 5.5: Speech signal with hypothesized (a) VLRs and (b) glottal activity regions. Missed VLRs and the sonorant consonants are detected by using glottal activity detection

the knowledge of pitch and jitter information The detailed method for detection of voiced regions as follows:

Method 5 : Detection of Voiced Regions Using GAD

Procedure for detection of voiced region using ZFF is described in the following way:
→ Obtain the epoch locations of speech using ZFF
→ Add 20 dB of white Gaussian noise to the clean speech
→ Obtain the epoch locations of noisy speech using ZFF
→ If the difference between the epoch locations obtained in the two cases are less than 2 ms, then retain it. otherwise discard the epoch
→ Calculate the instantaneous pitch period using the epoch locations and eliminate the epoch if its pitch period is more than 15 ms
→ Calculate the instantaneous jitter and eliminate the epoch if jitter is more than 1 ms

Fig. 5.5 shows speech signal along with the hypothesized VLRs (in (a)) and detected glottal activity region (in (b)). It is seen that there are some missed VLRs at the end of speech utterance (around 2.4 sec). These missed VLRs are detected by the glottal activity detection algorithm. Further, there is one nasal sound in between 1 - 1.2 sec which is not considered as VLR. GAD will be helpful if such sonorant is present at the beginning or at the end of speech utterance. Detection of such sonorant sound is required for appropriately detecting the end points.

5.2.5 Obstruent Information

Stop consonants, fricatives and affricates are known as obstruents. There may be an obstruent consonant at the begin and end of speech utterance. The obstruents are not detected by the VLR detection. To include them in the speech region, an obstruent detection algorithm is required. In [182], a method is proposed to detect the dominant aperiodic regions in speech using sub fundamental frequency filtering. Since aperiodic component is dominant in burst and frication region of most of the obstruents, the same method is used in this work to detect the obstruents. The method is a modification of ZFF method. The steps involved in the method are described below.

Method 6 : Detect the Dominant Aperiodic Regions in Speech

Procedure for detection of dominant aperiodic region is described in the following way:

→ Pass the speech signal through a 0 Hz resonator four times

→ Obtain the sub fundamental frequency filtered signal by removing the trend four times using a window size of three pitch periods

→ Enhance the filtered signal in the dominant aperiodic region by squaring it.

→ Smooth the enhanced signal using a 5 ms window

→ For further smoothing, locate the peaks in the signal using a peak picking algorithm and repeat the peak value till the next peak is obtained

→ A FOGD of 100 ms length is convolved with the signal to get the final evidence

→ A positive peak in the final evidence gives the probable onset of an obstruent.

→ The first positive peak present within 100 ms before a VLROP is considered as the onset of obstruent and

→ The last negative peak present within 100 ms after the VLREP is considered as the offset of an obstruent.

Fig. 5.6 (a) shows a speech signal with hypothesized end points detected by the VLR detection method. Vertical dotted lines show the ground truth end points. The first VLROP and last VLREP are much deviated from the ground truth marking. Fig. 5.6 (b) shows the final obstruents evidence. The first positive peak present within 100 ms of first VLROP is considered as the refined begin point and the last negative peak present within 100 ms region of the last VLREP is considered as the refined end point. In Fig. 5.6 (c), it can be observed that the end points are detected with more accuracy after using obstruent detection.

5.2.6 Speech Duration Knowledge

Sometimes database is collected from an unknown environment without any restriction on the testing condition. The subject is not restricted to speak the given sentence in a proper way, but left free to speak in a very casual way. In such cases, the subject may try to speak one or two words

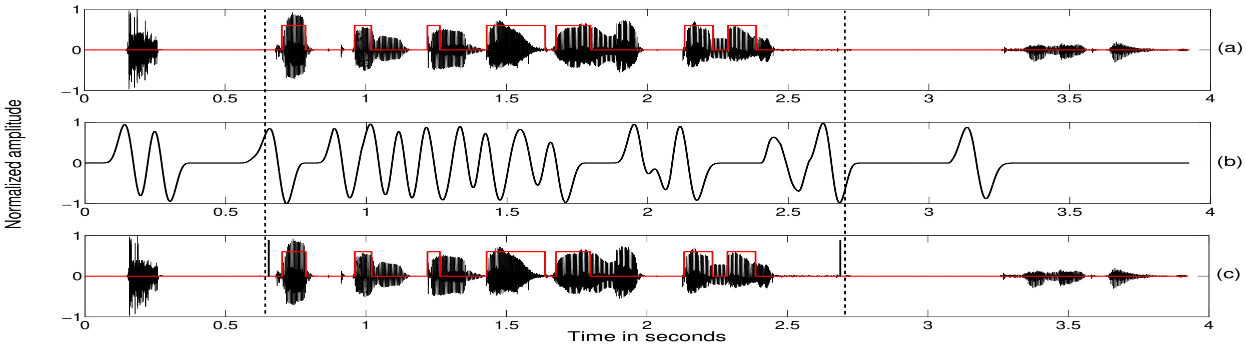


Figure 5.6: (a) Speech with hypothesized VLRs after spurious removal. (b) Final obstruent evidence. (c) Speech with refined begin and end points. Vertical dark lines show the detected end points and long vertical dotted lines show ground truth end points. Burst region at the beginning and the frication region at the end of the signal are included after the refinement

several times. For example instead of saying "*Lovely picture can only be drawn*", one may say "*Lovely [pause] lovely picture can only be drawn*". Sometimes one may repeat a word twice and can speak something which is not part of the sentence. When these extra words are uttered without any pause, it is not possible to do any further rectification. However, if these words are uttered at the beginning or at the end leaving some silence in between the actual utterance and the extra word, it is possible to remove the extra utterance by using SDK. In normal speech distance between two subsequent VLRs cannot be more than 300 ms [49]. If it is more than 300 ms, then it is not part of the given sentence. This knowledge will also help to remove other non-speech spurious VLRs which are detected 300 ms away from the speech utterance. Following procedure is followed to use the SDK.

Method 7 : Detection of Speech Regions Using Speech Duration Knowledge

Procedure for detection of refined end point regions is described in the following way:

→ First, location of all the VLROPs are obtained

→ Average of the euclidean distances from one VLROP to all other VLROPs are computed.

→ The VLROP having the minimum average euclidean distance is considered as the center (C) of the speech utterance.

$C = \arg\min_i \frac{1}{N-1} \sum_{j=1,\neq i}^{N} (V_i - V_j)^2$

· where, $1 \leq i \leq N$, $1 \leq j \leq N$, C is the detected center of speech utterance, N is the number of detected VLRs and V_i is the i^{th} VLROP in number of samples

→ Now, starting from the center of the speech, duration between two successive VLRs are calculated on either side until the duration is found to be greater than 300 ms or the peripheral VLR is reached.

→ If the duration between two successive VLRs is found to be greater than 300 ms, then between the two VLRs, the one which is closure to C is declared as the peripheral VLR.

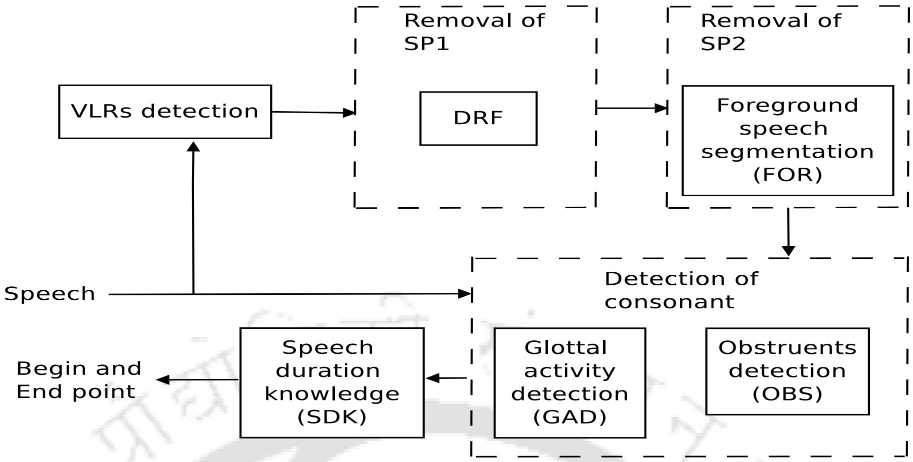


Figure 5.7: Block diagram of proposed end point detection. SP1 and SP2 denotes spurious VLRs in non-speech and speech background, respectively.

5.3 Robust End Point Detection using Speech Specific Knowledge

The speech specific knowledge described in the previous section can be systematically used to detect the end points in practical conditions containing both speech and non-speech degradation. Fig. 5.7 shows the block diagram for the proposed method. First VLRs are detected from the speech signal. Spurious VLRs in the non-speech background is removed by using DRF information. Spurious VLRs in speech background is removed using foreground speech segmentation. Then, missed VLRs and sonorant consonants present at the begin or end are detected using GAD. The obstruent consonants present at the begin or end are detected and end points of peripheral VLRs are modified. Finally SDK is used to further remove the spurious VLRs. The VLROP of first VLR and VLREP of last VLR are declared as the begin and end points, respectively.

Fig. 5.8 illustrates the proposed detection procedure. The same speech files used in the previous sections are used for the illustration. Fig. 5.8 (a) shows the speech signal with non-overlapping non-speech noise and speech background. First step is to detect the VLRs. Fig. 5.8 (b) shows the detected VLRs. Second step is to remove the non-speech noise using the DRF information. The spurious VLRs where the DRF is exceeding 1 kHz are removed and refined VLRs are shown in Fig. 5.8 (c). Third step is to remove the speech background which is done using foreground speech segmentation. The detected boundaries and refined VLRs obtained after foreground speech segmentation are shown in Fig. 5.8 (d). In the next step the glottal activity regions are detected. The detected glottal activity regions are added to VLRs and are shown in Fig. 5.8 (e). This includes some sonorant regions

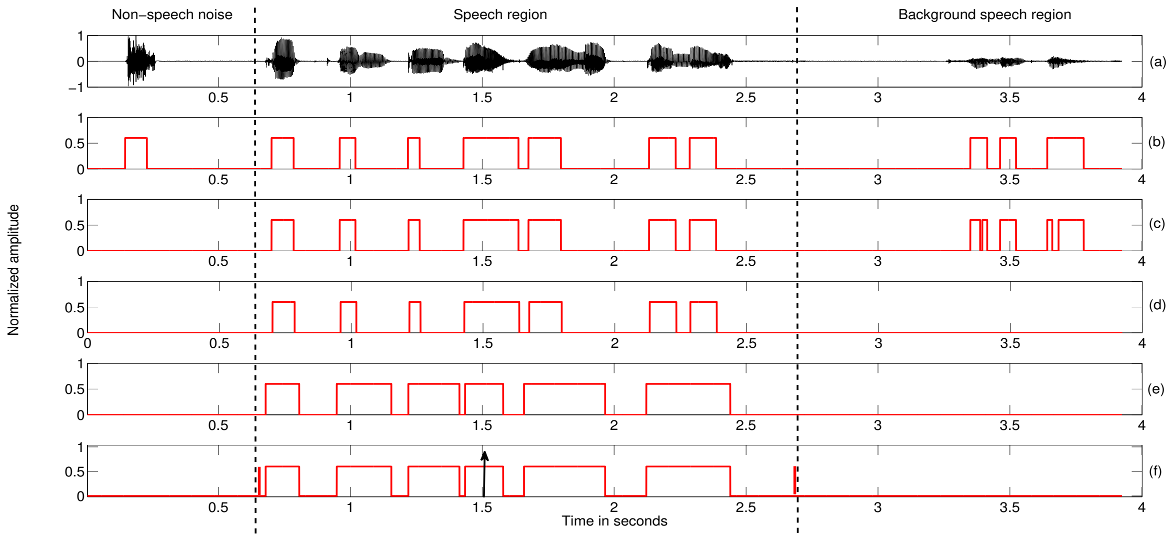


Figure 5.8: Illustration of the begin and end point detection procedure. (a) Speech signal with non-speech noise and background speech. (b) Detected VLRs. (c) VLRs after removing the non-speech noise using DRF information. (d) VLRs after removing the speech background using foreground speech segmentation. (e) Detected glottal activity regions added to the VLRs. (f) Refined begin and end point using obstruent information. The arrow around 1.5 s shows the center C of the speech utterance. Duration between successive VLRs are less than 300 ms. Therefore, no further modification is made using SDK knowledge. Dotted line shows ground truth manual marking.

which are not detected by the VLR detection algorithm. Fifth step is to detect the obstruents at the beginning and at the end of the speech utterance. The modified begin and end points using obstruent evidence are shown in Fig. 5.8 (f). The arrow in Fig. 5.8 (f) shows the center of speech utterance detected. Starting from this point, the duration between successive VLRs are computed in both sides. For this utterance, the inter VLR time interval is less than 300 ms for all cases and hence, no further modification is incorporated using the SDK knowledge.

5.4 Experimental Evaluation of Proposed EPD

5.4.1 Databases

In order to evaluate the robustness of proposed algorithm in degraded speech conditions, we conducted the TDSV experiments on the two databases, namely, the IITG database [38, 46] and the RSR2015 database [14]. The RSR2015 database is mainly designed for TDSV with pre-defined pass phrases. This database contains 71 hours of audio recording, recorded on a set of six portable devices including mobile phones and tablets, in a closed office environment. The speech data is recorded from 300 English speakers (143 females and 157 males) decorating the variety of accents spoken in Singapore. These 300 speakers were selected to be representative of the ethnic distribution of Singaporean population, with age ranging from 17 to 42 years old. Each of the speakers recorded 9 sessions, each session consists of 30 pre-defined pass phrases. In this work, three pass phrases are optionally have been selected for the experimental studies. The duration of the utterances vary between 3-5 sec. Out of 9 utterances (sessions) for every pass phrase, 3 utterances are used for training, and remaining 6 utterances are used for testing. The pass phrases used in this study are shown below.

TD-1 : *"Only lawyers love millionaires"*.

TD-2 : *"I know I did not meet her early enough"*.

TD-3 : *"The Birthday party has cupcake and ice-cream"*.

As the RSR2015 database is collected in a closed office environment, it is considered as clean database. To evaluate the proposed method in degraded condition, the speech signal is corrupted with Babble noise [110]. The energy level of the noise is scaled such that the overall SNR of the degraded speech varies from 20-0 dB, in steps of 5 dB.

The IITG database was collected from the course students with attendance as an application to address issues in the practical deployment of TDSV system [38, 46, 105, 183]. The speech biometric based attendance system developed for course students of IITG was used on a regular basis for marking attendance [38, 46, 184]. The students used to call to a voice-server based toll free number to mark attendance using fixed set of mobile handsets kept at the department office. In the enrollment phase, three pre-defined pass phrases were recorded for each enrolled speaker in an anechoic chamber. The data was recorded in three different sessions using a set of mobile handsets. The three pass phrases from the IITG database are:

TD-1 : *"Don't ask me to walk like that"*.

TD-2 : *"Lovely pictures can only be drawn"*.

TD-3 : *"Get into the hole of tunnels"*.

During the testing phase, the speakers are prompted to utter one of the three pre-defined utterances randomly selected by the voice-server. Collection of the testing data was continued for an entire academic semester from a population of 325 speakers (286 males and 39 females). On an average, 30 utterances are available per speaker for each phrase. These speakers constitute an age group in between 25-35 years. The duration of the utterances of the database vary between 3-5 sec. During collection of the testing database, the speakers could move freely within, and in and out of the open hall in the department office and also are allowed in free environment. This makes the testing data more practical because the data includes background noise, background speech and other environmental conditions. Due to mismatches in sensor, style of speech and environment between the enrollment and testing phases, the TDSV task becomes more challenging in IITG database. Hence, there is no need to add artificial noise in the IITG database for the assessment of the TDSV system.

5.4.2 Performance Evaluation

The performance of the EPD algorithm is evaluated in terms of two metrics : *Identification rate* (IR) and *Identification error* (IE) [185]. These utterances are marked manually for ground truth begin and end points of the speech signal. These begin and end points are called as reference points. The begin and end points detected by applying different automatic detection algorithms are called as estimated points. Let l_b^{ref} and l_e^{ref} denote the midpoints (in ms) of begin and end frames of a speech utterance/ signal. All speech frames before the begin frame and after the end frame are considered as silence frames. Let l_b^{est} and l_e^{est} denote the midpoints (in ms) of the begin and end frames of speech signal, as estimated by the EPD algorithm. If $l_b^{est} \in l_b^{ref} \pm 500$ ms, then the begin point is considered to be correctly estimated. Similarly, if $l_e^{est} \in l_e^{ref} \pm 500$ ms, then the end point is considered to be correctly estimated. We may define IR and IE as :

IR (%) : Percentage of speech utterances for which the begin point (end point) is correctly estimated.

IE (ms) : For every begin point (end point) that is correctly estimated, the absolute difference between the reference begin point (end point) and its estimate is calculated. In other words, $\mid l_b^{est} - l_b^{ref} \mid$ ($\mid l_e^{est} - l_e^{ref} \mid$) is calculated. The average value of the absolute difference for all the correctly estimated begin points (end points) is denoted as IE.

5.4.2.1 Performances on the RSR2015 Database under Clean and Degraded Condition

The performance of proposed begin and end point detection method is evaluated for clean speech using 300 speakers data from the RSR2015 database for three utterances *"Only lawyers love millionaires"*, *"I know I did not meet her early enough"*, and *"The Birthday party has cupcake and ice-cream"*. The begin and end points are marked manually to obtain the ground truths.

In Table 5.1, performance of automatically detected begin and end points are shown for TD-1 phrase of clean RSR2015 database. The proposed method is compared with different existing methods in terms of IR and IE. The evaluation is also carried out at different stages in the block diagram of the

Table 5.1: Performance of begin and end points detection in terms of Identification Rate and Identification Error, evaluated on clean RSR2015 database.

Method	Metric	Energy	FSS	GAD	VLR	VLR+DRF	VLR+DRF +FSS	VLR+DRF +GAD	VLR+DRF +GAD+FSS	VLR+DRF+GAD +FSS+OBS	VLR+DRF+GAD +FSS+OBS+SDK
Begin	IR (%)	96.64	96.52	98.39	98.39	97.82	97.80	98.46	98.28	99.02	**99.17**
	IE (ms)	163	53	177	73	61	177	52	51	48	**39**
End	IR (%)	96.82	80.00	90.00	94.93	95.64	95.93	95.46	96.02	96.64	**98.78**
	IE (ms)	**69**	115	323	207	215	299	238	178	199	204

proposed method so that comparison is made with the individual methods as well as with different possible combinations. It can be seen that the VLR based detection gives around 1.75 % improvement over Energy based detection for begin and end points. In case of GAD based detection, performance increases by around 1.75 % for begin points and decreases by around 7 % for end points compared to Energy based detection. In case of FSS, the performance decreases for both begin and end points. It is also observed that performance of begin point detection increases for different combinations. For the proposed detection method which uses combination of all individual methods, the performance increases by around 2.5 % and 2 % for begin and end points, respectively. In terms of IE, the proposed method shows better performance compared to the individual methods for begin points. However, the IE is found to be very high in case of the end points.

Table 5.2: Performance of begin and end points detection in terms of Identification Rate (IR) and Identification Error (IE). The performance metrics are evaluated on the TD-1 utterance of the RSR2015 database. The testing utterances are corrupted by Babble noise, with SNR varying from 20-0 dB in steps of 5 dB.

SNR	Method	Metrics	Energy	FSS	GAD	VLR	VLR+DRF	VLR+DRF +FSS	VLR+DRF +GAD	VLR+DRF +GAD+FSS	VLR+DRF+GAD +FSS+OBS	VLR+DRF+GAD +FSS+OBS+SDK
20 dB	Begin	IR (%)	84.46	96.07	83.57	82.14	96.07	**99.64**	98.04	97.86	98.75	96.78
		IE (ms)	165	73	147	155	73	**54**	224	218	224	147
	End	IR (%)	88.93	**93.39**	78.00	59.00	62.00	68.00	58.05	86.43	78.00	86.96
		IE (ms)	220	204	171	168	168	**166**	171	197	173	224
15 dB	Begin	IR (%)	77.14	79.82	**99.82**	97.68	99.82	96.43	98.21	97.25	97.68	96.61
		IE (ms)	355	200	**97**	286	193	357	304	332	322	229
	End	IR (%)	49.82	**89.28**	80.00	50.00	61.14	68.9	77.50	76.16	74.1	77.14
		IE (ms)	354	199	**169**	170	177	171	174	184	175	248
10 dB	Begin	IR (%)	60.00	71.61	99.28	96.25	95.72	**99.29**	96.96	96.98	96.42	96.25
		IE (ms)	376	**275**	298	303	406	298	345	352	363	310
	End	IR (%)	45.89	**79.46**	48.04	48.60	51.60	58.00	70.02	73.12	67.30	72.32
		IE (ms)	366	240	180	170	178	180	178	**168**	179	272
5 dB	Begin	IR (%)	59.00	68.75	**98.03**	95.35	95.71	95.17	97.14	96.81	96.43	95.17
		IE (ms)	376	**296**	440	292	396	419	392	399	407	381
	End	IR (%)	45.83	**72.85**	40.50	52.00	41.40	40.01	53.40	58.31	61.40	67.85
		IE (ms)	376	269	177	168	178	179	178	**159**	179	277
0 dB	Begin	IR (%)	57.00	61.61	95.00	94.10	95.17	95.89	**96.25**	95.78	95.00	94.46
		IE (ms)	376	327	494	**271**	443	433	408	423	458	446
	End	IR (%)	45.75	**72.14**	24.8	53.90	27.90	29.10	34.60	46.28	55.90	58.92
		IE (ms)	366	266	169	**164**	180	180	171	192	181	276
20-0 dB	Average Begin & End	IR	61.38	78.50	74.71	72.9	72.65	74.98	78.02	82.49	82.09	**84.25**
		IE	333	235	234	**215**	239	244	255	263	266	281

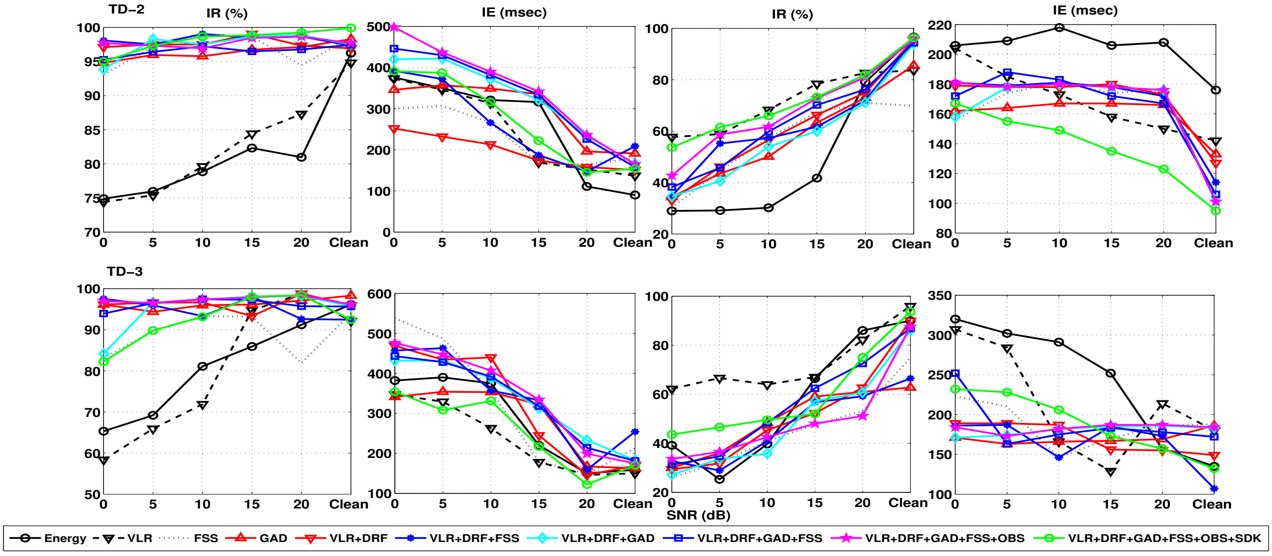


Figure 5.9: The performance of the begin and end point detection using four individual methods and their six different combinations are summarized on TD-2 and TD-3 test utterances in terms of the IR and IE on RSR2015 database. The RSR2015 database is corrupted by Babble noise, with SNR varying from 20-0 dB in steps of 5 dB.

To evaluate the robustness of the proposed begin and end point detection algorithm, the testing utterances of the RSR2015 database are corrupted with Babble noise from NOISEX-92 database [110]. The SNR is varied from 20-0 dB in steps of 5 dB. In Table 5.2, the performance is shown for TD-1 phrase of the RSR2015 database. Due to degradation in the speech signal, the performance is quite low compared to the clean condition. The energy based EPD shows around 10 % decrement in performance for 20 dB SNR. Similar decrement is observed for other individual methods. However, for the proposed method using combination of all the algorithms, an overall improvement is observed across all 5 levels of degradation. In some cases we can observe that the individual method is performing better than the proposed method in terms of IR or IE or both. For example in case of 10 dB SNR, GAD is showing 99.28 % IR and 298 ms IE for begin points which is slightly better than the performance of the proposed method. However, for the GAD method with the same SNR, we can observe that the performance of the proposed method is significantly higher than the GAD for end point detection. Similarly, although FSS is performing better than the proposed method for end points, it is under performing with a significant margin for the begin points. On an average, the performance of the proposed method is found to be better than the individual methods and different intermediate combinations of the individual methods. Complementary advantages of the individual methods lead to improved performance for the proposed method. Results for TD-2 and TD-3 phrases of RSR2015 database are shown in Fig. 5.9 for both clean and degraded speech conditions, and similar trend is observed.

In Table 5.2, it can be observed that even though IR is better for the proposed methods, most of the times the IE in case of begin points is poorer compared to some of the individual methods. This is because of the method used in this work for computing IE, where only detected end points are considered. The end points beyond ±500 ms of the reference mark are not considered for IE computation.

5.4.1.2 Performances on the IITG Database

Table 5.3: Performance of begin and end points detection in terms of Identification Rate and Identification Error. The performance metrics are evaluated for the TD-1, TD-2 and TD-3 sentences of the IITG database. These testing utterances are collected in the practical environment.

Technique			Energy	VLR	FSS	GAD	VLR+DRF	VLR+DRF +FSS	VLR+DRF +GAD	VLR+DRF +GAD+FSS	VLR+DRF+GAD +FSS+OBS	VLR+DRF+GAD +FSS+OBS+SDK
SNR	Method	Metrics										
TD1	Begin	IR (%)	94.35	94.93	94.93	97.27	95.32	97.46	93.95	95.00	**97.85**	97.08
		IE (ms)	**37**	62	47	47	61	42	68	59	42	44
	End	IR (%)	87.13	77.78	76.80	88.11	78.55	93.17	91.42	92.00	93.56	**97.07**
		IE (ms)	**75**	126	115	97	123	96	148	86	80	79
TD2	Begin	IR (%)	95.79	96.78	95.06	98.17	95.25	99.08	87.56	89.76	**99.08**	96.16
		IE (ms)	**27**	30	50	48	73	37	83	63	31	30
	End	IR (%)	87.75	86.67	77.69	92.50	79.52	93.60	91.59	93.24	94.15	**96.89**
		IE (ms)	89	**76**	118	88	165	105	169	142	96	95
TD3	Begin	IR (%)	98.36	96.73	96.73	97.75	96.94	98.97	96.53	97.52	99.18	**99.39**
		IE (ms)	**32**	69	50	61	74	45	84	74	42	40
	End	IR (%)	86.73	83.06	81.22	90.61	83.47	92.86	95.51	94.28	94.28	**97.35**
		IE (ms)	**64**	134	89	84	135	79	135	113	68	66
-	Average Begin & End	IR	91.68	89.33	87.07	94.06	88.18	95.86	92.76	93.63	96.35	**97.32**
		IE	**54**	83	78	71	105	67	115	90	60	59

Same experiments are reported for the IITG database which is more practical data in terms of session, sensor, background degradation and environmental conditions. Table 5.3 shows the performance of different detection algorithms on IITG database. It can be observed that for the practical data, the performance of the Energy based begin point detection is similar to that of the RSR2015 database. However, in case of end points, performance is around 10 % less than the RSR2015 database. Among different individual algorithms, GAD shows the best performance for both begin and end points. The proposed method using combination of different individual methods shows significantly improved performance for both begin and end points detection. On an average, around 2 % and 8 % improvement is observed for begin and end points respectively.

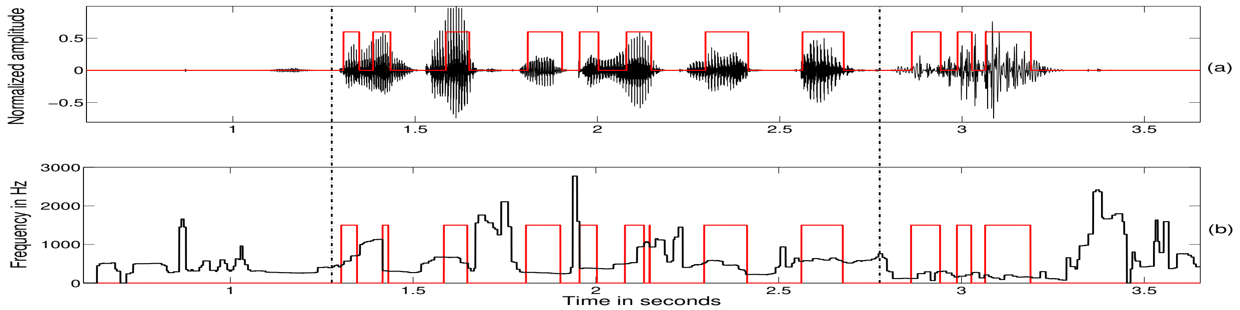


Figure 5.10: (a) Speech signal with non-overlapping background noise along with hypothesized VLRs. (b) DRF contour of the speech along with refined VLRs. Two vertical dotted lines show the begin and end points. Spurious VLRs in non-speech background region are not removed after using the DRF information because the noise have DRF value less than 1000 Hz.

5.4.3 Analysis of Failure Cases

The end points which are not detected by the proposed method are analyzed. The major failure is due to the impulse-like non-speech background having higher low frequency energy content. Due to impulse-like nature, both the VLR detection algorithm and GAD capture these regions as speech region. The DRF value computed from the HNGD spectrum is less than 1000 Hz and therefore cannot remove these spurious detections. Fig. 5.10 explains the situation. Fig. 5.10 (a) shows a speech signal with non-overlapping background noise. Fig. 5.10 (b) shows the DRF contour. The non-speech background has DRF values less than 1000 Hz and are not removed by the proposed method.

In a few cases, the speaker is repeating some words without leaving sufficient space in between the repeated words. SDK is failing to remove spurious detection in such cases. The proposed EPD method is a combination of different algorithms and every individual algorithm has limitations. Sometimes both VLR and GAD algorithms fail to detect some voiced speech regions. Similarly the sub fundamental frequency filtering method may fail to detect some obstruent region. FSS algorithm is threshold based and sometimes fails to detect background speech. Due to the failure in the individual level, the proposed algorithm fails giving a lesser identification rate in such cases.

5.5 Development of Text-Dependent Speaker Verification

The TDSV systems are usually based on template/ model-sequence matching techniques in which the time axes of an input speech utterance and reference models of registered speakers are aligned, and the similarities between them are accumulated from the beginning to the end of utterance [50, 60, 94, 96, 186]. When the linguistic content of the utterance is known, SV systems performs better.This is because such systems can better able to model the characteristics of specific phonetic content contained in the speech signal. To make a SV system more suitable for practical applications, the proposed system should be designed such that, it should be language, sensor independent and session independent and robust to background degradation.

From the commercial speech based authentication systems as an application point of view, the enrolled user is free to provide the test speech at the testing phase with no constraints on the quality, recording conditions, duration, channel mismatch and other environmental noises. Accordingly, the performance of speaker verification systems is influenced by many of these possible variabilities. Among these variabilities, lexical content and channel variations are the most detrimental [14]. Compared to channel variability due to uncontrolled environmental factors, lexical variability can be manageable. In case of text-dependent scenario, lexical variability is not present. Matching of short duration phrases during training and testing with high accuracy makes it an attractive option for commercial speech based person authentication systems.

The accuracy of a SV system depends on the match between the speaker model and test features [107]. The performance may improve by using the proposed EPD algorithm. Fig. 5.11 (a) and (b) show the DTW warping path for an utterance before and after using the proposed EPD algorithm. It can be seen that detection of end points brings the DTW warping path closure to the regression line.

In this work, a TDSV system is developed using 39 dimensional MFCC features and DTW as a template matching technique. The training phase of speakers include 3 sessions of 3 different texts. Features of these training templates are extracted and kept as reference templates for further experiments. In the testing phase, the test data is processed in a similar way as the training data and MFCC features of these test data are extracted. DTW algorithm is used for matching the train and test templates.

The dynamic programming algorithm provides a many-to-many mapping between training and testing utterances while minimizing the overall accumulated distance. This minimum accumulated distance is called as DTW Distance [108, 109]. The testing feature vectors are compared with reference to the claimed model reference template by the DTW algorithm. To find the best match between the train and test feature vectors, dynamic warping path minimizes the total distance between the train and test feature vectors using DTW algorithm. An efficient solution of this minimization is found

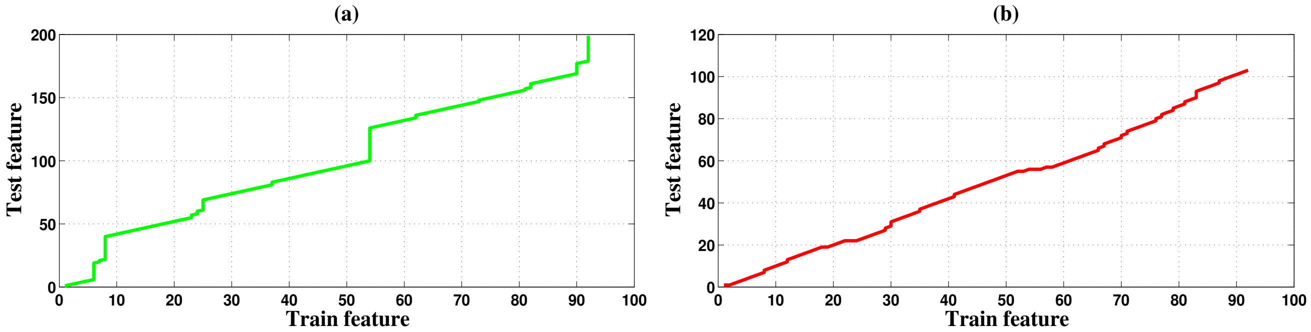


Figure 5.11: Evidences for degraded and clean speech utterances for *Don't ask me to walk like that* (a) speech utterances containing non-overlapping background noise and background speech and showed DTW warping path deviated away from the regression line, (b) speech utterances containing clean speech region after detecting the begin and end points using algorithms and showed DTW warping path along the regression line.

by the dynamic programming due to its elastic distance measure. DTW algorithm computes the accumulated distance score between the reference model of the train template X and testing template T of different lengths by taking the DTW warping path as,

$$DTW_\psi(X,T) = \sum_{l=1}^{M} DTW(\psi_x(l), \psi_t(l))s(l)/S_\psi \tag{5.1}$$

where, $DTW_\psi(X,T)$ is the accumulated distance score between the test and reference template, $DTW(\psi_x(l),\psi_t(l))$ is the shortest time spectral distortion, $s(l)$ is a non-negative warping path weighting co-efficient and S_ψ is the DTW path normalizing factor. The test speech is matched with the claimed reference model using DTW algorithm as given by Eq. 5.1 to give the minimum distance score. The decision is taken with respect to a set of four cohort speakers by comparing the distance score against them. The cohort approach is mainly based on a subset of enrolled speaker's models which represents the closest score to the claimed speaker. This cohort selection is made in advance (off-line), which consists closest model to the claimed speaker. The cohort based method is faster than the traditional method of testing the model against the imposters, reducing the unnormalized EER. This is because the utterance must be from a set of pre-defined phrases.

The performance evaluation of the TDSV system is based on two standard metrics - EER and mDCF. Let $\mathrm{S}c = \{\mathrm{s}c_1, \mathrm{s}c_2, ..., \mathrm{s}c_S\}$ represent the entire set of DTW scores obtained after matching the testing vs. the training utterances. The set $\mathrm{S}c$ is normalized so that $\{\mathrm{s}c_i \in [0,1] \mid i = 1, 2, ..., S\}$. Let $\theta \in [0,1]$ be the decision threshold, above which the claim of the test utterance against the speaker is accepted. Let S_G denote the number of genuine/true claims that have been accepted, S_I denote the number of imposter/false claims that have been rejected, S_M denote the number of genuine/true claims that have been rejected and S_F denote the number of imposter/false claims that have been accepted. Therefore, $S = S_G + S_I + S_M + S_F$. We may now define EER and DCF as follows :

EER (%) : For any given threshold, θ, the *Miss Rate* is given as MR $= \frac{S_M}{S_G+S_M} \times 100$ %. The *False Alarm Rate* is given as FAR $= \frac{S_F}{S_I+S_F} \times 100$ %. At a particular threshold, $\theta = \theta_0$, MR = FAR. This error is known as the EER. In other words, EER = MR = FAR, at $\theta = \theta_0$.

mDCF : For any given threshold, θ, the *Probability of Miss* is given as $P_M = \frac{S_M}{S_G+S_M}$. The *Probability of False Alarm* is given as $P_F = \frac{S_F}{S_I+S_F}$. Two parameters, C_M and C_F, assign costs to the event of a *miss* (a genuine claim rejected), and that of a *false alarm* (an imposter claim is accepted) respectively. Also, an *a priori* probability P_T is assigned, which assumes that out of all the test claims against a speaker, only a fraction P_T are genuine claims. The cost parameter of the mDCF, at any given θ, is given by,

$$C_\theta = C_M \times P_M \times P_T + C_F \times P_F \times (1 - P_T) \tag{5.2}$$

The mDCF is then given by,

$$\text{mDCF} = \min_{\theta \in [0,1]} C_\theta \tag{5.3}$$

In this work, $C_M = 10$, $C_F = 1$, and $P_T = 0.01$ are considered.

5.6 Experimental Evaluation of Proposed EPD in TDSV System

In this section, we present the experimental results for the TDSV systems implemented separately on the two databases, the RSR2015 and the IITG database. The RSR2015 is a clean database, therefore, to examine the robustness of the proposed begin and end point detection algorithm, the data is degraded by adding babble noise.

5.6.1 Performance on the RSR2015 Database Under Clean Speech

The performance of proposed TDSV system is evaluated on the clean RSR2015 database. Table 5.4 presents the performance metrics of the individual algorithms and their combinations used for the begin and end points detection.

For the TD-1 utterance of the RSR2015 database [184], the Energy based and the VLR based method provides an EER of 7.59 %, and 6.79 %, respectively. The proposed method without SDK

Table 5.4: Performances of the TDSV systems using different end point detection methods evaluated on the TD-1 phrase of the clean RSR2015 database. The TDSV systems is built using 39 dimensional MFCCs and DTW, and performance is evaluated in terms of EER and DCF.

Technique		Ground-truth	Energy	FSS	GAD	VLR	VLR+DRF	VLR+DRF	VLR+DRF	VLR+DRF	VLR+DRF+GAD	VLR+DRF+GAD
Method	Metric	end points						+FSS	+GAD	+GAD+FSS	+FSS+OBS	+FSS+OBS+SDK
Clean	EER (%)	**4.54**	7.59	7.3	7.28	6.79	7.59	7.3	6.96	6.2	5.51	**5.37**
	DCF	**0.0307**	0.0388	0.0353	0.0406	0.0361	0.0347	0.0353	0.0407	0.0386	0.038	**0.0341**

$(VLR + DRF + GAD + FSS + OBS)$ provides an EER of 5.51 % and the proposed method with SDK $(VLR + DRF + GAD + FSS + OBS + SDK)$ provides an EER of 5.37 % which is around 2 % improved result compared to the Energy based detection. The ground-truth end points detection gives better performance as compared to the existing techniques. Among the individual methods, VLR based method gives the best performance which is comparable to the performance of the proposed method. This result is expected in case of RSR2015 database, because except the channel effect, there is no other type of degradation present in the database. The telephonic handsets used for collecting the training and testing speech are same for maximum number of speakers and the non-speech regions are mostly silence regions. For such type of speech, the begin and end point detection can be performed accurately without much difficulty. Similar performances are observed for the TD-2 and TD-3 utterances as shown in Fig 5.12.

5.6.2 Performances on the RSR2015 Database Under Degraded Condition

The testing utterances of the RSR2015 database are corrupted with Babble noise of the NOISEX-92 database. Table 5.5 presents the performances of different methods for the TD-1 phrase. As can be seen from the table, the performance of TDSV decreases as the SNR increases. Different individual methods such as, FSS, VLR and GAD show improved performance compared to the Energy based detection in case of all different levels of degradation. The proposed method gives the best performance among all different combinations. On an average, 4 % improvement in EER across five levels of degradation is observed for the proposed method compared to the Energy based method. Experiments are performed for TD-2 and TD-3 phrases and performances are shown in Fig. 5.12. As can be seen from the figure, TD-2 and TD-3 phrases show similar trend as TD-1 phrase in terms of EER.

Table 5.5: Performances of the TDSV systems using different end point detection methods evaluated on the TD-1 phrase of the RSR2015 database under degraded speech condition. The TDSV systems is built using 39-dimensional MFCCs and DTW, and performance is evaluated in terms of EER and DCF.

Technique / SNR	Metrics	Energy	FSS	GAD	VLR	VLR+DRF	VLR+DRF +FSS	VLR+DRF +GAD	VLR+DRF +GAD+FSS	VLR+DRF+GAD +FSS+OBS	VLR+DRF+GAD +FSS+OBS+SDK
20 dB	EER (%)	8.17	7.64	7.3	7.64	8.84	7.37	7.08	8.37	6.96	**6.2**
	DCF	0.0394	0.0411	0.0353	0.0411	0.0381	0.0341	**0.033**	0.0341	0.0407	0.0365
15 dB	EER (%)	9.64	8.45	10.37	10.75	11.36	10.85	8.84	9.59	11.55	**7.64**
	DCF	0.0387	0.034	0.0425	0.066	0.0428	0.0428	0.0381	**0.0337**	0.0428	0.0412
10 dB	EER (%)	11.13	12.43	10.55	12.43	11.55	11.55	11.48	12.6	13.36	**9.49**
	DCF	0.0448	0.0522	0.0604	0.0522	0.0735	0.0649	0.0653	0.0648	0.054	**0.0402**
5 dB	EER (%)	16.57	17.15	16.15	15.99	16.57	16.18	17.89	17.68	17.68	**13.37**
	DCF	0.0688	0.0657	0.0671	0.0646	0.0647	0.0707	0.0725	0.0669	0.0648	**0.0509**
0 dB	EER (%)	20.29	17.94	19.84	16.18	17.17	21.64	17.94	17.73	17.73	**16.47**
	DCF	0.0693	0.0684	0.0694	0.0826	0.0758	0.0781	0.084	0.0948	0.0669	**0.0577**

5.6.3 Performance on IITG Database

Similar to the RSR2015 database, performance of the TDSV system is evaluated on IITG database. The challenge in this case is to discriminate a large set of speakers, considering the testing conditions are quite different from the training conditions. Table 5.6 presents the performance metrics for the three phrases namely, TD-1, TD-2 and TD-3. VLR, FSS and GAD based methods are giving 3-5 % improvement in EER over the Energy based method. The EER is further reduced when different individual methods combined. Approximately 7 % improvement in EER is observed for the proposed method compared to the Energy based method. Robustness of the proposed method in detecting the begin and end points is helping the TDSV system to achieve the improved performance.

5.7 Summary and Conclusions

This work proposes a begin and end point detection algorithm using some speech specific information. The proposed algorithm aims at removing non-overlapping speech as well as non-speech background degradations. The method is vowel-like region detection based, and uses vocal tract and source related information to remove background noise. Dominant resonant frequency information is used to remove non-speech background and foreground speech segmentation is performed to remove the speech background. To detect the consonants at the beginning and at the end of the VLRs, glottal

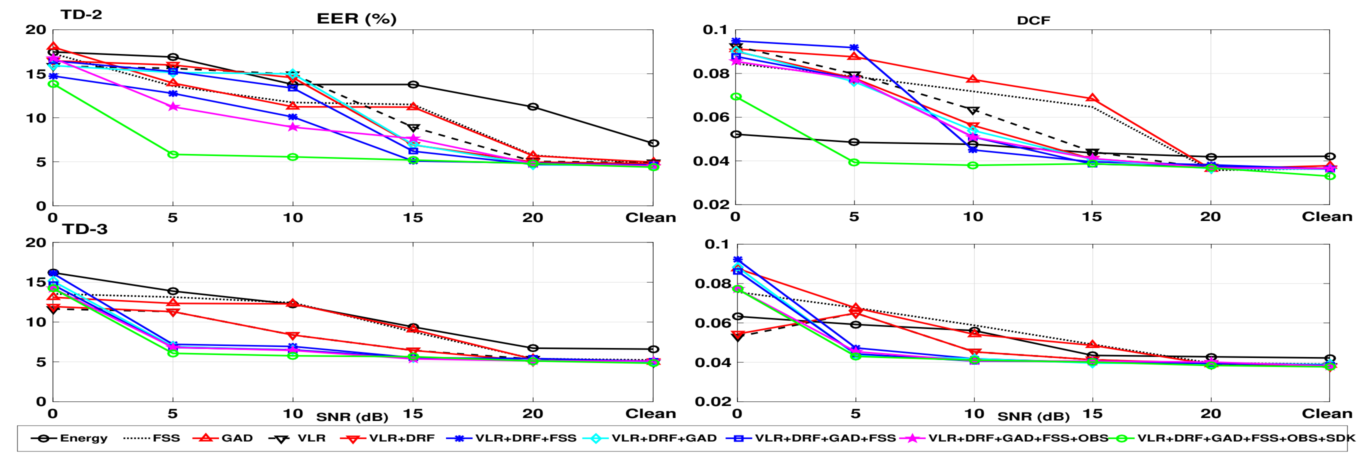


Figure 5.12: Performances of the TDSV systems using different end point detection methods evaluated on the TD-2 and TD-3 phrases of the RSR2015 database.The testing utterances are corrupted by Babble noise, with SNR varying from 20-0 dB in steps of 5 dB. The TDSV systems is built using 39-dimensional MFCCs and DTW, and performance is evaluated in terms of EER and DCF.

Table 5.6: Performances of the TDSV systems using different end point detection methods evaluated on the three phrases of the practical IITG database. The TDSV systems is built using 39-dimensional MFCCs and DTW, and performance is evaluated in terms of EER and DCF.

Technique		Energy	FSS	GAD	VLR	VLR+DRF	VLR+DRF	VLR+DRF	VLR+DRF	VLR+DRF+GAD	VLR+DRF+GAD
SNR	Metrics						+FSS	+GAD	+GAD+FSS	+FSS+OBS	+FSS+OBS+SDK
TD-1	EER (%)	13.19	8.69	9.50	8.64	7.78	8.10	7.70	7.61	6.96	**6.28**
	DCF	0.0422	0.0365	0.0367	0.0363	0.0355	0.0369	0.037	0.0343	0.0353	**0.0339**
TD-2	EER (%)	11.37	8.20	6.69	7.21	6.72	6.56	6.45	6.33	6.36	**6.20**
	DCF	0.0404	0.04	0.0338	0.0391	0.0352	0.0352	0.0343	0.0324	**0.0316**	0.0338
TD-3	EER (%)	18.42	9.74	9.84	10.61	8.35	10.87	9.38	10.53	9.07	**8.20**
	DCF	0.0526	0.0405	0.0395	0.0423	**0.0379**	0.0429	0.0386	0.0422	0.0426	0.0399

activity and dominant aperiodic regions detection are performed. Further, speech duration knowledge is used to remove some spurious VLRs. The proposed algorithm is used to evaluate a TDSV system. The TDSV system with the proposed method shows improved performance than the individual methods such as, Energy, FSS, GAD, and VLR based end point detection.

The proposed end point detection method works when there is an issue of non-overlapping background speech and non-speech noise. In case of overlapping background noise, the temporal and spectral enhancement methods may be useful for robust TDSV system. Future work will be to use temporal and spectral enhancement methods to remove the overlapping background degradation followed by the use of speech specific knowledge for removal of non-overlapping background degradation.

6

Robust Text Dependent Speaker Verification

Contents

Objective

In this chapter, we explore ways to combine different approaches proposed earlier for developing a TDSV system. A speech-specific knowledge based robust EPD technique is used for noise compensation at signal level. Feature level compensation is done by using robust features extracted from HS of the IMFs obtained from MEMD of speech. We also explore a combined temporal and spectral speech enhancement technique prior to the end points detection for enhancing speech regions embedded in noise. All experimental studies are conducted using two databases, namely, RSR2015 and IITG database. It is found that the use of robust end points detection improves the performance of TDSV system compared to the energy based end points detection in both clean as well as degraded speech conditions. Use of noise robust HS features augmented with MFCCs further improves the performance of system. It is also found that the use of speech enhancement prior to signal and feature level compensation results in further improvement in performance for low SNR cases. The final combined system obtained by using the three robust methods provides a relative improvement from 6% to 25% in terms of the EER on the RSR2015 database corrupted with Babble noise of varying strength. Also, the combined system provides 30% to 45% relative improvement in EER on the IITG database.

6.1 Introduction

To deal with degraded conditions, the compensation is done at the signal level, feature level, model level, score level or all of them. Compensation at the signal level involves detection of speech regions. In most of the methods in the literature, the signal is processed to detect the voiced regions by taking evidence from speech/ non-speech frames [19]. Some of the old methods used energy, amplitude, zero-crossing rate, duration, linear prediction error energy, energy based VAD and pitch for detection of speech regions [187]. These methods cannot distinguish between the true speaker's speech and other speaker's speech. Henceforth, statistical modeling methods like HMM, GMM-VAD, Laplacian-Gaussian model and gamma models were also used to detect speech regions during verification [188, 189]. In [60], the authors used GAD for detection of speech end points in TDSV system. Several other works in the literature detected VLRs and used the detected regions for SV [24, 49]. Authors in [51] proposed a method that uses independent processing of VLRs and non-VLRs for achieving better SV performance under clean as well as degraded conditions. One of the recent method uses different speech specific information for robust detection of speech end points [1]. This method used VLRs for detection of speech end points. Some spurious detection in the non-speech noise regions were removed using DRF information [182], whereas some spurious detection in the speech background were removed using a FSS algorithm [102]. The detected end points were further refined using glottal activity and dominant aperiodic region detection. This begin and end points detection method is used in the current work for signal level compensation.

Feature level compensation is done by using noise robust features along with the conventional features. The MFCCs, are considered as the baseline features for various speech processing applications [12, 14, 51]. Most of the state-of-art SV systems also use the MFCCs derived exclusively from the magnitude spectrum of speech utterance, while neglecting its phase spectrum [3]. However, the phase spectrum of speech is equally critical to speech perception as the magnitude spectrum, and has found important use in many speech processing applications [190]. Moreover, the MFCCs may not always constitute the optimum features for all man-machine applications [93, 112]. In fact, with this viewpoint, there have been many alternate features introduced to the Mel filterbank, allowing to improve the performances in speech processing tasks [113]. To improve the feature extraction process, a standard strategy consisting of designing filterbanks using data driven optimization procedure was introduced in [191]. In the same direction, different other approaches based on non-stationary data analysis techniques [120] and *Wavelet Transform* have been found to be useful in different speech processing applications [50]. In [42], an attempt has been made to explore new features for characterizing speakers, obtained from a non-linear and non-stationary data analysis technique called EMD [134], and its variants called MEMD [185]. MEMD is a complete data-adaptive and AM-FM analysis based technique which can decompose any real world speech signal into its oscillatory or AM-FM components called the IMFs. The objective in [42] was to investigate the data-adaptive filterbank nature of EMD/ MEMD that could complement the Mel filterbank in the TISV task. In a recent work [184], we have an investigation on the effect of modifying the process of extracting IMFs with lesser mode-mixing, and better reflecting the higher frequency content of speech. The same MEMD based feature extraction method is used in this work for feature level compensation.

In practical field deployable scenarios, the speech utterances are affected by different degradations like background noise, background speech, speech from other speakers, sensor and channel mismatch, emotional conditions and other environmental conditions, resulting in degraded speech. It has been noticed that the performance of the system falls significantly under such condition, especially when using MFCCs as the features [1, 51, 123, 168, 184]. These observations on the TDSV system highlight the issues related to the development of a system under such conditions and motivate a solution for achieving better system performance. In this work, modification is performed at two stages to improve the system performance under degraded conditions. In the first stage, a signal level compensation methods are used for removing the effect of noise from degraded speech utterances by applying the robust end point detection using speech-specific knowledge to detect the speech regions [1]. Then in the second stage, the detected speech regions are passed through the HS, constructed by using the IMFs obtained from the MEMD of speech utterance.

Further, in this work, we also conduct experiments to combine another spectral processing method with the temporal processing method to obtain better noise reduction for improved system performance. The speech signal is spectrally enhanced using the method described in [28]. The enhanced speech signals are then passed through the robust EPD using speech-specific knowledge to detect the refined begin and end points of the speech utterances [1]. Finally the speaker specific features are extracted from the detected speech regions. The contribution of this chapter is by combining the following three methods. 1) Robust EPD using speech-specific knowledge. 2) Robust features extracted from IMFs of HS obtained from MEMD and 3) Temporal and spectral techniques for speech enhancement. The popular RSR2015 database [14] and IITG database [1, 38, 184] have been used in this work for conducting all the experiments.

The rest of the chapter is organized as follows : Section 6.2 discusses the Robust End Point Detection. Section 6.3 discusses the robust features extracted from HS. Section 6.4 describes the speech enhancement techniques. Section 6.5 describes the experimental setup. Section 6.6 presents the experimental results and analysis. Finally, Section 6.7 summarizes and concludes this work.

6.2 Robust End Point Detection

Prior to extracting the features from the speech signal, robust EPD using speech-specific knowledge is performed to eliminate the silence regions and background noise regions at the beginning and end of the speech utterance. The EPD method is shown in Figure 6.1. The method is based on VLRs detection. The idea is to detect the VLR onset and end points correctly so that begin and end points of speech can be searched near the onset point of first VLR and offset point of last VLR, respectively. Some of the spurious VLRs are removed by using DRF information and a FSS algorithm. Some speech regions at the beginning and end of speech region are detected using GAD and dominant aperiodic region detection.

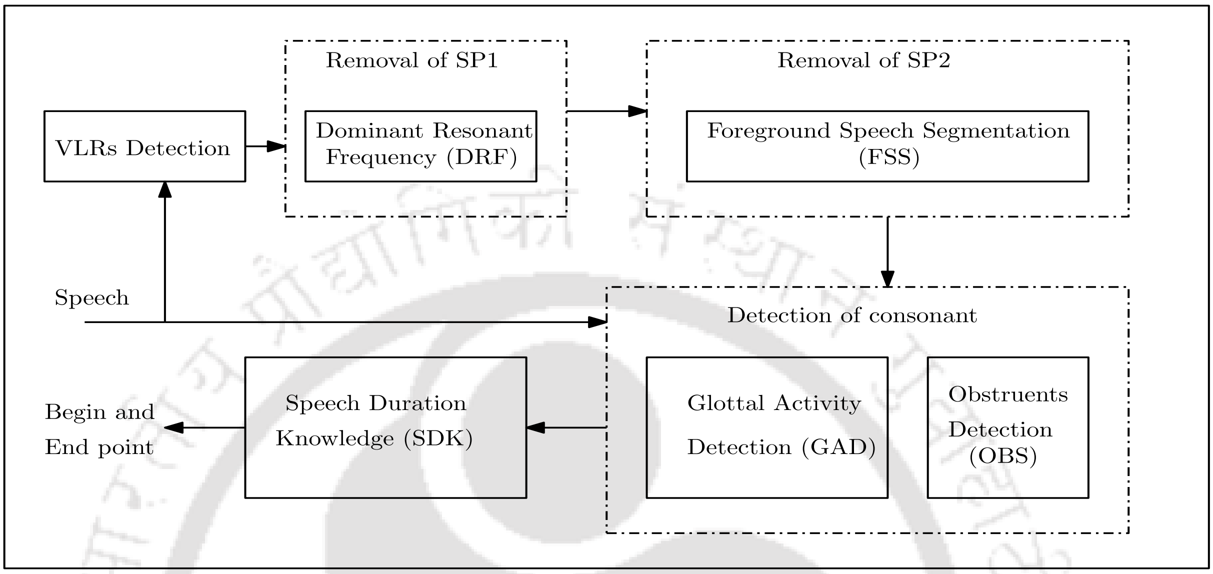


Figure 6.1: Block diagram of the robust EPD. SP1 and SP2 denotes spurious VLRs detected in non-speech and speech background, respectively [1].

6.2.1 Vowel-like region (VLR) detection

Motivation behind the use of VLR for begin and end point detection is the high energy nature of VLRs which makes them high SNR region and therefore they are less affected by noise degradation [49]. VLR detection involves detection of its begin and end points, namely, the VLROPs and VLREPs [24, 51].

The detection of VLROP and VLREP is performed by using excitation source information derived from ZFFS [59] and HE of LP residual of speech [56]. The evidences acquired from these two methods are enhanced by adding amplitude envelope evidence [51] and the begin and end points are detected more accurately.

The final VLROP and VLREP information is derived by adding the two evidences [51]. The combined evidence is then normalized by the maximum value of the sum. The locations of peaks between two successive positive to negative zero crossings of combined evidence represent the hypothesized VLROP or VLREP.

6.2.2 Removal of spurious VLRs in non-speech region using DRF

The vocal tract information are captured from the spectrum in the form of dominant resonances associated with the shape of particular cavity in the vocal tract responsible for the production of speech segment. These resonance peaks are called DRF. DRF is the frequency which is resonated most by the vocal tract. DRF is computed from the Hilbert envelope of numerator group delay spectrum of zero time windowed speech [179]. For VLRs, DRF value is mostly less than 1 kHz and the non-speech noises mostly contain high-frequency components. This knowledge is used for identifying and removing the spurious VLRs in the non-speech region [1]. The VLRs having DRF more than 1 kHz are removed from the output of VLR detection.

6.2.3 Removal of spurious VLRs in background speech using FSS

To remove the background speech, a FSS is used. FSS algorithm was proposed in [180], which was further modified in [102]. In this chapter, the modified version of FSS method is used to remove the background speech [1]. The method uses both source and system information. Excitation source information is extracted using ZFFS analysis and vocal tract system features extracted from the modulation spectrum energy. The spurious VLRs in the background speech region are removed and the VLRs present in the foreground speech region are retained.

6.2.4 Glottal activity and obstruents detection

Once the VLRs are detected, GAD can be explored to add a sonorant consonant at the begin or at the end of speech utterance. The GAD method proposed in [181] is used in this work. GAD detects VLRs as well as sonorant consonant regions. Therefore, it enables better localization of voiced regions and helps to minimize the misses in the VLR detection output. This, in turn, helps to detect the appropriate end points of speech utterance.

There may be an obstruent consonant at the begin and end of speech utterance. To include them in the speech region, obstruent detection is performed. Since the aperiodic component is dominant in burst and frication region of most of the obstruents, the dominant aperiodic region detection method proposed in [182] is used to detect the obstruents. The first VLROP and last VLREP are considered as the refined begin and end points and are detected with more accuracy after using obstruent detection.

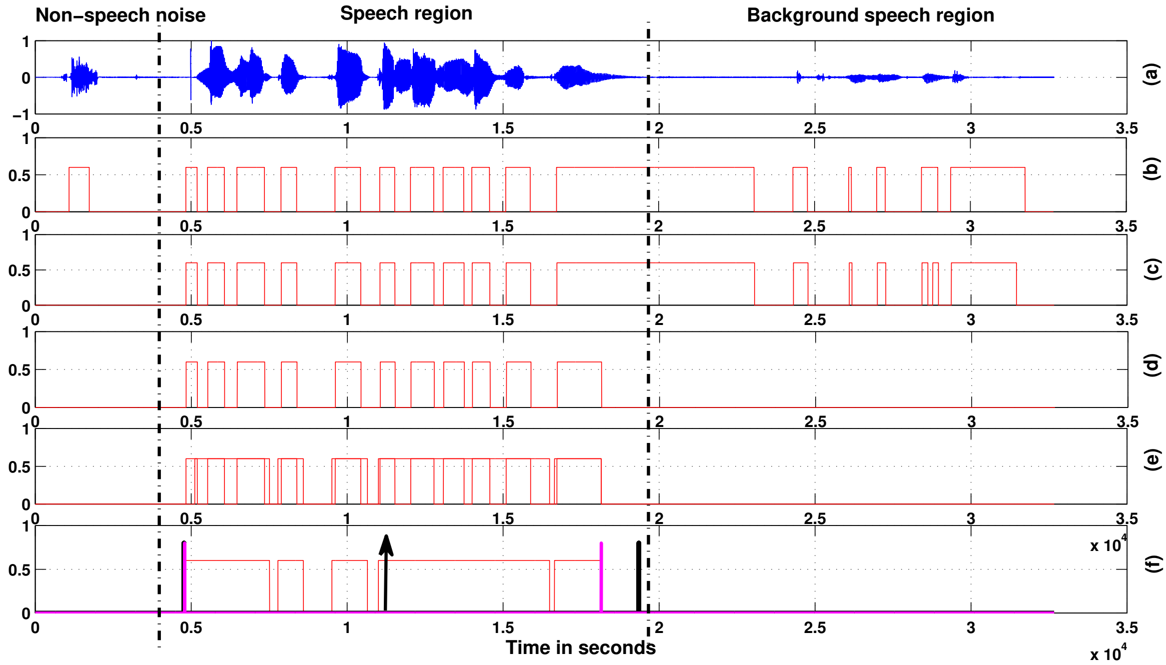


Figure 6.2: Illustration of the begin and end point detection procedure. (a) Speech signal containing non-overlapping background noise and background speech. (b) Detected speech regions of VLROP and VLREP are used to obtain the VLRs. (c) VLRs after discarding the background noise using DRF information. (d) VLRs after discarding the background speech using FSS. (e) Detected GADs added to the VLRs. (f) Refined begin and end points using obstruent information. The center C of the speech utterance shows the arrow around 1.2 sec. The duration between successive VLRs are detected, either less than 300 msec or the peripheral VLR is reached. Finally, the VLROP of first VLR and VLREP of last VLR are detected using SDK knowledge. Dotted line shows the manually marked VLROP and VLREP points in the speech region.

6.2.5 Speech duration knowledge for further refining the end points

Finally, SDK is used to further remove the spurious VLRs. In many cases, one may repeat the utterance twice and can talk something which is not part of speech utterance. In these situations, it is not feasible to make any rectifications. However, if the user uttered at the begin or end leaving some silence region between actual speech utterance and extra word. Then it is viable to discard the extra word using SDK. The SDK method for refining begin and end point detection is as follows:

First, identify all the locations of VLROPs in the speech utterance and compute the average of Euclidean distances from one VLROP to all other VLROPs. The VLROP having the minimum average Euclidean distance is marked as the center. Then, starting from the center of the speech, the duration between two successive VLRs are computed on either side until the peripheral VLR or the duration between two successive VLRs is greater than 300 ms [49]. At this point among the two VLRs, the one which is closer to center is declared as the peripheral VLR. All VLRs outside the peripheral VLRs are removed. Finally, the VLROP of first VLR and VLREP of last VLR are declared as the begin and end points of the speech utterance, respectively [1]. Figure 6.2 shows speech signal for the utterance "Lovely pictures can only be drawn" with background noise include both speech and non-speech background noise. Figure 6.2 (a) shows the speech utterance with non-overlapping background noise and background speech. Figure 6.2 (b) shows the VLROP and VLREP evidences and detected VLRs, respectively. The non-speech background noise present in between 0-0.5 sec and background speech present in between 2.25-3.5 sec is also detected as VLRs due to impulse-like characteristics. Second step is to remove such missed spurious VLRs, DRF information is used. If DRF is less than 1 kHz then consider as VLRs and if more than 1 kHz then considered as non-speech region. Later remove non-speech region and then refined VLRs are shown in Figure 6.2 (c). The third step is to remove the background speech which is obtained using FSS. With the help of FSS, the detected spurious VLRs in the background speech region are removed, and refined VLRs are obtained after FSS are shown in Figure 6.2 (d). In the next step, the glottal activity regions are explored to add a sonorant consonant at the begin or at the end of speech utterance for better localization. The detected GADs are obtained for appropriately detecting the begin and end points are shown in Figure 6.2 (e). The fifth step is to detect the obstruent consonant at the begin and end of speech utterance. The refined begin and end points are detected with more accuracy using obstruent evidences are shown in Figure 6.2 (f). The arrow in Figure 6.2 (f) shows the center of detected VLRs of speech utterance. Starting from the center of speech, the duration between two successive VLRs are computed on either sides until either the duration is found to be greater than 300 msec or the peripheral VLR is reached. Thereafter, no further modification is incorporated using SDK knowledge. The dotted lines on either sides show ground truth manual marking.

6.3 Robust Features From Hilbert Spectrum of MEMD

The EMD is a data-adaptive technique, which can decompose the speech signal into a finite number of components, called IMFs, without the need of any *a priori* basis [93]. Every speech signal has its unique and meaningful decomposition. Again, the signal changes dynamically and varies with resonant structure of the speech signal. The changed resonant structure distributed among its unique set of IMFs, which are obtained without any *a priori* basis. MEMD explored in different real-world applications [42, 138, 141]

6.3.1 Distribution of mode-mixing for MEMD

The mode-mixing makes the IMFs less narrowband, leading to less accurate estimation of their instantaneous frequencies and amplitude envelopes. Moreover, this makes it difficult to segregate or characterize a certain subset of the IMFs, as being useful for analysis, for a particular task. Hence, in this work, we utilize a recently proposed variants of EMD - the MEMD -which reduces mode-mixing in the IMFs [93].

Practically, the decomposition is stopped when a user-defined maximum M number of IMFs, has been extracted. For a digital speech signal, $s(n)$.

$$s(n) = r_M(n) + \sum_{i=1}^{M} h_i(n) = \sum_{i=1}^{M+1} h_i(n) \tag{6.1}$$

where, $h_i(n)$ represents the decomposition of the signal in its IMFs, and $r_M(n)$ the final residue, which is a trend-like signal and M is the total number of IMFs extracted. In Figure 6.3, the first 5 IMFs obtained from MEMD of a digital speech signal $s(n)$, taken from the RSR2015 database. This represents low and high frequency oscillations present at different instants within the same IMF or distributed among multiple IMFs. This process is called *mode-mixing*. These components spread across different IMFs at different instants of time, leading to a less accurate number of extrema and number of zero crossings differ by utmost one. The reliable IMFs of MEMD may be used to be better suited for AM-FM analysis. To avoid unnecessary generation and processing of higher-order IMFs, the decomposition reduced to a maximum of 10 components ($M = 9$), for the MEMD method [184].

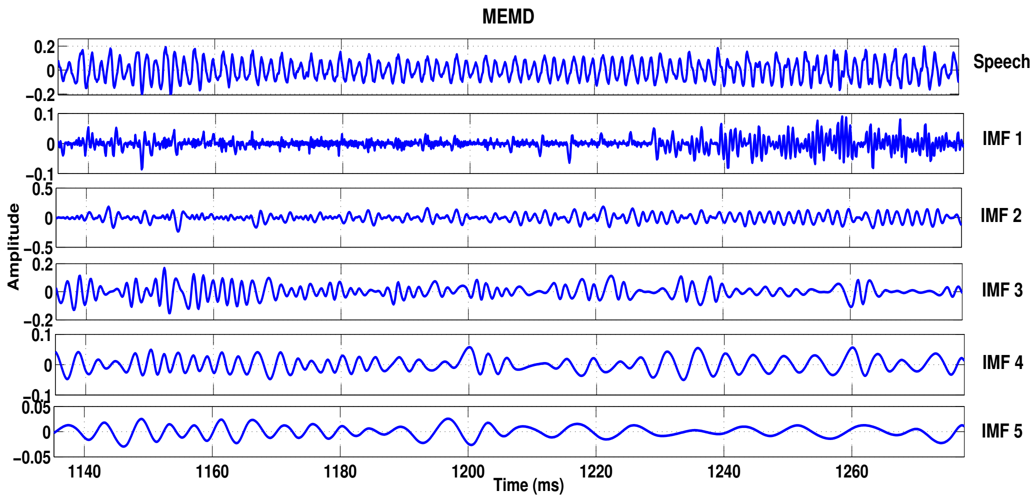


Figure 6.3: IMFs 1-5 generated from a speech signal, $s(n)$, using MEMD.

6.3.2 Dynamic changes of instantaneous frequencies and amplitude envelopes

Having obtained the IMF components whose center frequencies decreases as the order of IMF increases. The Hilbert transform is applied to each component to compute dynamically changing instantaneous frequency and amplitude envelope. The Hilbert transform, $H[x(t)]$, of a signal $x(t)$, is computed from Fourier transform.

$$x(t) \longleftrightarrow X(f) \ , \ \frac{1}{\pi t} \longleftrightarrow -j \ sgn(f) = \begin{cases} -j \ , \ f > 0 \\ j \ , \ f < 0 \end{cases} ,$$

$$H[x(t)] = x(t) * \frac{1}{\pi t} \longleftrightarrow -j \ sgn(f)X(f) = \begin{cases} -jX(f) \ , \ f > 0 \\ jX(f) \ , \ f < 0 \end{cases}$$

The instantaneous frequency function, $f(t)$, and amplitude envelope function, $A(t)$, is derived from the analytical signal, $x_A(t)$, which is free from any negative frequency Fourier components.

$$A(t) = |x_A(t)| \ , \ \phi(t) = \arctan \frac{\Im\{x_A(t)\}}{\Re\{x_A(t)\}} \ , \ f(t) = \frac{1}{2\pi} \frac{d\phi(t)}{dt} \tag{6.2}$$

Correspondingly, the DFT method [3] is used for estimating the instantaneous frequency and amplitude envelope of any discrete-time signal, $x(n)$ [3]. Henceforth, if $A_k(n)$ and $f_k(n)$ represent the amplitude envelope and instantaneous frequency of $h_k(n)$, respectively. The time-frequency distribution of energy envelope is the squared magnitude of amplitude envelope. This formulation, when represented in a complete, compact and adaptive Fourier representation in terms of an image, is called the HS [134, 192].

$$H(f,t) = \left\{A_k^2(n) \mid f_k(n), n\right\} \ , \ k = 1, ..., K \leq M + 1 \tag{6.3}$$

Figure 6.4 represents the Hilbert spectra of speech signal used in Figure 6.3, obtained using MEMD, where fixed number of components ($K = M + 1 = 10$) are used. Figure 6.4 shows the HS for a section of the speech utterance of last few components, which are low-frequency trend-like wave forms, which are excluded from the spectrum, as they have high energy and obscure the image. This is evident from the spectrum, most of the energy in the spectrum lies within 100-700 Hz (particularly, one can observe more dots), which is the pitch frequency range, i.e., the frequency range of vibration of vocal folds in the glottis (during the production of voiced speech) [93]. This spectrum can be post-processed to obtain instantaneous pitch frequency, which constitutes the HS, obtained from MEMD, for carrying

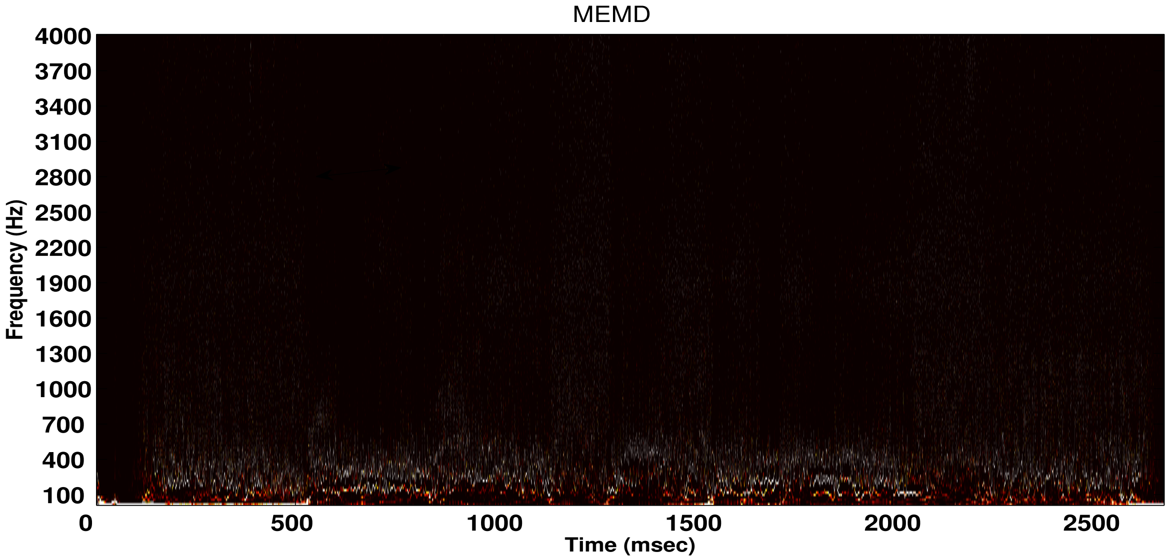


Figure 6.4: Hilbert spectrum of the speech signal used in Figure 6.3, constructed from the IMFs of MEMD.

speaker-specific information [184].

6.3.3 Characteristics of HS and MFCC features

The Mel filterbank has a fixed structure whereas, the MEMD filterbank has an adaptive structure at every frame, both capture speaker-specific information using their specific methods. The HS represents the instantaneous frequencies and amplitude envelopes of IMFs, are processed in short time segments to generate features. These features capture different sort of speaker information for characterizing the speakers.

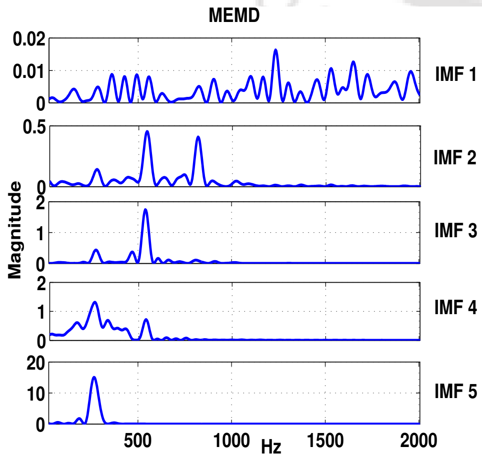


Figure 6.5: Squared magnitude spectra of each of IMFs 1-5 generated by MEMD of a 20 msec segment of the speech signal used in Figure 6.3.

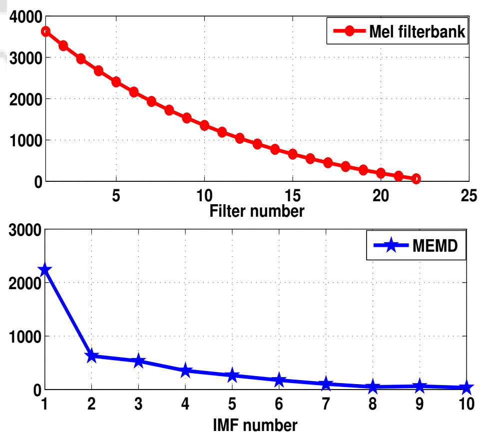


Figure 6.6: (a) Center frequencies of a 22-filter Mel filterbank ; (b) Mean frequency of the IMFs derived from the speech signal of Figure 6.3.

Figure 6.5 depicts the power spectra of first 5 IMFs, corresponding to a 20 msec segment of speech signal of Figure 6.3. As is evident from the figure, the power spectra of IMFs represent different portions of speech spectrum, as if they have been band pass filtered. However, in this case the entire process takes place in the time domain to manifest the characteristics of an adaptive filterbank. To illustrate the difference between this adaptive and the Mel filterbank, the centre frequencies of a 22 filter Mel filterbank are plotted, in Figure 6.6, along with the *mean frequencies* of IMFs [185]. For a 20 msec segment of speech signal, $s(n)$, the mean frequency of corresponding segment of its k^{th} IMF is obtained as,

$$F_k^m = \sum_{f=0}^{F_s/2} \frac{f \times S_k(f)}{\sum_{f=0}^{F_s/2} S_k(f)} \ , \ k = 1, ..., M+1 \ , \tag{6.4}$$

where, $F_s = 8$ kHz is the *sampling frequency* of $s(n)$, and $S_k(f)$ is the power spectrum of 20 msec segment of k^{th} IMF. f represents the analog frequencies corresponding to the digital frequencies of the DFT spectrum of IMF. From Figure 6.6, the MEMD and Mel filterbank are different. The center frequencies are changes as the nature of signal changes [133].

6.3.4 Instantaneous frequencies and energies of the Hilbert spectrum

The features extracted from HS constitutes the instantaneous energy envelopes and instantaneous frequencies (derived from the instantaneous phases) of the IMFs. For visual clarity, Figure 6.7 shows the first 4 IMFs, corresponding to a 20 msec segment of speech from Figure 6.3 is constructed. Thus, the instantaneous energy values are normalized at each time instant, and the frequencies are represented in kHz to reduce the dynamic range. Hence, it is beneficial to extract features from them, after some degree of averaging over the time segment. This reduces the feature space, and allows the features derived to be concatenated with the MFCCs, which are obtained for every frame, after dividing the entire speech utterance into overlapping frames.

However the instantaneous frequencies and energies need to be redistributed. If no mode-mixing is present, then, at every speech frame, the first IMF, $h_1(n)$, would produce the highest instantaneous frequency. The second highest frequency would be given by $h_2(n)$, the third highest by $h_3(n)$, and the fourth highest by $h_4(n)$. Based on this, the four instantaneous frequencies at every time instant are sorted in descending order of frequency. These sorted instantaneous frequencies and energies are then rearranged. Figure 6.8 shows the rearranged IMFs as shown in Figure 6.7, contain speaker-specific

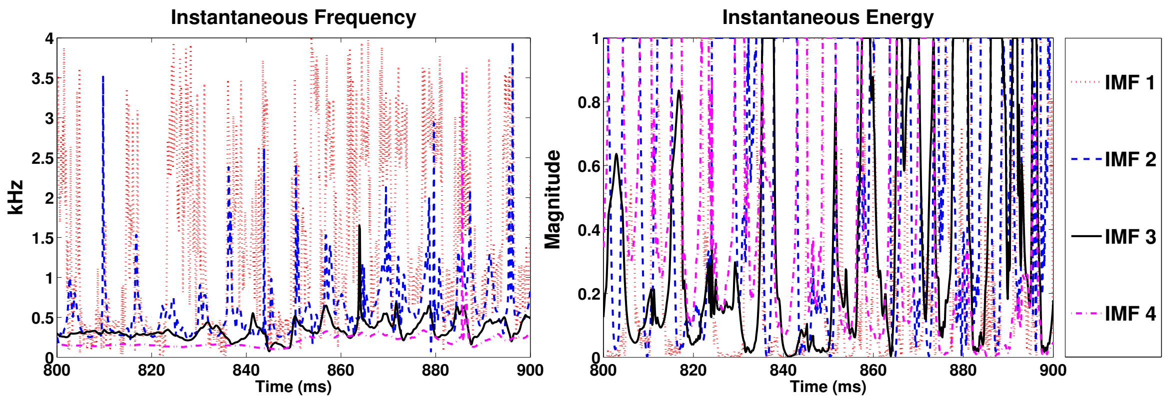


Figure 6.7: Instantaneous frequencies (kHz) and normalized Instantaneous energies of the first 4 IMFs (derived using MEMD) corresponding to the HS shown in Figure 6.4.

information, which may be used for extracting features for the task of TDSV.

6.3.5 Robust feature extraction from the instantaneous frequencies and energies

Based on the preceding discussion, different features are extracted from the IMFs of speech. Let $K \leq M+1$ be the number of IMFs from which the instantaneous frequencies, $\{f_k(n)\ ,\ k=1,...,K\}$, and the instantaneous energies, $\{|A_k(n)|^2\ ,\ k=1,...,K\}$ are extracted. From 20 msec frame size and with a frame shift of 10 msec, $s^j(n)$ represent the j^{th} frame of speech signal $s(n)$. Then, $h_k^j(n)$ represents the j^{th} frame of the k^{th} IMF of $s(n)$. Correspondingly, $f_k^j(n)$ and $|A_k^j(n)|^2$ are the instantaneous frequencies and energies, respectively, of $h_k^j(n)$, extracted using HS. Let N_f be the number of samples in a 20 msec frame. The following features are extracted from each speech frame. It

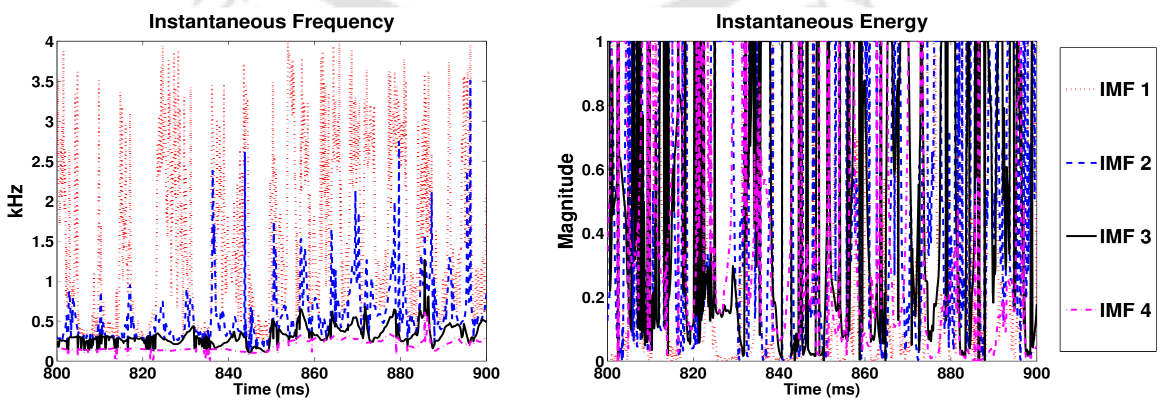


Figure 6.8: Instantaneous frequencies (kHz), sorted in decreasing order, and the corresponding rearranged normalized instantaneous energies of the first 4 IMFs (derived using MEMD) corresponding to Figure 6.7.

has been shown that the first few IMFs of the speech signal show the vocal tract resonances of the speech utterance. In [184], the IMFs obtained from HS have been utilized to estimate the vocal tract resonances produced by the different cavities that are formed when the speaker utters. These cavities depends on the sound produced and physical structure of the vocal tract, which never be same for two different speakers. Hence, the first few IMFs are used to carry important speaker-specific information. More specifically, HS constitutes different frequency bands and their corresponding energy bands to represent useful speaker characteristics that can complement the MFCCs.

Mean Instantaneous Frequency: It is derived as,

$$F_k^j = \frac{1}{N_f} \sum_{n=0}^{N_f-1} f_k^j(n)/1000 \ , \ k = 1, ..., K \leq M+1 \tag{6.5}$$

$$F_K^j = [F_1^j, F_2^j, ..., F_K^j]^T \ , \quad F_K = \{\Gamma\{F_K^j\}, \forall j \in \mathbb{N}\} \tag{6.6}$$

The F_K feature is used to capture the dominant frequencies of the different frequency bands of the IMFs obtained from MEMD. This feature may be expected to carry speaker-specific cues for characterizing the speakers.

Absolute Deviation of Instantaneous Frequency: It is derived as,

$$\Delta F_k^j = \frac{1}{N_f} \sum_{n=0}^{N_f-1} |\frac{f_k^j(n)}{1000} - F_k^j| \ , \ k = 1, ..., K \leq M+1 \tag{6.7}$$

$$\Delta F_K^j = [\Delta F_1^j, \Delta F_2^j, ..., \Delta F_K^j]^T \ , \ \Delta F_K = \{\Gamma\{\Delta F_K^j\}, \forall j \in \mathbb{N}\} \tag{6.8}$$

For a speech signal, some frequency bands show large variations, whereas other frequency bands show comparatively steady in the HS. The ΔF_K feature is used to capture these variations, spreads or widths of the first K frequency bands arranged in the decreasing order of frequencies. This ΔF_K feature might be useful in characterizing the speaker information in a better way.

Correlation between Instantaneous Frequencies: It is derived as,

$$\sigma F_k^j = \sum_{n=0}^{N_f-1} [\ \{f_k^j(n) - F_k^j\}\{f_{k+1}^j(n) - F_{k+1}^j\}\ ] \ , \ k = 1, ..., K-1 \leq M \tag{6.9}$$

$$\sigma F_{K-1}^j = [\sigma F_1^j, \sigma F_2^j, ..., \sigma F_{K-1}^j]^T \ , \sigma F_{K-1} = \{\Gamma\{\sigma F_{K-1}^j\}, \forall j \in \mathbb{N}\} \tag{6.10}$$

There is a possibility that certain frequency bands may vary in a particular manner. On the other hand, a different pair of nearest frequency bands may be closely related with each other. For such a speech utterance, σF_{K-1} feature capture the dependence or correlation between successive frequency bands of the HS. The K frequency bands and their $K-1$ successive frequency bands correlation values are obtained. Therefore, σF_{K-1} inter-band relations could serve as an useful speaker-specific cues.

Mean Instantaneous Energy: It is derived as,

$$|A_k^j(n)|^2 := \frac{|A_k^j(n)|^2}{\underset{k}{max}\ |A_k^j(n)|^2}\ ,\ k = 1, ..., K \leq M+1 \tag{6.11}$$

$$E_k^j = \frac{1}{N_f} \sum_{n=0}^{N_f-1} |A_k^j(n)|^2\ ,\ k = 1, ..., K \leq M+1 \tag{6.12}$$

$$E_K^j = [E_1^j, E_2^j, ..., E_K^j]^T\ ,\ E_K = \{\Gamma\{E_K^j\}, \forall j \in \mathbb{N}\} \tag{6.13}$$

The E_K feature denotes the average energy (amplitude envelope) of the different frequency bands. For the first K frequency bands there are equivalent number of energy bands. Therefore, E_K represents the mean value of these K energies at different frequencies. Hence, E_K feature could be useful in discriminating the speakers.

Absolute Deviation of Instantaneous Energy: It is derived as,

$$\Delta E_k^j = \frac{1}{N_f} \sum_{n=0}^{N_f-1} \mid \{|A_k^j(n)|^2 - E_k^j\} \mid\ ,\ k = 1, ..., K \leq M+1 \tag{6.14}$$

$$\Delta E_K^j = [\Delta E_1^j, \Delta E_2^j, ..., \Delta E_K^j]^T\ ,\ \Delta E_K = \{\Gamma\{\Delta E_K^j\}, \forall j \in \mathbb{N}\} \tag{6.15}$$

A particular speaker can emit certain frequencies at varying strengths, whereas other speaker can only emit other frequencies at specific strengths. One can observe large energy variations in certain frequency bands (energy bands), whereas other frequency bands (energy bands) may be relatively steady. Therefore, ΔE_K feature represents the variation of energy in different frequency bands. These cues may be serve as an important speaker specific cue.

Correlation between Instantaneous Energies: It is derived as,

$$\sigma E_k^j = \sum_{n=0}^{N_f-1} \left[\ \{|A_k^j(n)|^2 - E_k^j\}\{|A_{k+1}^j(n)|^2 - E_{k+1}^j\}\ \right]\ ,\ k = 1, ..., K-1 \leq M \tag{6.16}$$

$$\sigma E_{K-1}^j = [\sigma E_1^j, \sigma E_2^j, ..., \sigma E_{K-1}^j]^T\ ,\ \sigma E_{K-1} = \{\Gamma\{\sigma E_{K-1}^j\}, \forall j \in \mathbb{N}\} \tag{6.17}$$

The σE_{K-1} feature captures the relation between successive frequency bands in terms of their energy variations. Hence, the increase in energy in a particular energy band affects its succeeding energy band. Therefore, these energy variations may be speaker specific, and hence serve as an important feature for each speaker.

Instantaneous Energy weighted Instantaneous Frequency: It is derived as

$$\Upsilon_k^j = \sum_{n=0}^{N_f-1} \frac{|A_k^j(n)|^2}{\sum_{k=0}^{K} |A_k^j(n)|^2} f_k^j(n) \ , \ k = 1, ..., K \leq M+1 \tag{6.18}$$

$$\Upsilon_K^j = [\Upsilon_1^j, \Upsilon_2^j, ..., \Upsilon_K^j]^T \ , \ \Upsilon_K = \{\Gamma\{\Upsilon_K^j\}, \forall j \in \mathbb{N}\} \tag{6.19}$$

For different speakers, the correlation may vary. To capture such correlation Υ_K feature merges the information contained in instantaneous frequencies and energies. The first K frequency bands (in decreasing order of frequency),and their corresponding K energy bands can be weighted. Therefore, the Υ_K feature may be expected to capture speaker-specific cues [184].

Figure 6.9 represents the block diagram of the feature extraction process used for the TDSV system. The features are extracted from the instantaneous frequencies and energies. Along with HS features, the conventional 39 dimensional MFCCs, 51 dimensional Ext. MFCCs, and the cG_K [42, 133, 184] features are also derived from every speech frame. These features are used with the expectations that their combinations with MFCCs would significantly enhances the capability of the TDSV system.

6.4 Speech Enhancement Techniques for TDSV System

The main concern of the work presented in this section is to use a combined temporal and spectral enhancement method for enhancing speech under degraded conditions. This method can be successfully used for identifying and enhancing speech-specific components from the degraded speech. The temporal processing method involves identification and enhancement of speech-specific components present at the gross and fine levels [28]. The evidences obtained by using gross level components are as follows: first the gross level components are used to detect high SNR regions using the sum of the 10 largest peaks in the DFT spectrum which represents the vocal tract information. The second evidence is obtained from the smoothed HE of LP residual of speech representing the excitation source information. The third evidence is the modulation spectrum which represents the suprasegmental information of speech. The origin of these three approaches are different and hence combining them together improves the robustness and detection accuracy as compared to individual processing methods [28]. The gross weight function $w_g(n)$ computed from these three evidences are summed up together, normalized and non-linearly mapped using the mapping function [28].

$$w_g(n) = \frac{1}{1 + e^{-}\lambda(s_i(n) - T)} \tag{6.20}$$

where, λ is the slope parameter and $w_g(n)$ is the non-linearly mapped values of the normalized sum $s_i(n)$ and T is the average value of $s_i(n)$. The gross weight function is obtained by computing the deviation between spectrally processed speech with respect to direct speech.

The fine level components are identified using the knowledge of the instants of significant excitation which mostly correspond to the epoch locations [28]. Therefore, using HE of LP residual one can extract the robust epoch locations from the speech signal. From HE of the LP residual perspective, an approximate location of instants is sufficient because the enhancement is commonly achieved by emphasizing the residual signal in the speech regions around the instants of significant excitation. The speech regions around the instants of significant excitation are used as fine level evidence. The epoch locations are used for obtaining the fine weight function [28]. The region around the instants of significant excitation are convolved with a Hamming window which has a 3 ms temporal duration. Therefore, the fine weight function is derived as

$$w_f(n) = (\sum_{k=1}^{N_K} \delta(n - i_k)) * h_w(n) \tag{6.21}$$

where, N_k is the total number of epochs located in speech frame, i_k is estimated location of speech.

A total weight function $w(n)$ is obtained by multiplying the gross weight function $w_g(n)$ with the fine weight function $w_f(n)$ which is represented as

$$w(n) = w_g(n) * w_f(n) \tag{6.22}$$

The temporally processed speech can be obtained by synthesizing as follows:

$$S_k(z) = \frac{R_w(z)}{1 + \sum_{n=1}^{M} a_n z^{-n}} \tag{6.23}$$

where, $S_k(z)$ represents the temporally processed speech and $R_w(z)$ is the weighted LP residual used to excite the time-varying all-pole filter derived from the degraded speech to generate the enhanced speech and a_n is the LP filter coefficients.

The temporally processed speech is further subjected to improve the vocal-tract characteristics at the spectral domain. Whereas the temporal processing approach enhances the speech region around the instants of significant excitation, spectral processing approach enhances speech-specific components and suppresses the noise components in the spectral domain.

The short-term magnitude of the degradation and degraded speech are estimated. The MMSE-LSA estimator is applied to the magnitude spectra for obtaining the enhanced spectra from the degraded speech [103]. The spectral gain function for the MMSE-LSA estimator is expressed as follows:

$$H(n) = \frac{\xi_n}{1+\xi_n} exp(\frac{1}{2}\int_{\upsilon_n}^{\infty}\frac{e^{-x}}{x}dx) \tag{6.24}$$

where

$$\upsilon_n = \frac{\xi_n}{1+\xi_n}\gamma_n \tag{6.25}$$

where ξ_n and γ_n are *a priori* SNR and *a posteriori* SNR, respectively.

The enhanced magnitude spectra and degraded phase spectra are then combined to produce an estimate of clean speech and the overlap-add method is used for the re-synthesis in the time domain. The re-synthesized speech is the enhanced speech and this enhancement may improve the TDSV system performance under degraded speech and challenging test conditions.

6.5 Experimental Setup

The proposed work is directed towards addressing issues related to practically deployable systems. In order to evaluate the robustness of the proposed framework under degraded speech conditions, we conducted the TDSV system experiments on two databases, namely, the RSR2015 database [14] and the IITG database [1, 38, 105, 183, 184]. The databases are described below.

6.5.1 Databases

The RSR2015 database mainly considered for TDSV with pre-defined speech utterances as described in the previous chapter. To evaluate the proposed framework under degraded speech condition, the clean speech utterances are corrupted with Babble noise to generate the noise mixed speech [110]. The energy level of the noise is scaled such that overall SNR of the degraded speech files is maintained at 20-0 dB in steps of 5 dB. The system performance is evaluated on clean and synthetically degraded RSR2015 database.

The IITG database was collected from the undergraduate and post-graduate course students with a speech biometric-based attendance system as an application to address the issues in the practical deployment of TDSV system as described in the previous chapter [1]. During the enrollment phase, three pre-defined speech utterances were recorded for each enrolled course student in an anechoic chamber. During the collection of testing database, the course students could move freely within, and in and out of the open hall in the department office, in the corridor, in front of the classroom and also in the free environment. The data collection process was more practical because the database includes background noise, background speech, and other environmental conditions. The recorded data contains mismatches between the enrollment and the testing phases, in terms of the sensors, mobile handsets, style of speech and environment. Due to these conditions, IITG database becomes more challenging. Henceforth, there is no compelling reason to add artificial noise in the IITG database for the evaluation of the TDSV system.

6.5.2 Feature extraction

During the training and testing process, the given speech signal is processed in frames of 20 msec duration at a 10 msec frame shift. Prior to extracting the features, EPD is performed to eliminate the background speech/background noise regions at the beginning and end of the given speech utterance. The speech utterance between the detected end points are considered and used for the feature extraction. For each 20 msec frame size Hamming windowed speech frame, MFCCs are computed using 22 logarithmically spaced Mel-filterbanks. The features extracted from every frame are stacked and then CMVN is performed, for every feature dimension. This extracted feature set, is then used for the task of TDSV system, using the technique of DTW [1,3,94,184].

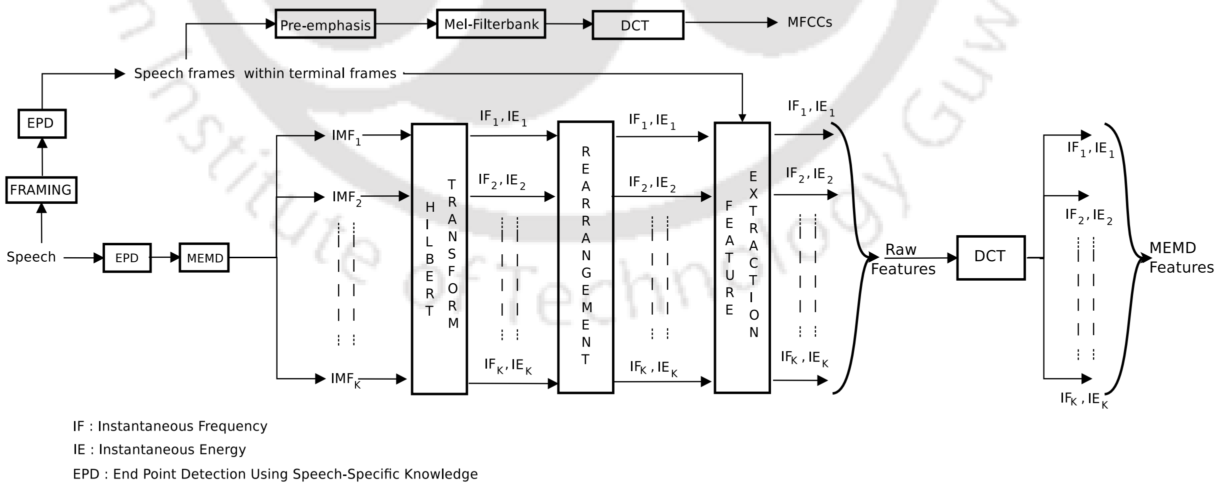


Figure 6.9: Feature Extraction Process for the TDSV system

In this work, the seven different feature sets obtained from IMFs derived from HS of a given speech utterance are used independently, and in combination with the feature set corresponding to 39 dimensional MFCCs, for the TDSV task, using DTW. The dimensions of each of the 7 feature sets are varied, by varying the number of IMFs from $K = 10$ to 2, in steps of 2, to observe the importance of each IMF. This can be done in order to find out the range of IMFs which are useful for the TDSV task. From every speech frame, two different sets of MFCCs are obtained. For extracting the MFCCs from a speech frame, the speech frame is first passed through the pre-emphasis filter, and then 22 logarithmically spaced filters (Mel filterbank) are applied on its DFT spectrum. Then, 39 dimensional MFCCs and an extended 51 dimensional MFCCs (Ext. MFCCs) feature vector are obtained from each speech frame. The standard 39-dimensional MFCCs feature vector comprises of the first 13 cepstral (excluding the 0^{th} coefficient), the first 13 Δ cepstral, and the first 13 $\Delta\Delta$ cepstral coefficients. The Ext. MFCCs feature vector comprises of the first 17 cepstral (excluding the 0^{th} coefficient), the first 17 Δ cepstral, and the first 17 $\Delta\Delta$ cepstral coefficients. Comparing the system performances of the Ext. MFCCs with that of the 39-dimensional MFCCs enables us to notice how useful the higher dimensions of the MFCCs are in the TDSV system task. Also, it empower us to ascertain the combinations of the seven different feature sets obtained from HS, with the 39-dimensional MFCCs. Apart from the aforementioned feature sets, the cG_K, is obtained from each speech frame using its raw IMFs of the given speech signal as described in [93]. However, considering this enormity of our database, we derive the cG_K feature for every frame of the speech. Further, we refine the feature vector using DCT [42].

6.5.3 Performance metrics

The performance of proposed TDSV system is evaluated on the RSR2015 and the IITG databases, based on two standard metrics- EER and mDCF as described in the previous chapter [1, 184].

6.6 Results and Analysis

In this section, we present the experimental results and analysis, in terms of EER and mDCF, for the TDSV system implemented separately on the RSR2015 and the IITG databases.

6.6.1 Performances obtained when robust EPD followed by extraction of MEMD features

The performances of the seven HS features derived from MEMD, the conventional 39 dimensional MFCCs features, the Ext. 51 dimensional MFCCs, the cG_K features and their different combinations are evaluated. The effectiveness of the robust EPD algorithm with different features is shown by comparing the performance with the energy based EPD.

6.6.1.1 Performances on the RSR2015 database under clean conditions

The performance metrics of the proposed methods for the TDSV system is evaluated on the clean RSR2015 database. Table 6.1 presents the performance metrics for each of the seven extracted experimental features, derived using all the 10 IMFs. For comparison, the performance metrics for MFCCs, Ext. MFCCs and cGk features are also shown. As it is observed from the table, for all the three speech utterances, the performance for MFCCs are better than the rest of the features. For the TD-1 speech utterance, the 39 dimensional MFCCs provide an EER of 6.37 %. The 51 dimensional Ext. MFCCs shows improvement as compared to the standalone MFCCs, with an EER of 4.93 %. This improved performance of Ext. MFCCs over 39 dimensional MFCCs in TDSV task was observed in one of our previous energy based EPD work reported in [184]. In Table 6.1, the numbers shown in the brackets are the results obtained by using energy based EPD and the numbers shown without brackets are the results obtained by using the speech specific knowledge based robust end points detection method. It can be seen that the performance for the robust end points detection method is significantly better than the energy based method. For all the three utterances, standalone MFCC features and Ext. MFCCs are far better than cGk and seven experimental HS features. Having observed the performances of the standalone features, we need to proceed towards combining them with the MFCCs and evaluating the effect of the combinations on the TDSV system.

Table 6.2 presents the performance metrics of the feature combinations for the TD-1 sentence. The seven features extracted from HS and cG_K feature are augmented with MFCCs, and the number of IMFs from which the features are extracted are reduced by changing K values from 10 to 2, in steps of 2. It can be observed from the table that the best performance of the combination of the features are obtained for $K \leqslant 4$. The combination of different HS features and cG_K features are compared to the standalone MFCCs, an absolute improvement in EER of around 0.5-1.0 % are obtained at $K = 4$. The additional 51 dimensions of the Ext. MFCCs also significantly improves the EER as compared to the standard MFCCs. Similar trends are noticed for mDCF values in each case. Figure 6.10 shows the DET curves for different feature combinations, considering K=4. It can be observed from the figure that the Ext. MFCCs outperforms the MFCCs and MFCCs+cG_K combination [193]. The MFCCs+HS combinations outperform the standard MFCCs for most of the cases. However, they are not always competitive with the 51 dimensional Ext. MFCCs. The results obtained using the speech specific knowledge based robust EPD method are compared with the results obtained using energy

Table 6.1: The performance of the TDSV system evaluated using Robust EPD and Energy Based EPD results are shown in brackets, using 39-dimensional MFCCs, 51-dimensional Ext. MFCCs, cG_K feature and the proposed seven HS features derived from MEMD, using all the IMFs ($K = 10$), on the RSR2015 database. The numbers shown in brackets are obtained by using energy-based VAD and without brackets are obtained by using speech specific knowledge based robust EPD method.

Technique		Mel filterbank		MEMD filterbank	MEMD-Hilbert Spectrum						
Feature		MFCCs 39 dim	Ext. MFCCs 51 dim	cG_{10}	F_{10}	E_{10}	σF_9	σE_9	ΔF_{10}	ΔE_{10}	Υ_{10}
Sentence	Metric										
TD-1	EER (%)	6.37 (7.59)	4.93 (6.22)	13.22 (27.71)	6.83 (15.48)	11.76 (12.57)	8.81 (11.65)	12.14 (13.14)	13.39 (22.61)	21.67 (19.33)	16.88 (16.22)
	mDCF	0.0375 (0.0388)	0.035 (0.0378)	0.0382 (0.0555)	0.0341 (0.0454)	0.0516 (0.0428)	0.0294 (0.0421)	0.0317 (0.0433)	0.0432 (0.0513)	0.0469 (0.0484)	0.0494 (0.046)
TD-2	EER (%)	5.9 (6.29)	4.9 (6.4)	11.23 (23.16)	6.42 (13.33)	10.81 (16.75)	8.27 (9.02)	8.64 (8.85)	8.57 (12.63)	12.87 (13.43)	7.71 (11.84)
	mDCF	0.0442 (0.0421)	0.0367 (0.0423)	0.0437 (0.0546)	0.0432 (0.0474)	0.0546 (0.0497)	0.0452 (0.044)	0.0453 (0.0439)	0.0359 (0.0469)	0.0457 (0.0472)	0.0419 (0.0463)
TD-3	EER (%)	5.87 (6.6)	4.94 (6.25)	11.89 (17.5)	6.67 (11.4)	6.45 (8.18)	7.33 (15.17)	8.56 (15.93)	7.03 (7.83)	8.39 (9.9)	7.26 (10.19)
	mDCF	0.039 (0.0422)	0.0382 (0.0418)	0.0414 (0.058)	0.039 (0.0492)	0.0336 (0.0444)	0.0411 (0.0544)	0.0393 (0.0555)	0.0301 (0.0441)	0.0367 (0.0469)	0.0378 (0.0475)

based VAD method, shown in brackets. It is observed that the results are improved after using the robust end points detection method. Figure 6.11 presents the performance metrics of different feature combinations for the sentences TD-2 and TD-3, respectively. In the case of TD-2 and TD-3 sentences, the Ext. MFCCs marginally improves the system performance after applying the robust EPD and also very less improvement is observed in HS feature combinations as compared to our previous work [184]. However, the MFCCs+cG_K and MFCCs+HS feature, again, obtain best performances at K=4. This study clearly suggests that HS features and cG_K can naturally complement the MFCCs performances, under clean conditions. The additional 51 dimensional Ext. MFCCs may not always improve the system performance. Moreover, the use of robust EPD method further improves the performance of the system. It is clearly observed that only the first few IMFs are useful in characterizing the speakers which is similar to the observations reported in [93, 184].

6.6.1.2 Performances on the RSR2015 database under degraded conditions

Having demonstrated the performances of the features and their combinations under clean conditions, we can now evaluate how robust the features are to external interference. For this reason, the testing speech utterances are corrupted with Babble noise [110] prior to extracting feature sets

Table 6.2: The performance of the TDSV system using robust EPD followed by MEMD feature extraction. The results are shown for 39-dimensional MFCCs, 51-dimensional Ext. MFCCs, and the combinations of the 39-dimensional MFCCs with the cG_K feature and each of the seven HS experimental feature obtained from MEMD. The performance metrics are evaluated for the TD-1 speech utterance of the RSR2015 database. The dimensions of the MEMD features are varied by changing K from 10 to 2, in steps of 2. The numbers shown in brackets are obtained by using energy-based VAD and without brackets are obtained by using speech specific knowledge based robust EPD method.

Technique		Mel filterbank		MEMD filterbank	MEMD-Hilbert Spectrum						
Feature		MFCCs	Ext. MFCCs	MFCCs +	MFCCs +	MFCCs +	MFCCs +	MFCCs +	MFCCs +	MFCCs +	MFCCs +
K	Metric	39 dim	51 dim	cG_K	F_K	E_K	σF_{K-1}	σE_{K-1}	ΔF_K	ΔE_K	Υ_K
10	EER (%)	6.37 (7.59)	**4.93** (6.22)	8.72 (8.65)	6.9 (6.48)	6.82 (8.17)	10.46 (9.36)	7.33 (8.82)	8.08 (8.41)	7.88 (8.4)	7.72 (7.11)
	mDCF	0.0375 (0.0388)	**0.035** (0.0378)	0.0413 (0.0398)	0.037 (0.038)	0.0375 (0.0392)	0.042 (0.0402)	0.0385 (0.0397)	0.0383 (0.0396)	0.037 (0.0394)	0.0372 (0.0385)
8	EER (%)	-	-	7.28 (7.34)	6.99 (6.54)	6.86 (7.07)	8.85 (9.7)	7.61 (8.54)	6.69 (7.58)	7.98 (7.39)	7.9 (7.41)
	mDCF	-	-	0.0346 (0.0387)	0.0353 (0.038)	0.0353 (0.0383)	0.0391 (0.0404)	0.039 (0.0395)	0.0357 (0.0389)	0.0378 (0.0385)	0.0367 (0.0388)
6	EER (%)	-	-	5.96 (6.96)	6.44 (6.54)	6.68 (7.56)	7.96 (8.27)	6.69 (7.02)	6.77 (6.95)	7.76 (7.04)	6.7 (7.21)
	mDCF	-	-	0.0295 (0.0384)	0.0344 (0.038)	0.0359 (0.0387)	0.0383 (0.0393)	0.0367 (0.0382)	0.0359 (0.0384)	0.039 (0.0382)	0.0336 (0.0386)
4	EER (%)	-	-	**5.15** (**5.03**)	**6.18** (6.57)	**6.23** (6.62)	6.67 (6.91)	**6.28** (6.39)	6.88 (7.28)	6.45 (6.71)	**6.27** (7.2)
	mDCF	-	-	**0.0249** (0.0368)	**0.0351** (0.0381)	**0.0361** (0.0379)	**0.0338** (0.0381)	**0.0353** (0.0377)	**0.0389** (0.0386)	**0.0349** (0.038)	**0.0349** (0.0386)
2	EER (%)	-	-	5.87 (6.66)	6.67 (6.96)	7.3 (6.93)	7.41 (6.85)	6.75 (7.7)	6.78 (6.75)	6.67 (7.42)	7.09 (7.27)
	mDCF	-	-	0.0269 (0.0381)	0.0348 (0.0384)	0.0358 (0.0382)	0.0357 (0.0381)	0.0367 (0.0388)	0.0367 (0.0382)	0.0353 (0.0386)	0.0346 (0.0386)

from the RSR2015 database. Table 6.3 presents the performances of the HS features and cG_K feature combination with MFCCs for the TD-1 sentence. As observed from the clean speech, the best performances are observed at K=4 and therefore this value is used for all the experiments conducted under degraded conditions. This can be noticed from the table that the performance of the TDSV system degrades significantly with the increase in SNR levels. The 51-dimensional Ext. MFCCs feature show slightly better performance as compared to the MFCCs. Most of the feature combinations (MFCCs + HS) shows improved performance compared to the MFCCs alone. Numbers within the brackets are the results obtained using energy based end points detection. Clearly, use of robust end points detection in the proposed method is giving improved results compared to the energy based end points detection, for different levels of degradation.

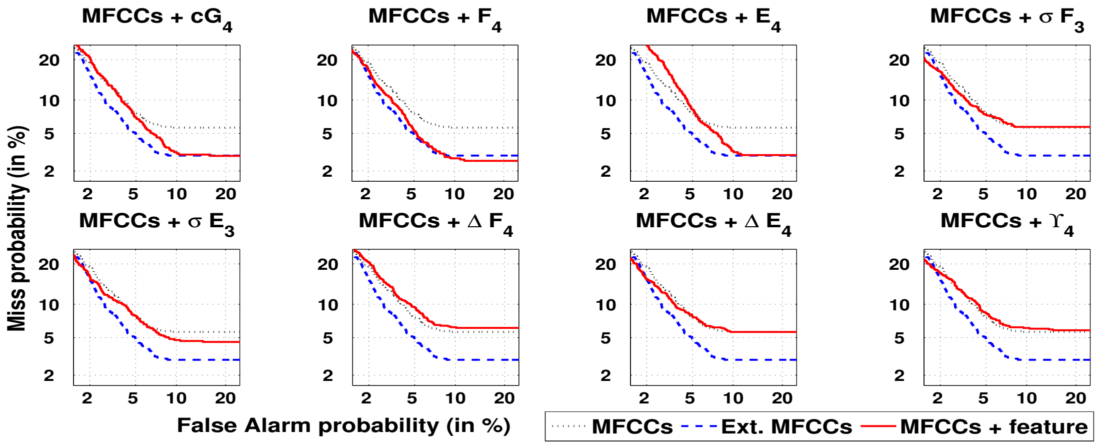


Figure 6.10: Presents the DET curves of the 39-dimensional MFCCs, 51-dimensional Ext. MFCCs, and the combinations of the MFCCs with the cG_K feature and each of the seven experimental features. The DET curves are plotted for the TD-1 sentence of the RSR2015 database. The dimensions of the cG_K and seven HS features are fixed at $K = 4$.

Figure 6.12 presents the performance metrics for the feature combinations for the TD-2 and TD-3 sentences. The performances under different SNR levels have been plotted for better representation of the improvement in the system performance. This can be noticed in the figure that similar to the observations made in Table 6.3, in most of the cases, the combined features are showing improved performance compared to the MFCCs alone.

6.6.1.3 Performances on the IITG database

So far we have evaluated the speech corrupted by artificially inserted Babble noise. Now in this section, we present the performance of the IITG system using speech utterances which are naturally affected by background speech, background noise, telephone channel, interference from other speakers and other environmental conditions. For this purpose, we considered the IITG database for evaluating robustness of the proposed system under practical conditions. This system was used for practical deployment at the institute level for marking attendance as an application.

Table 6.4 presents the performance metrics of the MFCCs+HS features and the MFCCs+cG_K feature combination for each of the three speech utterances. Taking characteristics of the speaker-specific cues from the examinations made for the RSR2015 database, the dimensions of the cG_K and different HS features are remain fixed by setting K=4. From the table, one can notice that the 51-dimensional Ext. MFCCs and MFCCs+cG_K feature combination provide the improved performance of the system with respect to the MFCCs for the TD-1 and TD-3 speech utterances, but MFCCs+cG_K feature combination for the TD-2 speech utterance case show slightly degraded performance as compared to

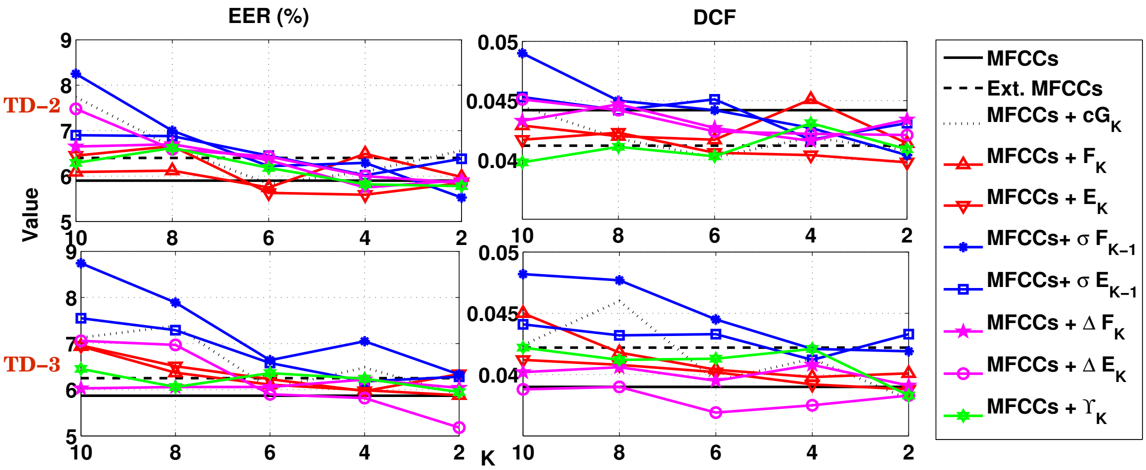


Figure 6.11: The performance of the TDSV system using robust EPD followed by MEMD feature extraction. The results are shown for the 39-dimensional MFCCs, 51-dimensional Ext. MFCCs, and the combinations of the 39-dimensional MFCCs with the cG_K feature and each of the seven experimental features. The performance metrics are evaluated for the TD-2 and TD-3 sentences of the RSR2015 database. The dimensions of the cG_K and seven HS features are varied by changing K from 10 to 2.

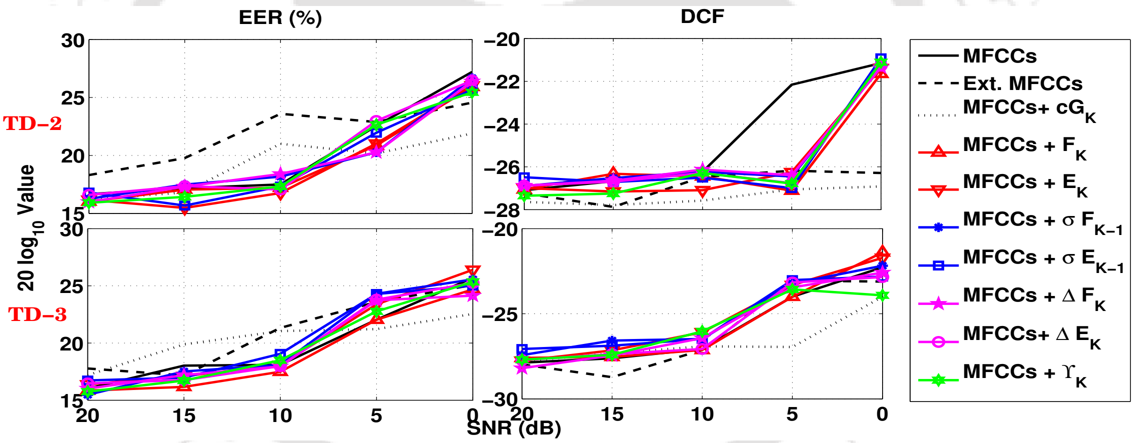


Figure 6.12: The performance of the TDSV system using robust EPD followed by MEMD feature extraction. The results are shown for the 39-dimensional MFCCs, 51-dimensional Ext. MFCCs, and the combinations of the 39-dimensional MFCCs with the cG_K feature and each of the seven experimental features. The performance metrics are evaluated for the TD-2 and TD-3 sentences of the RSR2015 database. The testing utterances are corrupted by Babble noise, with SNR varying from 20-0 dB. The dimensions of the cG_K seven features are kept constant, with $K = 4$.

the MFCCs and Ext. MFCCs. The MFCCs + HS feature combinations inconsistently provide the improved performance, for all the three speech utterances. The best performances are spread across different feature combinations and hence fusion of various low-dimensional features could be useful. It is also observed that the robust end points detection is helping in most of the cases by improving the EER compared to the energy based end points detection. In the table, results obtained by using energy based end points detection are shown inside brackets.

Table 6.3: The performance of the TDSV system using robust EPD followed by MEMD feature extraction. The results are shown for the 39 dimensional MFCCs, 51 dimensional Ext. MFCCs, and the combinations of the 39 dimensional MFCCs with the cG_K feature and each of the seven HS experimental feature obtained from MEMD. The performance metrics are evaluated for the TD-1 speech utterance of the RSR2015 database. The testing utterances are corrupted with Babble noise, with SNR varying from 20-0 dB, in steps of 5 dB. The dimensions of the MEMD features are kept constant, with $K = 4$. The numbers shown in brackets are obtained by using energy-based VAD and without brackets are obtained by using speech specific knowledge based robust EPD method.

Technique		Mel filterbank		MEMD filterbank	MEMD-Hilbert Spectrum						
Feature		MFCCs	Ext. MFCCs	MFCCs +	MFCCs +	MFCCs +	MFCCs +	MFCCs +	MFCCs +	MFCCs +	MFCCs +
SNR	Metric	39 dim	51 dim	cG_4	F_4	E_4	σF_3	σE_3	ΔF_4	ΔE_4	Υ_4
20 dB	EER (%)	6.81 (8.1)	6.55 (11.91)	6.46 (6.55)	6.33 (6.71)	6.44 (6.67)	6.5 (6.98)	6.92 (6.48)	6.3 (7.34)	6.76 (7.4)	**6.25** (7.84)
	mDCF	0.0442 (0.0394)	**0.0332** (0.0425)	0.0415 (0.038)	0.0444 (0.0382)	0.0447 (0.0379)	0.0452 (0.0382)	0.0474 (0.0378)	0.0453 (0.0387)	0.0453 (0.0385)	0.0429 (0.0391)
15 dB	EER (%)	7.23 (9.26)	6.9 (12.47)	6.93 (11.13)	7.12 (7.77)	**5.94** (6.95)	7.49 (7.64)	6.08 (7)	7.34 (7.37)	7.37 (7.56)	6.63 (8.06)
	mDCF	0.0464 (0.0404)	0.0453 (0.0429)	**0.0408** (0.0418)	0.0483 (0.039)	0.0439 (0.0382)	0.0472 (0.0387)	0.0462 (0.0382)	0.0465 (0.0387)	0.0463 (0.0387)	0.0434 (0.0393)
10 dB	EER (%)	7.5 (11.23)	7.38 (12.9)	11.23 (13.22)	7.26 (13.23)	**6.89** (10.31)	8.1 (10.27)	7.35 (10.29)	8.32 (9.89)	7.13 (10.34)	7.35 (10.48)
	mDCF	0.0488 (0.042)	0.046 (0.0433)	**0.0418** (0.0436)	0.0476 (0.0436)	0.0442 (0.0409)	0.0488 (0.0409)	0.0473 (0.0409)	0.0494 (0.0408)	0.0485 (0.041)	0.0485 (0.0413)
5 dB	EER (%)	13.42 (12.51)	10.73 (20.93)	**10.21** (14.18)	11.05 (14.35)	11.25 (12.6)	10.32 (12.37)	12.52 (12.59)	10.32 (13.31)	14.09 (13.51)	13.56 (13.14)
	mDCF	0.048 (0.0431)	0.0453 (0.0499)	0.0444 (0.0443)	**0.0441** (0.0445)	0.0487 (0.0428)	0.0476 (0.0427)	0.0447 (0.0428)	0.0479 (0.0436)	0.0459 (0.0436)	0.0458 (0.0435)
0 dB	EER (%)	22.93 (20.29)	**16.9** (23.29)	18.43 (21.04)	18.76 (21.6)	20.04 (20.42)	21.56 (20.76)	22.22 (20.1)	20.86 (22.17)	21.04 (20.68)	18.71 (22.03)
	mDCF	0.0476 (0.0493)	0.0485 (0.0522)	0.051 (0.05)	0.0527 (0.0505)	**0.0453** (0.0493)	0.0486 (0.0496)	0.0489 (0.049)	0.0556 (0.051)	0.0464 (0.0495)	0.0488 (0.0508)

6.6.2 Performances obtained by using Speech Enhancement followed by Robust EPD method

In this experiment, speech enhancement is performed before the end points detection and feature extraction stage. The performance on the RSR2015 database is presented not only under clean speech conditions but also under Babble noise of varying strength. Table 6.5 presents the performances of the RSR2015 database under clean and different noise levels. Numbers inside the bracket shows the best results from Table 6.3. These results were obtained without performing speech enhancement and are shown in Table 6.5 for comparison.

From the table, one can observe that introduction of the temporal and spectral enhancement on the speech utterances provides slightly better and comparable performances, under various conditions.

Table 6.4: The performance of the TDSV system using robust EPD followed by MEMD feature extraction. The results are shown for the 39-dimensional MFCCs, 51-dimensional Ext. MFCCs, and the combinations of the 39-dimensional MFCCs with the cG_K feature and each of the seven HS experimental feature obtained from MEMD. The performance metrics are evaluated for the TD-1, TD-2 and TD-3 speech utterances of the IITG database. The testing utterances are collected in the practical scenarios. The dimensions of the MEMD features are kept constant, with $K = 4$. The numbers shown in brackets are obtained by using energy-based VAD and the results shown without brackets are obtained by using speech specific knowledge based robust EPD method.

Technique		Mel filterbank		MEMD filterbank	MEMD-Hilbert Spectrum						
Feature		MFCCs 39 dim	Ext. MFCCs 51 dim	MFCCs + cG_4	MFCCs + F_4	MFCCs + E_4	MFCCs + σF_3	MFCCs + σE_3	MFCCs + ΔF_4	MFCCs + ΔE_4	MFCCs + Υ_4
Sentence	Metric										
TD-1	EER (%)	10.81 (13.19)	9.86 (12.84)	10.48 (10.91)	**9.42** (11.46)	10.96 (10.62)	11.77 (10.41)	12.76 (10.25)	11.23 (11.83)	12.68 (10.68)	11.3 (11.48)
	mDCF	0.0416 (0.0422)	0.0459 (0.0412)	0.0411 (0.035)	0.0423 (0.0368)	0.0423 (0.0339)	0.0429 (0.0332)	0.0443 (0.0327)	0.043 (0.0379)	**0.041** (0.0341)	0.0419 (0.0368)
TD-2	EER (%)	11.21 (11.37)	10.65 (11.65)	12.56 (11.56)	10.89 (11.3)	11.85 (10.33)	14.59 (10.71)	13.45 (9.86)	**9.1** (11.1)	10.86 (10.86)	10.38 (10.38)
	mDCF	0.0406 (0.0404)	0.0413 (0.0415)	0.0404 (0.0412)	0.0433 (0.0403)	0.0441 (0.0366)	0.0449 (0.038)	0.0435 (0.0349)	**0.04** (0.0395)	0.0417 (0.0385)	0.0415 (0.037)
TD-3	EER (%)	16.87 (18.42)	15.93 (17.93)	**14.69** (17.96)	15.88 (19.26)	16.37 (17.86)	16.82 (17.48)	17.45 (17.35)	15.7 (17.3)	17.25 (17.9)	15.92 (17.93)
	mDCF	0.0491 (0.0526)	0.0526 (0.0513)	0.0497 (0.0514)	**0.045** (0.0551)	0.0518 (0.0509)	0.0544 (0.0498)	0.0538 (0.0494)	0.0485 (0.0495)	0.0511 (0.051)	0.0512 (0.0513)

Table 6.5: The performance of the TDSV system using enhanced 39-dimensional MFCC features. The performance metrics are evaluated for the TD-1, TD-2 and TD-3 speech utterances of the RSR2015 database in terms of the EER and mDCF. The experimental analysis done on the clean and testing utterances are corrupted with varying strength from 0-20 dB, in steps of 5 dB.The numbers shown in brackets are obtained by MFCCs+HS best performing results and withut brackets are obtained by using speech enhancement followed by robust EPD method.

Sentence	Metric	Clean	**20 dB**	**15 dB**	**10 dB**	**5 dB**	**0 dB**
Dimension		39	39	39	39	39	39
TD-1	EER (%)	4.83 (6.18)	5.12 (6.25)	5.21 (5.94)	6.24 (6.89)	11.36 (10.32)	17.36 (18.71)
	mDCF	0.0386 (0.0351)	0.0378 (0.0429)	0.0402 (0.0439)	0.0421 (0.0442)	0.0653 (0.0479)	0.0756 (0.0488)
TD-2	EER (%)	4.77 (5.53)	5.16 (6.3)	6.11 (5.94)	7.73 (7.35)	13.15 (14.09)	19.05 (21.71)
	mDCF	0.0367 (0.0404)	0.042 (0.0453)	0.0432 (0.0439)	0.0498 (0.0473)	0.0674 (0.0684)	0.0772 (0.088)
TD-3	EER (%)	4.85 (5.18)	5.49 (5.94)	5.84 (6.43)	7.52 (7.49)	10.26 (12.65)	15.79 (18.31)
	mDCF	0.0384 (0.0383)	0.0396 (0.0425)	0.0429 (0.042)	0.0525 (0.044)	0.0634 (0.0631)	0.073 (0.0721)

Table 6.5 show the experimental results that consistently providing improved results as compared to the MFCCs augmented with HS features and MFCC standalone features. Table 6.6 shows the same experiments for the IITG database. The IITG database is already affected by environmental noise, telephone channel and interference from other speakers, and hence noise is not added artificially.

Table 6.6: The performance metrics are evaluated for the TD-1, TD-2 and TD-3 speech utterances of the IITG database. The testing utterances are collected in the practical scenarios. The results are evaluated in terms of EER and mDCF. The numbers shown in brackets are obtained by MFCCs+HS best performing results and without brackets are obtained by using speech enhancement followed by robust EPD method.

Sentence	TD-1	TD-2	TD-3
Metric	39 dim	39 dim	39 dim
EER (%)	6.82 (9.42)	7.38 (9.1)	8.54 (14.69)
mDCF	0.0331 (0.0423)	0.0343 (0.04)	0.0398 (0.0497)

Similar trend in performance is observed for the IITG database as well. Speech enhancement improves the performance of the TDSV system significantly.

6.6.3 Performances obtained by using speech enhancement followed by robust EPD and extraction of MEMD features

The final combined system can be obtained by using all three robust methods, where robust EPD is performed on the enhanced speech and the robust MEMD features are extracted from the detected speech regions. Performances are evaluated on the RSR2015 and the IITG databases.

Table 6.7: The performance of the TDSV system using speech enhancement followed by robust EPD and extraction of MEMD features. Results are shown for 39-dimensional MFCCs, 51-dimensional Ext. MFCCs, and the combinations of the 39-dimensional MFCCs with the cG_K feature and each of the seven HS experimental feature obtained from MEMD. The performance metrics are evaluated for the TD-1 speech utterance of the RSR2015 database. The testing utterances are corrupted with Babble noise, with SNR varying from 20-0 dB, in steps of 5 dB. The dimensions of the MEMD features are kept constant, with $K = 4$.

Technique		Mel filterbank		MEMD filterbank	MEMD-Hilbert Spectrum						
Feature		MFCCs	Ext. MFCCs	MFCCs +	MFCCs +	MFCCs +	MFCCs +	MFCCs +	MFCCs +	MFCCs +	MFCCs +
SNR	Metric	39 dim	51 dim	cG_4	F_4	E_4	σF_3	σE_3	ΔF_4	ΔE_4	Υ_4
20 dB	EER (%)	7.3	5.37	**5.08**	**5.08**	5.12	5.37	**5.08**	**5.08**	5.51	5.83
	mDCF	0.0353	0.0341	0.0341	**0.033**	0.0338	0.0341	**0.033**	0.0341	0.038	0.0365
15 dB	EER (%)	10.37	7.59	11.55	11.36	10.75	10.85	8.84	11.55	8.45	**7.01**
	mDCF	0.0571	0.0337	0.0428	0.0428	0.066	0.0428	**0.0381**	0.0428	0.044	0.0441
10 dB	EER (%)	**10.55**	12.6	13.36	11.55	12.43	11.55	11.48	13.36	10.75	11.22
	mDCF	0.0704	0.0648	0.054	0.0535	0.0522	0.0649	0.0635	0.054	0.066	**0.0482**
5 dB	EER (%)	15.68	17.68	17.68	16.57	16.18	16.18	17.89	17.68	16.53	**15.49**
	mDCF	0.0725	0.0669	0.0648	0.0647	0.0707	0.0707	0.0725	0.0648	0.0617	**0.0602**
0 dB	EER (%)	17.24	17.73	17.73	**17.17**	21.64	21.64	17.94	17.73	17.68	20.14
	mDCF	0.0864	0.0748	**0.0669**	0.0758	0.0781	0.0781	0.084	**0.0669**	0.082	0.0777

Table 6.7 shows the results of the final combined system obtained by using three robust methods applied on the speech utterances in a sequential manner. The obtained results shows improvement in

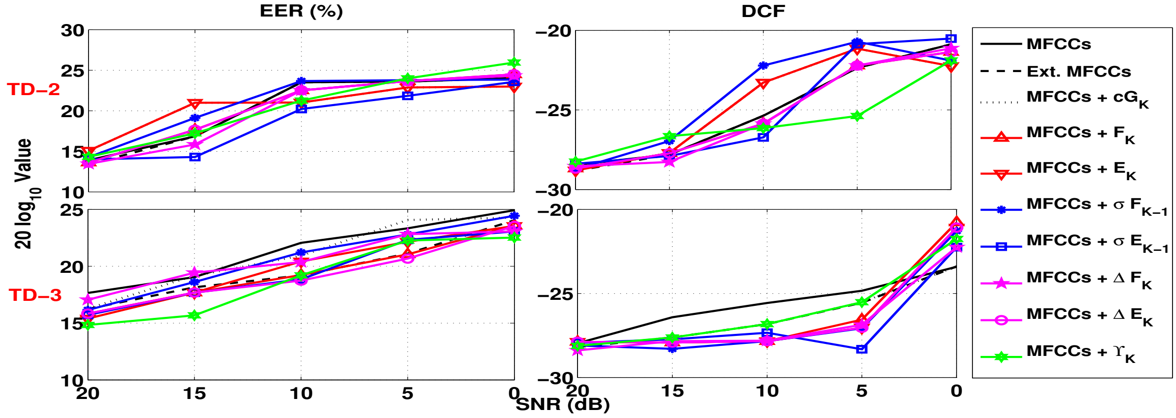


Figure 6.13: The performance of the TDSV system using speech enhancement followed by robust EPD and extraction of MEMD features. Results are shown for 39-dimensional MFCCs with the cG_K feature and each of the seven experimental features. The performance metrics are evaluated for the TD-2 and TD-3 sentences of the RSR2015 database. The testing utterances are corrupted by Babble noise, with SNR varying from 20-0 dB. The dimensions of the cG_K seven features are kept constant, with $K = 4$.

performances compared to the performances with the signal and feature level compensation techniques shown in Table 6.3. The improvements are consistently observed for the low SNR cases namely, 0 dB and 5 dB SNR. For 10-20 dB SNR, improvements are observed in some of the cases. Similar trend in performances are observed for TD-2 and TD-3 speech sentences as shown in Figure 6.13.

Table 6.8: The performance of the TDSV system using speech enhancement followed by robust EPD and extraction of MEMD features. Results are shown for 39-dimensional MFCCs, 51-dimensional Ext. MFCCs, and the combinations of the 39-dimensional MFCCs with the cG_K feature and each of the seven HS experimental feature obtained from MEMD. The performance metrics are evaluated for the TD-1, TD-2 and TD-3 speech utterances of the IITG database. The testing utterances are collected in the practical scenarios. The dimensions of the MEMD features are kept constant, with $K = 4$.

Technique		Mel filterbank		MEMD filterbank	MEMD-Hilbert Spectrum						
Feature		MFCCs	Ext. MFCCs	MFCCs +	MFCCs +	MFCCs +	MFCCs +	MFCCs +	MFCCs +	MFCCs +	MFCCs +
Sentence	Metric	39 dim	51 dim	cG_4	F_4	E_4	σF_3	σE_3	ΔF_4	ΔE_4	Υ_4
TD-1	EER (%)	9.5	8.69	6.82	7.07	6.73	6.28	6.46	**6.12**	6.18	7.07
	mDCF	0.0367	0.0365	0.0341	0.0341	0.0339	0.0339	0.0333	**0.0309**	0.0331	0.0341
TD-2	EER (%)	6.45	6.69	6.26	6.43	6.23	6.2	**6.01**	6.91	6.26	6.43
	mDCF	0.0343	0.0338	0.0336	0.0341	0.0338	0.0331	0.0334	0.0354	0.0346	**0.0328**
TD-3	EER (%)	9.84	9.74	8.35	9.69	9.38	8.17	**7.63**	8.2	8.35	9.67
	mDCF	0.0395	0.0405	**0.0379**	0.0401	0.0386	0.038	0.0387	0.0399	**0.0379**	0.0397

Table 6.8 shows the same set of features and same experiments for the IITG database. Similar trend in the system performance is observed for the IITG database as well, which can be observed by comparing the results in Table 6.8 with the corresponding results shown in Table 6.4. The combination of three robust methods improves the performance of the TDSV system for most of the cases.

6.7 Summary and Conclusion

This work focused on using different techniques for TDSV system under degraded speech and challenging test conditions. The process of TDSV system includes several stages such as, pre-processing, feature extraction, modeling and decision. In this work, several robust methods are explored in different stages of the speaker verification system. In the pre-processing stage, a robust end points detection using speech specific knowledge is used instead of energy based VAD. Similarly, in the feature extraction stage, robust features extracted from HS of IMFs obtained from MEMD are used in addition to the conventional MFCC features. We observed that when we perform the robust end points detection and use the robust HS features as additional features in the TDSV system, there is improvement in performance as compared to the standalone MFCCs extracted from the speech regions detected by energy based VAD. The seven features in combination with the MFCCs i.e., MFCCs+HS features show improved performance compared to MFCCs alone in different levels of degradation. This improvement is observed for HS features obtained from low dimensions of IMFs as expected and the best performances are spread across different feature combinations with MFCCs.

Moreover, we also explored a combined temporal and spectral technique for speech enhancement. The enhanced speech utterances are passed through robust EPD to obtain the refined begin and end points. The results obtained after performing speech enhancement show slightly better performance compared to the best performance obtained by using MFCCs augmented with HS features and standalone MFCCs, without performing speech enhancement. Finally, we used all three methods in a sequential manner, where robust EPD is performed on the enhanced speech and then MEMD features are extracted from the regions between the detected end points. The combined method significantly improves the system performance for the test utterances with 0 dB, 5 dB and 10 dB SNRs. On the RSR2015 database, the proposed method provides a relative improvement in EER by 25%, 23% and 6%, for 0 dB, 5 dB and 10 dB, SNRs, respectively. On the IITG database, it provides the relative improvement in EER from 30% to 45%. The classification performance is improved in terms of mDCF as well. On the RSR2015 database, the combined method shows relative improvement in mDCF by 8%, 23% and 14%, for 0 dB, 5 dB and 10 dB, SNRs, respectively, whereas, the mDCF is relatively improved by 24% on the IITG database.

7

Summary and Conclusions

Contents

Objective

This chapter, the final chapter of this thesis towards development of a robust text-dependent speaker verification (TDSV) system under clean, degraded speech and challenging test conditions using noise robust techniques. This chapter summarizes the experiments done as a part of this thesis and highlights the improvements obtained. The conclusions are presented after evaluation. Towards the end, possible future research directions also mentioned for researchers.

7.1 Summary

The objective of this thesis work is to develop a robust TDSV system by using noise robust techniques for achieving better system performance under clean and degraded speech conditions. To achieve this, three different directions are explored for a TDSV task. First, a TDSV system is developed by using combined temporal and spectral speech enhancement for enhancing speech regions embedded in background noise followed by EPD. Another approach is developed based on robust features extracted from HS of IMFs obtained from MEMD of the speech signal. The third one corresponds to use of speech-specific knowledge based robust EPD. In the final working chapter of the thesis, a robust TDSV system is developed by exploiting the evidences from these three different proposed approaches. The summary and conclusions of the work in each chapter of the thesis is presented as follows:

The first chapter of the thesis - Introduction - is dedicated to the speaker knowledge which is utilized for authenticating a person from the speech utterance. The system implemented over an online telephone network which is handled by an IVR system. This is developed through CTI card for marking student attendance. The first chapter also discusses the modular representation of the work, motivation behind the implementation and some issues present in the practical deployment of the TDSV systems are explored. Finally, the chapter concludes with a brief description of the organization of framework of this thesis.

The first chapter of this thesis also presents some issues related to implementation of a robust TDSV system. Based on these issues, we will present some reviews in the **the second chapter of the thesis - End point detection, speech enhancement and feature extraction - A Review** - presents possible directions for the development of person authentication system from the perspective of field deployable systems. Based on the main issues addressed in the **first chapter** of this thesis, the reviews are explored in three different directions based on the characteristics of the speech signal. The first issue is related to end point detection under degraded speech conditions. Second issue exploited for the removal of degradation effect by using speech enhancement under degraded speech. The third direction is review of alternative/ complementary features for enhancing the system performance further. The chapter concluded by outlining the framework of this thesis.

The third chapter of the thesis - Processing Degraded Speech for Text-Dependent Speaker Verification - attempted to achieve the first objective of the thesis. The objective is to *"use speech enhancement for enhancing degraded speech followed by EPD for improved TDSV system performance"*. Different combination of experimental studies are conducted, out of which one is able to successfully improved system performance when compared to other combinations. Among all the experiments conducted, the best performance is obtained by using the combination of temporal and spectral enhancement followed by EPD which helps in noise suppression throughout the utterance. The obtained speech utterances attributes the better end points detected on the enhanced speech.

The fourth chapter of the thesis - Analysis of the Hilbert Spectrum for Speaker Verification -attempted to fulfill the second objective of the thesis. The objective is to *"investigate the utility of the HS of the speech signal, obtained using MEMD in combination with MFCCs for the purpose of augmenting the performance of a TDSV system under practical conditions"*. More specifically, features extracted from the instantaneous frequencies and energies of the IMFs, which constitute its HS, obtained from MEMD. The subset of IMFs are useful for extracting the speaker specific features. This allows us to limit the number of IMFs extracted and find the seven different useful feature sets for augmenting the MFCCs in extracting speaker-specific information. These extracted features were then tested individually and in conjunction with the MFCCs, on two different databases for the TDSV task, using DTW technique. The features were tested on RSR2015 database not only under clean conditions, but also under Babble noise condition of varying strength. The testing conditions for the IITG database also provided high degree of mismatch from the training conditions.

The fifth chapter of the thesis - Significance of Speech Specific Knowledge for End Point Detection -attempted to establish the third objective of the thesis. The objective is to *"show that the proposed begin and end point detection algorithm using some speech specific information aims to remove non-overlapping speech as well as non-speech background degradations"*. The VLR detection method uses vocal tract and source related information to remove background noise. DRF information is used to remove non-speech background and FSS is performed to remove the background speech. To detect the consonants at the beginning and at the end of the VLRs, glottal activity and dominant aperiodic regions detection are performed. Further, SDK is used to remove some spurious VLRs. The proposed method performs better than some of the existing techniques such as energy-, FSS, GAD and VLR based EPDs. The system performance is evaluated not just for clean speech signals, but also conducted on the speech signals affected by noise with varying strength and telephone channel conditions and different environmental conditions.

The sixth chapter of the thesis - Robust Text-Dependent Speaker Verification - attempted to address the fourth objective of the thesis. The objective is to *"use of different approaches explored earlier in different combinations to improve the performance of the TDSV system"*. In this chapter, several noise robust methods are explored in different stages of TDSV system. In the pre-processing stage, a robust EPD using speech specific knowledge is used instead of energy based EPD. Similarly, in the feature extraction stage, robust features are extracted from HS of IMFs obtained from MEMD are used in combination with MFCC features. MFCCs+HS features show robust and better improved performance in different levels of degradation, on two different databases. This improvement is observed for HS features obtained from low dimensions of IMFs. The best system performances are spread across different feature combinations with MFCCs. Finally, the system is developed for the practical deployment of the biometric based attendance system at the institute level for marking their attendance.

7.2 Conclusions

- The energy based end point detection will perform better on the enhanced speech compared to directly using on the degraded speech utterances.
- The concise but adaptive filterbank nature of MEMD features captures different speaker specific information in a different manner than the fixed Mel filterbank features which may be useful in enhancing the performance of the TDSV task.
- The use of different speech specific knowledge provides a robust end point detection method.
- The overall improvement can be achieved by combining different approaches tried like speech enhancement, new HS features and end points detection using speech-specific knowledge.

7.3 Directions for Future Work

- For the detection of begin and end points, we have used VLROPs, VLREPs and some speech specific knowledge. The detection accuracy of the begin and end points of the speech utterance may be improved by using statistical techniques with the proposed signal processing techniques.
- In the present thesis work, we explored the combined temporal and spectral speech enhancement techniques which are used sequentially to obtain better suppression of noise where the performance of the TDSV system shows improvement. More work needs to be done in future to develop joint temporal and spectral speech enhancement where the degraded speech is processed in a parallel or simultaneous manner.
- The EPD using speech specific knowledge was tested for limited practical conditions. For instance,it is assumed that the speech regions are non-overlapping with non-speech and background speech regions. As such, the utility of new algorithms are required where non-speech and background speech regions are overlapped with speech regions under practical conditions may be an interesting investigation.
- The usefulness of excitation source features like MPDSS, RMFCC and DCTILPR found to be effective in text independent speaker verification can also be explored in the TDSV task.
- Recently New techniques are attempted to mimic the MEMD, *variational Mode Decomposition* (VMD), Empirical Wavelet algorithms, with a better mathematical framework and with reduced mode-mixing. This is an open area of research for speech community. More work can be done in this regard and would bring more credibility and acceptability to the work.

Bibliography

[1] R. K. Bhukya, B. D. Sarma, and S. R. M. Prasanna, "End point detection using speech-specific knowledge for text-dependent speaker verification," *Circuits, Systems, and Signal Processing*, pp. 1–33, 2018.

[2] L. R. Rabiner and R. W. Schafer, *Digital processing of speech signals.* Prentice Hall, 1978.

[3] L. R. Rabiner, R. W. Schafer *et al.*, "Introduction to digital speech processing," *Foundations and Trends® in Signal Processing*, vol. 1, no. 1–2, pp. 1–194, 2007.

[4] J. Benesty, M. M. Sondhi, and Y. Huang, *Springer handbook of speech processing.* Springer Science & Business Media, 2008.

[5] F. Bimbot, J.-F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-García, D. Petrovska-Delacrétaz, and D. A. Reynolds, "A tutorial on text-independent speaker verification," *EURASIP Journal on Advances in Signal Processing*, vol. 2004, no. 4, p. 101962, 2004.

[6] S. Marinov, "Text dependent and text independent speaker verification systems. technology and applications," *overview article*, 2003.

[7] B. Gold, N. Morgan, and D. Ellis, *Speech and audio signal processing: processing and perception of speech and music.* John Wiley & Sons, 2011.

[8] L. Boves and E. Den Os, "Speaker recognition in telecom applications," in *Interactive Voice Technology for Telecommunications Applications. IVTTA'98. Proceedings.* IEEE, 1998, pp. 203–208.

[9] S. Furui, "Recent advances in speaker recognition," *Pattern Recognition Letters*, vol. 18, no. 9, pp. 859–872, 1997.

[10] G. R. Doddington, M. A. Przybocki, A. F. Martin, and D. A. Reynolds, "The nist speaker recognition evaluation–overview, methodology, systems, results, perspective," *Speech Communication*, vol. 31, no. 2, pp. 225–254, 2000.

[11] H. Gish and M. Schmidt, "Text-independent speaker identification," *IEEE signal processing magazine*, vol. 11, no. 4, pp. 18–32, 1994.

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

[13] 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, pp. 3317–3318.

[14] 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.

[15] J. P. Campbell, "Speaker recognition: A tutorial," *Proceedings of the IEEE*, vol. 85, no. 9, pp. 1437–1462, 1997.

[16] M. Hébert, "Text-dependent speaker recognition," in *Springer handbook of speech processing.* Springer, 2008, pp. 743–762.

[17] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech communication*, vol. 52, no. 1, pp. 12–40, 2010.

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

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

[20] S. E. Bou-Ghazale and K. Assaleh, "A robust endpoint detection of speech for noisy environments with application to automatic speech recognition," in *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE Int. Conf. on*, vol. 4. IEEE, 2002, pp. IV–3808.

[21] V. Hautamäki, M. Tuononen, T. Niemi-Laitinen, and P. Fränti, "Improving speaker verification by periodicity based voice activity detection," in *Proc. 12th Int. Conf. Speech and Computer (SPECOM'2007)*, vol. 2, 2007, pp. 645–650.

[22] 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, 1999.

[23] T. Kinnunen and P. Rajan, "A practical, self-adaptive voice activity detector for speaker verification with noisy telephone and microphone data." in *ICASSP*. Citeseer, 2013, pp. 7229–7233.

[24] S. R. M. Prasanna and G. Pradhan, "Significance of vowel-like regions for speaker verification under degraded conditions," *IEEE Trans. on audio, speech, and language processing*, vol. 19, no. 8, pp. 2552–2565, 2011.

[25] J. Ortega-García and J. González-Rodríguez, "Overview of speech enhancement techniques for automatic speaker recognition," in *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth Int. Conf. on*, vol. 2. IEEE, 1996, pp. 929–932.

[26] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. on acoustics, speech, and signal processing*, vol. 27, no. 2, pp. 113–120, 1979.

[27] B. Yegnanarayana, C. Avendano, H. Hermansky, and P. S. Murthy, "Speech enhancement using linear prediction residual," *Speech communication*, vol. 28, no. 1, pp. 25–42, 1999.

[28] P. Krishnamoorthy and S. M. Prasanna, "Enhancement of noisy speech by temporal and spectral processing," *Speech Communication*, vol. 53, no. 2, pp. 154–174, 2011.

[29] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," in *Readings in speech recognition*. Elsevier, 1990, pp. 65–74.

[30] X. Huang, A. Acero, H.-W. Hon, and R. Reddy, *Spoken language processing: A guide to theory, algorithm, and system development*. Prentice hall PTR Upper Saddle River, 2001, vol. 1.

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

[32] M. D. Plumpe, T. F. Quatieri, and D. A. Reynolds, "Modeling of the glottal flow derivative waveform with application to speaker identification," *IEEE Trans. on Speech and Audio Processing*, vol. 7, no. 5, pp. 569–586, 1999.

[33] S. R. M. Prasanna, C. S. Gupta, and B. Yegnanarayana, "Extraction of speaker-specific excitation information from linear prediction residual of speech," *Speech Communication*, vol. 48, no. 10, pp. 1243–1261, 2006.

[34] H. Hermansky, "Should recognizers have ears?" *Speech communication*, vol. 25, no. 1-3, pp. 3–27, 1998.

[35] T. Kinnunen, "Joint acoustic-modulation frequency for speaker recognition," in *Acoustics, Speech and Signal Processing, 2006.*, vol. 1. IEEE, 2006, pp. I–I.

[36] 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.

[37] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. on acoustics, speech, and signal processing*, vol. 28, no. 4, pp. 357–366, 1980.

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

[39] 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.

[40] L. Rabiner, A. Rosenberg, and S. Levinson, "Considerations in dynamic time warping algorithms for discrete word recognition," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 26, no. 6, pp. 575–582, 1978.

[41] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Trans. on acoustics, speech, and signal processing*, vol. 26, no. 1, pp. 43–49, 1978.

[42] R. Sharma, S. R. M. Prasanna, R. K. Bhukya, and R. K. Das, "Analysis of the intrinsic mode functions for speaker information," *Speech Communication*, vol. 91, pp. 1–16, 2017.

[43] S. K. Sahoo, T. Choubisa, and S. M. Prasanna, "Multimodal biometric person authentication: A review," *IETE Technical Review*, vol. 29, no. 1, pp. 54–75, 2012.

[44] 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 *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[45] R. Ramos-Lara, M. López-García, E. Cantó-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.

[46] R. K. Das, S. Jelil, and S. R. M. Prasanna, "Development of multi-level speech based person authentication system," *Journal of Signal Processing Systems*, vol. 88, no. 3, pp. 259–271, 2017.

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

[48] R. K. Das and S. M. Prasanna, "Speaker verification from short utterance perspective: a review," *IETE Technical Review*, pp. 1–19, 2017.

[49] S. R. M. Prasanna, J. M. Zachariah, and B. Yegnanarayana, "Begin-end detection using vowel onset points," in *Workshop on Spoken Language Processing*, 2003.

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

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

[52] D. J. Hermes, "Vowel-onset detection," *The Journal of the Acoustical Society of America*, vol. 87, no. 2, pp. 866–873, 1990.

[53] J.-F. Wang, C.-H. Wu, S.-H. Chang, and J.-Y. L. Lee, "A hierarchical neural network model based on a c/v segmentation algorithm for isolated mandarin speech recognition," *IEEE Trans. on Signal Processing*, vol. 39, no. 9, pp. 2141–2146, 1991.

[54] J. F. Wang and S.-H. Chen, "Ac/v segmentation algorithm for mandarin speech signal based on wavelet transforms," in *Acoustics, Speech, and Signal Processing, 1999. Proceedings., IEEE Int. Conf. on*, vol. 1. IEEE, 1999, pp. 417–420.

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

[56] S. R. M. Prasanna and B. Yegnanarayana, "Detection of vowel onset point events using excitation information," in *Ninth European Conference on Speech Communication and Technology*, 2005.

[57] A. K. Vuppala, J. Yadav, S. Chakrabarti, and K. S. Rao, "Vowel onset point detection for low bit rate coded speech," *IEEE Trans. on Audio,Speech and lang Processing*, vol. 20, no. 6, pp. 1894–1903, 2012.

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

[59] 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.

[60] K. Ramesh, S. R. M. Prasanna, and R. K. Das, "Significance of glottal activity detection and glottal signature for text dependent speaker verification," in *Signal Processing and Communications (SPCOM), 2014 Int. Conf. on.* IEEE, 2014, pp. 1–5.

[61] C. Prakash, D. N. Gowda, and S. V. Gangashetty, "Analysis of acoustic events in speech signals using bessel series expansion," *Circuits, Systems, and Signal Processing*, vol. 32, no. 6, pp. 2915–2938, 2013.

[62] S. V. Vaseghi, *Advanced digital signal processing and noise reduction.* John Wiley & Sons, 2008.

[63] O. Cappé, "Elimination of the musical noise phenomenon with the ephraim and malah noise suppressor," *IEEE Trans. on Speech and Audio Processing*, vol. 2, no. 2, pp. 345–349, 1994.

[64] S. Kamath and P. Loizou, "A multi-band spectral subtraction method for enhancing speech corrupted by colored noise." in *ICASSP*, vol. 4. Citeseer, 2002, pp. 44 164–44 164.

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

[66] R. McAulay and T. F. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *Acoustics, Speech and Signal Processing, IEEE Trans. on*, vol. 34, no. 4, pp. 744–754, 1986.

[67] E. J. Balster, Y. F. Zheng, and R. L. Ewing, "Combined spatial and temporal domain wavelet shrinkage algorithm for video denoising," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 16, no. 2, pp. 220–230, 2006.

[68] H. Taşmaz and E. Erçelebi, "Speech enhancement based on un-decimated wavelet packet-perceptual filterbanks and MMSE–STSA estimation in various noise environments," *Digital Signal Processing*, vol. 18, no. 5, pp. 797–812, 2008.

[69] M. K. Hasan, S. Salahuddin, and M. R. Khan, "Reducing signal-bias from mad estimated noise level for DCT speech enhancement," *Signal Processing*, vol. 84, no. 1, pp. 151–162, 2004.

[70] J.-H. Chang, S. Gazor, N. S. Kim, and S. K. Mitra, "Multiple statistical models for soft decision in noisy speech enhancement," *Pattern Recognition*, vol. 40, no. 3, pp. 1123–1134, 2007.

[71] A. Gersho, "Advances in speech and audio compression," *Proceedings of the IEEE*, vol. 82, no. 6, pp. 900–918, 1994.

[72] M. Dendrinos, S. Bakamidis, and G. Carayannis, "Speech enhancement from noise: A regenerative approach," *Speech Communication*, vol. 10, no. 1, pp. 45–57, 1991.

[73] Y. Ephraim and H. L. Van Trees, "A signal subspace approach for speech enhancement," *IEEE Trans. on speech and audio processing*, vol. 3, no. 4, pp. 251–266, 1995.

[74] B. Yegnanarayana, S. M. Prasanna, and K. S. Rao, "Speech enhancement using excitation source information," in *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 2002, pp. I–541.

[75] S. R. M. Prasanna, "Event based analysis of speech," *Dept. of Computer Science and Engineering, Ph. D. Thesis: Indian Institute of Technology Madras, India*, 2004.

[76] U. Mittal and N. Phamdo, "Signal/noise KLT based approach for enhancing speech degraded by colored noise," *IEEE Trans. on Speech and Audio Processing*, vol. 8, no. 2, pp. 159–167, 2000.

[77] A. Rezayee and S. Gazor, "An adaptive klt approach for speech enhancement," *IEEE Trans. on Speech and Audio Processing*, vol. 9, no. 2, pp. 87–95, 2001.

[78] S. H. Jensen, P. C. Hansen, S. D. Hansen, and J. A. Sorensen, "Reduction of broad-band noise in speech by truncated QSVD," *IEEE Trans. on Speech and Audio Processing*, vol. 3, no. 6, pp. 439–448, 1995.

[79] F. Jabloun and B. Champagne, "Incorporating the human hearing properties in the signal subspace approach for speech enhancement," *IEEE Trans. on Speech and Audio Processing*, vol. 11, no. 6, pp. 700–708, 2003.

[80] 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, p. 258184, 2007.

[81] M. Li, J. Kim, P. K. Ghosh, V. Ramanarayanan, and S. Narayanan, "Speaker verification based on fusion of acoustic and articulatory information." in *INTERSPEECH*, 2013, pp. 1614–1618.

[82] T. Kinnunen and P. Alku, "On separating glottal source and vocal tract information in telephony speaker verification," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. Int. Conf. on.* IEEE, 2009, pp. 4545–4548.

[83] K. Murthy and B. Yegananarayana, "Combining evidence from residual phase and MFCC features for speaker recognition," *IEEE Signal Processing Letters*, vol. 13(1), pp. 52–55, 2006.

[84] 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, 2007.

[85] D. Pati and S. 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.

[86] D. Pati and S. R. M. Prasanna, "Speaker information from sub-band energies of linear prediction residual," in *National Conference on Communications (NCC), 2010*, Jan 2010, pp. 1–4.

[87] A. G. Ramakrishnan, B. Abhiram, and S. R. Mahadeva Prasanna, "Voice source characterization using pitch synchronous discrete cosine transform for speaker identification," *The Journal of the Acoustical Society of America*, vol. 137, no. 6, pp. EL469–EL475, 2015.

[88] R. K. Das, S. Abhiram, S. M. Prasanna, and A. Ramakrishnan, "Combining source and system information for limited data speaker verification," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[89] Z. Wu and N. E. Huang, "Ensemble empirical mode decomposition: a noise-assisted data analysis method," *Advances in adaptive data analysis*, vol. 1, no. 01, pp. 1–41, 2009.

[90] J.-R. Yeh, J.-S. Shieh, and N. E. Huang, "Complementary ensemble empirical mode decomposition: A novel noise enhanced data analysis method," *Advances in adaptive data analysis*, vol. 2, no. 02, pp. 135–156, 2010.

[91] M. A. Colominas, G. Schlotthauer, and M. E. Torres, "Improved complete ensemble EMD: A suitable tool for biomedical signal processing," *Biomedical Signal Processing and Control*, vol. 14, pp. 19–29, 2014.

[92] N. E. Huang, *Hilbert-Huang transform and its applications.* World Scientific, 2014, vol. 16.

[93] R. Sharma, L. Vignolo, G. Schlotthauer, M. A. Colominas, H. L. Rufiner, and S. R. M. Prasanna, "Empirical mode decomposition for adaptive AM-FM analysis of speech: a review," *Speech Communication*, vol. 88, pp. 39–64, 2017.

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

[95] 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.

[96] R. Piyare and M. Tazil, "Bluetooth based home automation system using cell phone," in *Consumer Electronics (ISCE), 2011 IEEE 15th International Symposium on.* IEEE, 2011, pp. 192–195.

[97] C. Onukwugha and P. Asagba, "Remote control of home appliances using mobile phone: A polymorphous based system," *African Journal of Computing & ICT*, vol. 6, no. 5, 2013.

[98] C. K. Das, M. Sanaullah, H. M. G. Sarower, and M. M. Hassan, "Development of a cell phone based remote control system: an effective switching system for controlling home and office appliances," *International Journal of Electrical & Computer Sciences IJECS*, vol. 9, no. 10, pp. 37–43, 2009.

[99] R. Shahriyar, E. Hoque, S. Sohan, I. Naim, M. M. Akbar, and M. K. Khan, "Remote controlling of home appliances using mobile telephony," *International Journal of Smart Home*, vol. 2, no. 3, pp. 37–54, 2008.

[100] C. Tsao and R. Gray, "An endpoint detector for LPC speech using residual error look-ahead for vector quantization applications," in *Acoustics, Speech, and Signal Processing, IEEE Int. Conf. on ICASSP'84.*, vol. 9. IEEE, 1984, pp. 97–100.

[101] M. H. Savoji, "A robust algorithm for accurate end pointing of speech signals," *Speech communication*, vol. 8, no. 1, pp. 45–60, 1989.

[102] K. T. Deepak and S. R. M. Prasanna, "Foreground speech segmentation and enhancement using glottal closure instants and mel cepstral coefficients," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 24, no. 7, pp. 1205–1219, 2016.

[103] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. on acoustics, speech, and signal processing*, vol. 33, no. 2, pp. 443–445, 1985.

[104] A. Paul, D. Mahanta, R. K. Das, R. K. Bhukya, and S. Prasanna, "Presence of speech region detection using vowel-like regions and spectral slope information," in *2017 14th IEEE India Council Int. Conf. (INDICON).* IEEE, 2017, pp. 1–5.

[105] D. Mahanta, A. Paul, R. K. Bhukya, R. K. Das, R. Sinha, and S. R. M. Prasanna, "Warping path and gross spectrum information for speaker verification under degraded condition," in *Communication (NCC), 2016 Twenty Second National Conference on.* IEEE, 2016, pp. 1–6.

[106] M. Pandit and J. Kittler, "Feature selection for a dtw-based speaker verification system," in *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE Int. Conf. on*, vol. 2. IEEE, 1998, pp. 769–772.

[107] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech communication*, vol. 52, no. 1, pp. 12–40, 2010.

[108] L. R. Rabiner, A. E. Rosenberg, and S. E. Levinson, "Considerations in dynamic time warping algorithms for discrete word recognition," *The Journal of the Acoustical Society of America*, vol. 63, no. S1, pp. S79–S79, 1978.

[109] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 26, no. 1, pp. 43–49, 1978.

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

[111] R. Sharma, L. Vignolo, G. Schlotthauer, M. Colominas, H. L. Rufiner, and S. R. M. Prasanna, "Empirical mode decomposition for adaptive AM-FM analysis of speech: A review," *Speech Communication*, vol. 88, pp. 39 – 64, 2017.

[112] R. S. Holambe and M. S. Deshpande, *Advances in non-linear modeling for speech processing.* Springer Science & Business Media, 2012.

[113] T. Ganchev, N. Fakotakis, and G. Kokkinakis, "Comparative evaluation of various MFCC implementations on the speaker verification task," in *Proceedings of the SPECOM*, vol. 1, 2005, pp. 191–194.

[114] F. Zheng, G. Zhang, and Z. Song, "Comparison of different implementations of MFCC," *Journal of Computer Science and Technology*, vol. 16, no. 6, pp. 582–589, 2001.

[115] M. Skowronski and J. Harris, "Exploiting independent filter bandwidth of human factor cepstral coefficients in automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 116, no. 3, pp. 1774–1780, Sept 2004.

[116] H. Yeganeh, S. Ahadi, S. Mirrezaie, and A. Ziaei, "Weighting of mel sub-bands based on SNR/Entropy for robust ASR," in *Signal Processing and Information Technology, 2008. ISSPIT 2008. IEEE International Symposium on*, Dec. 2008, pp. 292–296.

[117] Z. Wu and Z. Cao, "Improved MFCC-Based Feature for Robust Speaker Identification," *Tsinghua Science & Technology*, vol. 10, no. 2, pp. 158–161, 2005.

[118] T. Bäck, *Evolutionary algorithms in theory and practice: evolution strategies, evolutionary programming, genetic algorithms.* Oxford, UK: Oxford University Press, 1996.

[119] L. D. Vignolo, H. L. Rufiner, D. H. Milone, and J. C. Goddard, "Evolutionary Splines for Cepstral Filterbank Optimization in Phoneme Classification," *EURASIP Journal on Advances in Signal Proc.*, vol. 2011, pp. 8:1–8:14, 2011.

[120] C. Charbuillet, B. Gas, M. Chetouani, and J. Zarader, "Optimizing feature complementarity by evolution strategy: Application to automatic speaker verification," *Speech Communication*, vol. 51, no. 9, pp. 724–731, 2009.

[121] L. D. Vignolo, D. H. Milone, and H. L. Rufiner, "Genetic wavelet packets for speech recognition," *Expert Systems with Applications*, vol. 40, no. 6, pp. 2350–2359, 2013.

[122] L. Vignolo, H. Rufiner, and D. Milone, "Multi-objective optimisation of wavelet features for phoneme recognition," *IET Signal Processing*, March 2016.

[123] H. Beigi, *Speaker Recognition: Advancements and Challenges.* INTECH Open Access Publisher, 2012.

[124] J. Dang and K. Honda, "Acoustic characteristics of the piriform fossa in models and humans," *The Journal of the Acoustical Society of America*, vol. 101, no. 1, pp. 456–465, 1997.

[125] X. Lu and J. Dang, "An investigation of dependencies between frequency components and speaker characteristics for text-independent speaker identification," *Speech communication*, vol. 50, no. 4, pp. 312–322, 2008.

[126] T. Kitamura, K. Honda, and H. Takemoto, "Individual variation of the hypopharyngeal cavities and its acoustic effects," *Acoustical science and technology*, vol. 26, no. 1, pp. 16–26, 2005.

[127] S. Hayakawa and F. Itakura, "Text-dependent speaker recognition using the information in the higher frequency band," in *Acoustics, Speech, and Signal Processing, 1994. ICASSP-94., 1994 IEEE Int. Conf. on*, vol. 1. IEEE, 1994, pp. I–137.

[128] X. Lu and J. Dang, "Physiological feature extraction for text independent speaker identification using non-uniform sub-band processing," in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE Int. Conf. on*, vol. 4. IEEE, 2007, pp. IV–461.

[129] C. Jankowski, T. F. Quatieri, and D. A. Reynolds, "Measuring fine structure in speech: Application to speaker identification," in *Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 Int. Conf. on*, vol. 1. IEEE, 1995, pp. 325–328.

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

[131] M. S. Deshpande and R. S. Holambe, "Speaker identification based on robust AM-FM features," in *2nd Int. Conf. on Emerging Trends in Engineering and Technology (ICETET).* IEEE, 2009, pp. 880–884.

[132] T. Hasan and J. H. Hansen, "Robust speaker recognition in non-stationary room environments based on empirical mode decomposition." in *INTERSPEECH*, 2011, pp. 2733–2736.

[133] J.-D. Wu and Y.-J. Tsai, "Speaker identification system using empirical mode decomposition and an artificial neural network," *Expert Systems with Applications*, vol. 38, no. 5, pp. 6112–6117, 2011.

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

[135] R. Sharma and S. R. M. Prasanna, "A better decomposition of speech obtained using modified empirical mode decomposition," *Digital Signal Processing*, vol. 58, pp. 26 – 39, 2016.

[136] N. Chatlani and J. J. Soraghan, "EMD-based filtering EMDF of low-frequency noise for speech enhancement," *Audio, Speech, and Language Processing, IEEE Trans. on*, vol. 20, no. 4, pp. 1158–1166, 2012.

[137] K. Khaldi, A.-O. Boudraa, A. Bouchikhi, and M. T.-H. Alouane, "Speech enhancement via EMD," *EURASIP Journal on Advances in Signal Processing*, vol. 2008, no. 1, p. 873204, 2008.

[138] A. Bouchikhi and A.-O. Boudraa, "Multicomponent am–fm signals analysis based on EMD–B-splines ESA," *Signal Processing*, vol. 92, no. 9, pp. 2214–2228, 2012.

[139] K. Khaldi and A.-O. Boudraa, "On signals compression by EMD," *Electronics letters*, vol. 48, no. 21, pp. 1329–1331, 2012.

[140] K. Khaldi and A. Boudraa, "Audio watermarking via EMD," *Audio, Speech, and Language Processing, IEEE Trans. on*, vol. 21, no. 3, pp. 675–680, 2013.

[141] K. Khaldi, A.-O. Boudraa, and A. Komaty, "Speech enhancement using empirical mode decomposition and the teager–kaiser energy operator," *The Journal of the Acoustical Society of America*, vol. 135, no. 1, pp. 451–459, 2014.

[142] K. Khaldi, A.-O. Boudraa, B. Torresani, and T. Chonavel, "Hht-based audio coding," *Signal, image and video processing*, vol. 9, no. 1, pp. 107–115, 2015.

[143] K. Khaldi, A.-O. Boudraa, and M. Turki, "Voiced/unvoiced speech classification-based adaptive filtering of decomposed empirical modes for speech enhancement," *IET Signal Processing*, vol. 10, no. 1, pp. 69–80, 2016.

[144] J.-C. Cexus and A.-O. Boudraa, "Nonstationary signals analysis by teager-huang transform (tht)," in *Signal Processing Conference, 2006 14th European*. IEEE, 2006, pp. 1–5.

[145] K. Khaldi, M. T.-H. Alouane, and A.-O. Boudraa, "A new EMD denoising approach dedicated to voiced speech signals," in *Signals, Circuits and Systems SCS-2008. 2nd Int. Conf. on*. IEEE, 2008, pp. 1–5.

[146] K. Khaldi, A.-O. Boudraa, M. Turki, T. Chonavel, and I. Samaali, "Audio encoding based on the empirical mode decomposition," in *Signal Processing Conference, 2009 17th European*. IEEE, 2009, pp. 924–928.

[147] K. Khaldi, A.-O. Boudraa, B. Torresani, T. Chonavel, and M. Turki, "Audio encoding using huang and hilbert transforms," in *Communications, Control and Signal Processing (ISCCSP), 2010 4th International Symposium on*. IEEE, 2010, pp. 1–5.

[148] A. Bouzid and N. Ellouze, "Voiced speech analysis by empirical mode decomposition," in *Advances in Nonlinear Speech Processing*. Springer, 2007, pp. 213–220.

[149] Z. Yang, D. Huang, and L. Yang, "A novel pitch period detection algorithm based on hilbert-huang transform," in *Advances in Biometric Person Authentication*. Springer, 2005, pp. 586–593.

[150] H. Huang and J. Pan, "Speech pitch determination based on hilbert-huang transform," *Signal Processing*, vol. 86, no. 4, pp. 792–803, 2006.

[151] X. Li and X. Li, "Speech emotion recognition using novel HHT-TEO based features," *Journal of Computers*, vol. 6, no. 5, pp. 989–998, 2011.

[152] M. Przybocki and A. Martin, "The nist year 2003 speaker recognition evaluation plan," 2003.

[153] G. Wang, X.-Y. CHEN, F.-L. Qiao, Z. Wu, and N. E. Huang, "On intrinsic mode function," *Advances in Adaptive Data Analysis*, vol. 2, no. 03, pp. 277–293, 2010.

[154] Z. Wu and N. E. Huang, "Ensemble empirical mode decomposition: a noise-assisted data analysis method," *Advances in adaptive data analysis*, vol. 1, no. 01, pp. 1–41, 2009.

[155] R. Sharma and S. R. M. Prasanna, "Characterizing glottal activity from speech using empirical mode decomposition," in *Communications (NCC), 21 st National Conference on*. IEEE, 2015, pp. 1–6.

[156] S. D. Dhingra, G. Nijhawan, and P. Pandit, "Isolated speech recognition using MFCC and DTW," *International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering*, vol. 2, no. 8, pp. 4085–4092, 2013.

[157] J. Ding, C.-T. Yen, and D.-C. Ou, "A method to integrate GMM, SVM and DTW for speaker recognition," *International Journal of Engineering and Technology Innovation*, vol. 4, no. 1, pp. 38–47, 2014.

[158] A. Ouzounov *et al.*, "Cepstral features and text-dependent speaker identification-a comparative study," *Cybernetics and Information Technologies*, vol. 10, no. 1, pp. 3–12, 2010.

[159] I. Ding Jr, C.-T. Yen, and Y.-M. Hsu, "Developments of machine learning schemes for dynamic time-wrapping-based speech recognition," *Mathematical Problems in Engineering*, vol. 2013, 2013.

[160] L. Muda, M. Begam, and I. Elamvazuthi, "Voice recognition algorithms using mel frequency cepstral coefficient MFCC and dynamic time warping DTW techniques," *arXiv preprint arXiv:1003.4083*, 2010.

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

[162] P. Kenny, G. Boulianne, and P. Dumouchel, "Eigenvoice modeling with sparse training data," *IEEE Trans. on speech and audio processing*, vol. 13, no. 3, pp. 345–354, 2005.

[163] A. O. Hatch, S. Kajarekar, and A. Stolcke, "Within-class covariance normalization for SVM-based speaker recognition," in *Ninth Int. Conf. on spoken language processing*, 2006.

[164] J. J. Godfrey, E. C. Holliman, and J. McDaniel, "SWITCHBOARD: Telephone speech corpus for research and development," in *Acoustics, Speech, and Signal Processing ICASSP-92, IEEE Int. Conf. on*, vol. 1. IEEE, 1992, pp. 517–520.

[165] M. Brookes, "Voicebox," *Speech Processing Toolbox for Matlab, Department of Electrical & Electronic Engineering, Imperial College*, 2009.

[166] "https://onlinecourses.science.psu.edu/stat505/node/63." [Online]. Available: https://onlinecourses.science.psu.edu/stat505/node/63

[167] "http://htk.eng.cam.ac.uk/." [Online]. Available: http://htk.eng.cam.ac.uk/

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

[169] L. P. Wong and M. Russell, "Text-dependent speaker verification under noisy conditions using parallel model combination," in *Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01). 2001 IEEE Int. Conf. on*, vol. 1. IEEE, 2001, pp. 457–460.

[170] L. R. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition.* Prentice-Hall, 1993.

[171] L. Lamel, L. Rabiner, A. Rosenberg, and J. Wilpon, "An improved endpoint detector for isolated word recognition," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 29, no. 4, pp. 777–785, 1981.

[172] G. Saha, S. Chakroborty, and S. Senapati, "A new silence removal and endpoint detection algorithm for speech and speaker recognition applications," in *Proceedings of the 11th national conference on communications (NCC)*, 2005, pp. 291–295.

[173] M. Hamada, Y. Takizawa, and T. Norimatsu, "A noise robust speech recognition system," in *First Int. Conf. on Spoken Language Processing*, 1990.

[174] B. D. Sarma and S. R. M. Prasanna, "Analysis of spurious vowel-like regions (vlrs) detected by excitation source information," in *India Conference (INDICON), 2013 Annual IEEE.* IEEE, 2013, pp. 1–5.

[175] S. R. M. Prasanna, "Event-based analysis of speech," Ph.D. dissertation, Department of Computer Science and Engineering, IIT Madras, 2004.

[176] B. D. Sarma, S. S. Prajwal, and S. R. M. Prasanna, "Improved vowel onset and offset points detection using bessel features," in *Signal Processing and Commun (SPCOM), Int. Conf. on.* IEEE, 2014, pp. 1–6.

[177] B. Prasad, R. S. and Yegnanarayana, "Acoustic segmentation of speech using zero time liftering (ztl)." in *INTERSPEECH*, 2013, pp. 2292–2296.

[178] D. N. Gowda, "Signal processing for excitation-based analysis of acoustic events in speech," Ph.D. dissertation, Department of Computer Science and Engineering, IIT Madras, 2011.

[179] Y. Bayya and D. N. Gowda, "Spectro-temporal analysis of speech signals using zero-time windowing and group delay function," *Speech Communication*, vol. 55, no. 6, pp. 782–795, 2013.

[180] K. T. Deepak, B. D. Sarma, and S. R. M. Prasanna, "Foreground speech segmentation using zero frequency filtered signal," in *13th Annual Conference of the International Speech Communication Association*, 2012.

[181] N. Dhananjaya and B. Yegnanarayana, "Voiced/nonvoiced detection based on robustness of voiced epochs," *Signal Processing Letters, IEEE*, vol. 17, no. 3, pp. 273–276, 2010.

[182] B. D. Sarma, S. R. M. Prasanna, and P. Sarmah, "Consonant-vowel unit recognition using dominant aperiodic and transition region detection," *Speech Communication*, vol. 92, pp. 77–89, 2017.

[183] B. K. Khonglah, R. K. Bhukya, and S. R. M. Prasanna, "Processing degraded speech for text dependent speaker verification," *International Journal of Speech Technology*, vol. 20, no. 4, pp. 839–850, 2017.

[184] R. Sharma, R. K. Bhukya, and S. R. M. Prasanna, "Analysis of the hilbert spectrum for text-dependent speaker verification," *Speech Communication*, vol. 96, pp. 207–224, 2018.

[185] R. Sharma and S. R. M. Prasanna, "A better decomposition of speech obtained using modified empirical mode decomposition," *Digital Signal Processing*, vol. 58, pp. 26–39, 2016.

[186] T. Dutta, "Dynamic time warping based approach to text-dependent speaker identification using spectrograms," in *Congress on Image and Signal Processing, 2008. CISP'08.*, vol. 2. IEEE, 2008, pp. 354–360.

[187] J.-C. Junqua, B. Reaves, and B. Mak, "A study of endpoint detection algorithms in adverse conditions: incidence on a DTW and HMM recognizer," in *Second European conference on speech communication and technology*, 1991.

[188] H. Kremer, A. Cohen, and T. Vaich, "Voice activity detector VAD for HMM based speech recognition," in *Proceedings of ICSPAT*, 1999.

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

[190] L. D. Alsteris and K. K. Paliwal, "Further intelligibility results from human listening tests using the short-time phase spectrum," *Speech Communication*, vol. 48, no. 6, pp. 727–736, 2006.

[191] H. Boril, P. Fousek, and P. Pollák, "Data-driven design of front-end filter bank for lombard speech recognition," in *Ninth Int. Conf. on Spoken Language Processing*, 2006.

[192] N. E. Huang, "Empirical mode decomposition and hilbert spectral analysis," 1998.

[193] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The det curve in assessment of detection task performance," National Inst of Standards and Technology Gaithersburg MD, Tech. Rep., 1997.

List of Publications

Journal Publications

- Published Papers:

1. Ramesh K. Bhukya, S. R. M. Prasanna and Biswajit Dev Sarma, "Robust Methods for Text-Dependent Speaker Verification", **Circuits, Systems, and Signal Processing, (Springer)**, vol. 38, no. 11, pp. 5253–5288, November 2019.
2. Ramesh K. Bhukya, Biswajit Dev Sarma and S. R. M. Prasanna, "End Point Detection Using Speech Specific Knowledge for Text-Dependent Speaker Verification", **Circuits, Systems, and Signal Processing (Springer)**, vol. 37, no. 12, pp. 5507 - 5539, May 2018.
3. Rajib Sharma, Ramesh K. Bhukya and S. R. M. Prasanna, "Analysis of the Hilbert Spectrum for Text-Dependent Speaker Verification", **Speech Communication (Elsevier)**, vol. 96, pp. 207-224, February 2018.
4. Banriskhem K Khonglah, Ramesh K. Bhukya and S. R. M. Prasanna, "Processing Degraded Speech for Text-Dependent Speaker Verification," **International Journal of Speech Technology (Springer)**, vol. 20, pp. 839-850, December 2017.
5. Rajib Sharma, S. R. M. Prasanna, Ramesh K. Bhukya and R.K.Das, "Analysis of the Intrinsic Mode Functions for Speaker Information", **Speech Communication (Elsevier)**, vol. 91, pp. 1-16, July 2017.

- Manuscripts Under Review

1. Ramesh K. Bhukya and S. R. M. Prasanna, "Some Issues in the Practical Deployment of the Text-Dependent Speaker Verification", **Under Review in**, Journal of Signal Processing Systems, Springer, submitted in October 2018.
2. Ramesh K. Bhukya, S. R. M. Prasanna and Sarfaraz Jelil, "Text-Dependent Speaker Verification Using DTW-GMM Based Two-tier Authentication", **Under Review in**, Expert Systems with Applications, Elsevier, submitted in November 2018.
3. Ramesh K. Bhukya and S. R. M. Prasanna, "End Point Detection, Speech Enhancement and Feature Extraction - A Review", **to be submitted in IETE Technical Review**.
4. Ramesh K. Bhukya and S. R. M. Prasanna, " Approaches for Robust Methods for End-to-End Text-Dependent Speaker Verification Under Practical Conditions", **to be submitted in**, **Circuits, Systems, and Signal Processing, (Springer)**.
5. Ramesh K. Bhukya and S. R. M. Prasanna, "Deep Learning Approaches for End-to-End Robust Text-Dependent Speaker Verification Under Degraded Conditions", **to be submitted in**, **IET Biometrics**.
6. Ramesh K. Bhukya and S. R. M. Prasanna, " Robust Voice Liveness Detection and Text-Dependent Speaker Verification Using Throat Microphone", **to be submitted in**, **IET Signal Processing**.

Conference Publications

- Published Paper and Accepted Publication:

  1. S. Dey, S. Barman, Ramesh K. Bhukya, R.K.Das, Haris B. C, S. R. M. Prasanna and R. Sinha, "Speech Biometric Based Attendance System", *in Proc. National Conf. on Communication (NCC)*, IITK, Kanpur, India, February 2014.
  2. D. Mahanta, A. Paul, Ramesh K. Bhukya, R.K.Das, R.Sinha and S. R. M. Prasanna,"Warping Path and Gross Spectrum Information for Speaker Verification Under Degraded Condition", *in Proc. National Conf. on Communication (NCC),* IITG, Guwahati, India, 2016.
  3. A. Paul, D. Mahanta, R.K.Das, Ramesh K. Bhukya and S. R. M. Prasanna, " Presence of Speech Region Detection Using Vowel-like Regions and Spectral Slope Information", *in Proc. IEEE India International Conference (INDICON)*, IITR, Roorkee, India, December 2017.