Subsegmental, Segmental and Suprasegmental Processing of Linear Prediction Residual for Speaker Information

A

Thesis submitted

for the award of the degree of

DOCTOR OF PHILOSOPHY

By

DEBADATTA PATI

DEPARTMENT OF ELECTRONICS AND ELECTRICAL ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY GUWAHATI

GUWAHATI - 781 039, INDIA

DECEMBER 2011

To my guide **Dr. S. R. Mahadeva Prasanna**

for his help, guidance, inspiration and encouragement

and

My Wife and our Children

for their love, sacrifice and support

Certificate

This is to certify that the thesis entitled "**SUBSEGMENTAL, SEGMENTAL AND SUPRASEGMENTAL PROCESSING OF LINEAR PREDICTION RESIDUAL FOR SPEAKER INFORMATION**", submitted by **Debadatta Pati** (07610209), 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.

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

Acknowledgements

I take the honour and privilege to express my deep sense of gratitude to my esteemed guide Dr. S. R. Mahadeva Prasanna, Department of Electronics and Electrical Engineering, Indian Institute of Technology Guwahati for his scholarly instructive guidance, support and continued encouragement throughout my PhD work. Without our inspiring discussions on certain issues at times, I would have never been able to complete this work. I greatly admire his attitude towards research, creative thinking, hard work, dedication and punctuality in work. I am highly grateful to him for patiently checking all my manuscripts and thesis.

I sincerely thank to my doctoral committee chairman Prof. S. Dandapat for his support, encouragement and suggestions rendered during my research work. I am also very much thankful to my doctoral committee members Dr. Rohit Sinha and Dr. P. K. Das for their useful advice and for sparing their valuable time to evaluate the progress of my work.

I thank the Head of the Department, Prof. S. Majhi, and other faculty members for their help and support in carrying out my work. My thanks go to Mr. L. N. Sharma sir for providing an excellent facility and maintaining a good ambience in the Electro Medical and Speech Technology lab to carry out my work. His friendly and helpful nature has made my work easy. I sincerely thank Mr. Sanjib Das sir for his timely help. My thanks go to Mrs. Jharna madam for her help in setting the computers for my experiments in System Simulation lab.

I extend my sincere gratitude to my employer Balasore College of Engineering and Technology (BCET), Balasore, especially Sj. M. K. Biswal, Chairman, for sponsoring me, that made my PhD work possible at Indian Institute of Technology Guwahati (IITG), a prestigious institute in India. I also thank all staff of BCET for their constant support during this period.

I gratefully acknowledge All India Council for Technical Education (AICTE) for providing opportunity to do PhD at IIT Guwahati under QIP scheme.

I am thankful to the UK-INDIA Education and Research Initiative (UKIERI) project titled "Study of source features for speech synthesis and speaker recognition" between IIT Guwahati, IIIT Hyderabad and CSTR, University of Edinburgh, UK for providing the financial support

for attending and presenting my work in various conferences and an opportunity to work as a associate project engineer during the last part of my research work.

I sincerely thank Mr. R. C. Mishra sir who helped me a lot from all respects since the beginning of entering the IIT Guwahati campus. When I face any problem, his advice like, "*Don't worry, everything will be fine*", was indeed a quick relief to me.

I sincerely thank to Mrs. Padam Priyal for her help in teaching me for my minor comprehensive examination and also for my research work.

I sincerely thank Mrs. Nirmala madam for her help in my research work. She allowed me to use her computer as if it was mine.

I sincerely thank to Mr. Gaydhar Pradhan, Mr. Ratnakar Das, Mr. Mukul Bora and their family for brotherly care and help during my stay at IIT Guwahti.

I thank Mr. H. S. Jayanna, Mr. P. Krishnamurthy, Mr. M. Sabarimali Manikandan, Mr. K. Narasimha Murthy, Ms. Sumitra Shukla, Mr. D. Govind, Mrs. Shweta Ghai, Mr. Haris, Mr. Abhinav Mishra, Mr. Sumit Shulka and all other members in the lab for their help and support during my research work.

I thank my parents (Bou and Nana) for their affection, advice and moral support in my life and making me what I am today. I thank my brother, sisters, in-laws and all relatives for their love, affection and support in pursuing my research work.

I do not have words to express my gratitude to my wife *Bebina* and our children *Budha* and *Budhi* for their strong support, sacrifice, affection, patience and tolerance towards me during my stay at IIT Guwahati and hence I sincerely dedicate this work to them.

I am morally indebted to all the authors who have been quoted in this work. Without their words, to go by this research work would not have been feasible. My heartfelt thanks to all those who directly or indirectly assisted me during my PhD work whose names could not be mentioned here.

Debadatta Pati

Abstract

The speaker-specific information in speech is mostly attributed to the shape, size and dynamics of the vocal tract and excitation source. The excitation information can be viewed at *subsegmental* (3-5 msec), *segmental* (10-30 msec) and *suprasegmental* (100-300 msec) levels. These include glottal cycle activities, periodicity and strength of vocal folds vibration, and speaker learning habits. This work proposes methods to model these information from the linear prediction (LP) residual and uses them in a combined fashion to develop a speaker verification system.

The significance and different nature of the subsegmental, segmental and suprasegmental excitation information present in the LP residual are demonstrated by processing it directly in the time domain. The segmental excitation information provides the best performance followed by subsegmental level. Due to large intra-speaker variability and also text-independent mode, the suprasegmental excitation information provides the least performance. The combined evidence from each of these levels further improves the performance. For compact and effective representation, different methods of parameterizing the LP residual are explored. We found that Liljencrants-Fant (LF) parameters computed from the LP residual, combined use of spectral flatness measure and cepstral coefficients from LP residual mel warped spectrum, combined use of pitch, epoch strength and LP residual mel frequency cepstral trajectories are proposed as the possible ways of representing the subsegmental, segmental and suprasegmental excitation information, respectively. The vocal tract based system provides relatively good performance, but suffers severely in noisy conditions. In this sense, the proposed excitation information based system is relatively more robust and hence may be useful.

The contributions of the work reported in this thesis for subsegmental, segmental and suprasegmental processing of LP residual for speaker information include,

- Implicit processing of the the LP residual in the time domain with different frame size and shift for the extraction of subsegmental, segmental and suprasegmental speaker-specific excitation information.
- Implicit processing of the analytic representation of the LP residual in the time domain for independent modeling of amplitude and sequence information.
- Modification suggested to *zero-frequency filtering* method for accurate estimation of pitch and epoch strength from telephone speech.
- Proposed efficient approach for the computation of the LF parameters.
- Compact representation of the subsegmental excitation information by using LF parameters.
- Investigation on filter shapes for processing the LP residual in frequency and cepstral domains for compact representation of the segmental excitation information.
- Effective modelling of the suprasegmental excitation information by the combined use of pitch, epoch strength and cepstral trajectory vectors.
- Development of the speaker verification system using excitation information.

Keywords: Subsegmental, segmental, suprasegmental, LP residual, vocal tract and excitation information, LF parameters, mel warped spectrum, speaker recognition.

Contents

List of Figures

List of Tables

List of Acronyms

AANN	Auto Associative Neural Networks
APQ	Amplitude Perturbation Quotient
CMS	Cepstral Mean Substraction
CQ	Close Quotient
DET	Detection Error Trade-off
DFT	Discrete Fourier Transform
EER	Equal Error Rate
EGG	Electro Glotto Graph
ERR	Error Rate Reduction
FA	False Acceptance
FAR	False Acceptance Ratio
FFT	Fast Fourier Transform
FR	False Rejection
FRR	False Rejection Ratio
GCI	Glottal Closure Instant
GFD	Glottal Flow Derivative
GMM	Gaussian Mixture Models
GMM-UBM	GMM-Universal Background Model
GOI	Glottal Opening Instant
HE	Hilbert Envelope
IDFT	Inverse Discrete Fourier transform
LF	Liljencrants-Fant

LFCC	Linear Frequency Cepstral Coefficient
LLR	Log-Likelihood Ratio
LP	Linear Prediction
LPC	Linear Prediction Coefficient
LPCC	Linear Prediction Cepstral Coefficient
MFCC	Mel-Frequency Cepstral Coefficient
MPDSS	Mel Power Differences of Spectrum in Subbands
OQ	Open Quotient
PDSS	Power Differences of Spectrum in Subbands
PPQ	Period Perturbation Quotient
RAP	Relative Average Perturbation
RAPT	Robust Algorithm for Pitch Tracking
RCC	Residual Cepstral Coefficient
RFFTCC	Residual Fast Fourier Transform Cepstral Coefficient
RMFCC	Residual Mel-Frequency Cepstral Coefficient
RMSE	Residual Mel Subband Energy
RP	Residual Phase
RPDSS	Residual Power Differences in Subband Spectra
RRFCC	Residual Rectangular Filter Cepstral Coefficient
RRSE	Residual Rectangular Subband Energy
RTSE	Residual Triangular Subband Energy
RQ	Return Quotient
SBE	Subband Energy
SR	Speaker Recognition
UBM	Universal Background Model
VQ	Vector Quantization
WOCOR	Wavelet Octave Coefficient of Residue
ZFFS	Zero-Frequency Filtered Signal

List of Symbols

A	Average peak amplitude over N_s extracted pitch period
A_n	Peak amplitude of the n^{th} frame
A_0	Instantaneous epoch strength
$A(z)$	Inverse filter response
a	Wavelet scaling factor
a_k	Linear prediction (LP) coefficients
α	Growth factor of glottal glow derivative (GFD)
B	Inter-covariance matrix
b	Wavelet translation factor
β	Time constant of GFD
C	Number of cepstral coefficients
C_{ir}	LP residual cepstral coefficients using rectangular filter
C_j	Number of systems combined
$C(n)$	Cepstral representation of LP residual
$Comb_1$	Linear score level combination
$Comb_2$	Logical OR combination
C_t	Combined representation of c_{t1}, c_{t2}, c_{t4} and c_{t6} trajectory vectors
χ	Feature vectors set
c_t	$RMFCC$ trajectory
DIV	Divergence
Δ	Delta coefficient
$\Delta\Delta$	Delta-Delta coefficients

$\Delta X^n(m)$	Delta measurement of the n^{th} frame
$\Delta\Delta X^n(m)$	Delta Delta measurement of the n^{th} frame
E_e	Negative peak of GFD
E_o	Arbitrary gain constant of GFD
$e_{LF}(t)$	LF model of GFD
h_i	Upper band sample point of i^{th} subband
J_T	Jitter
K_w	Wavelet order
l_i	Lower band sample point of i^{th} subband
LLR_c	Log-likelihood score of the combined system
λ_s	Speaker s model
M_b	Number of subband filters
M_g	Number of Gaussian mixture models
M_w	Number of wavelet decmposition levels
m	Mean of the distribution of χ
m_s	Mean feature vector of speaker s
μ_i	Mean Gaussian mixture model
N	Number of DFT points
N_C	Number of sample points in the subband
N_F	Number of feature vectors
N_a	Number of samples used for mean substraction
N_c	Glottal closure instant in sample
N_e	Instant of the negative peak of the GFD in sample
N_j	Number of extracted pitch periods for jitter measurement
N_o	Glottal opening instant in sample
N_s	Number of extracted pitch periods for shimmer measurement
N_w	Window length
n_s	Number of d-dimensional feature vectors in speaker s

Ω_z	Discrete counter part of ω_z
$\omega(a, b)$	Daubechies wavelet function with a and b scaling and transformation parameters
ω_i	Gaussian mixture weights
ω_o	Reciprocal of the time that elapses during end of open phase of the GFD cycle
ω_z	Reciprocal of the time for first zero-crossing of the GFD cycle
P_g	Pitch period of the g^{th} glottal cycle
P_i	Identification performance of the system i
P_n	Pitch period of n^{th} frame
$P(\lambda_c)$	Likelihood given by claimed speaker
$P(X\|\lambda_s)$	Log-likelihood score given to model λ_s for test data X
$P(\lambda_u)$	Likelihood given by universal back ground model (UBM)
$P(x_t\|\lambda_s)$	Gaussian mixture density of feature x_t
P_0	Instantaneous pitch
$p(k)$	LP residual power spectrum
$\psi(n)$	Fourth order Daubechies wavelet basis function
$R(k)$	Discrete Fourier transform of $r(n)$
$R_h(w)$	Fourier transform of $r_h(n)$
$R(w)$	Fourier transform of $r(n)$
$r(n)$	LP residual
$r_a(n)$	Analytic representation of the LP residual
$r_h(n)$	Hilbert transform of the LP residual
S	Number of speakers
$\hat{S}(n)$	Linear predicted speech signal
Seg	LP residual segmental blocks
S_H	Shimmer
$S(n)$	Speech signal
Src	Proposed representation of the excitation information
Src_1	Combination of Sub and Seg

Src_2	Combination of Sub, Seg and $Supra$
Src_3	Combination of Seg and $Supra$
Src_4	Combination of $RMFCC$ and $MPDSS$
Src_5	Combination of T_0, A_0 and C_t vectros
Src_6	Combination of $RMFCC + \Delta + \Delta\Delta$ and $MPDSS$
Sub	LP residual subsegmental blocks
$Supra$	LP residual suprasegmental blocks
Σ_i	Gaussian mixture variance
T	Average time over N_j extracted pitch period
T_a	Time constant of the GFD return phase
T_c	Instant of the glottal closing
T_{cg}	Instant of g^{th} glottal closing
T_e	Instant of the negative peak of GFD
T_o	Instant of the glottal opening
T_{og}	Instant of g^{th} glottal opening
T_z	Instant of the first zero crossing of the GFD
T_0	Instantaneous pitch period
$\theta(n)$	Cosine phase of $r_a(n)$
V_i	Power difference in subband spectrum vectors
V_{nj}	Deviation in jitter measurement
W	Intra-covariance matrix
X_1	Performance of the baseline system
$x(n)$	Difference speech signal
x_t	Test feature vector
Y_1	Performance of the proposed system
$y(n)$	Zero-frequency filtered signal of $s(n)$
$y_1(n)$	Output of the zero-frequency resonator by passing $x(n)$ once
$y_2(n)$	Output of the zero-frequency resonator by passing $x(n)$ twice

1

Introduction

Contents

1.1 Objectives of the Thesis

Speech is produced from a time-varying vocal tract system excited by a time-varying excitation source [1]. The speaker-specific information in speech is mostly attributed to the shape, size and dynamics of the vocal tract and the excitation source [2]. State-of-the-art speaker recognition systems mostly use vocal tract information represented by features like, linear prediction cepstral coefficients (LPCC) or mel frequency cepstral coefficients (MFCC) [3–5]. These cepstral features have been successfully used for speaker recognition and demonstrated to provide good performance. The reason may be that, these cepstral features mainly characterize the smooth spectral envelope and thus mostly represent the vocal tract information.

The vocal tract related features are biased towards the content of the speech and are also sensitive to environmental variations [6, 7]. In applications where training and testing data are limited and of poor quality due to varied environmental effects, the performance degrades significantly. Hence there is a need for deriving robust features for speaker recognition. Motivated by this, the other component of the speech production, namely, the excitation source has been explored for speaker recognition. Both the physiological and behaviorial aspects of the speaker present in the excitation source component like pitch and intonation have been demonstrated to contain speaker information [2, 8]. Other major attempts to model the speaker-specific information from the excitation source include processing the linear prediction (LP) residual with proper LP order [9–17]. These attempts demonstrate that the LP residual contains speaker-specific excitation information. It is also demonstrated that features derived from the LP residual are relatively more robust and require less amount of data for speaker recognition [6, 12, 18, 19].

The performance of the speaker recognition system using the excitation information, in particular, the LP residual is not at par with the vocal tract information. It may happen that the proposed features from all the existing independent attempts may not represent the complete speaker-specific excitation information. For instance, some attempts focus on modeling speaker-specific excitation information present in 3-5 msec segments, termed as *subsegmental analysis* [12, 20]. Other attempts model speaker-specific excitation information present in

10-30 msec segments, termed as *segmental analysis* [9, 10]. Yet other attempts try to model speaker-specific excitation information present in the range more than 100 msec, termed as *suprasegmental analysis* [8]. All of these represent excitation information, but may be different and incomplete. The difficulty in the complete representation using a single feature may be due to the dynamic nature of the excitation source signal. The objective of this thesis is therefore to develop an unified framework for modeling speaker-specific excitation information. By this we mean to explore different approaches for effective modelling of the excitation information at subsegmental, segmental and suprasegmental levels and combine them for better representation of the speaker-specific excitation information.

The LP analysis is a mostly used speech signal processing technique for modeling the vocal tract information in terms of LP coefficients (LPCs) and then using them for inverse filtering to derive the error signal, termed more commonly as the LP residual. By inverse filtering of the speech signal with proper LP order (say, 8-20 for 8 kHz sampled speech signal), the vocal tract information is suppressed and the resulting LP residual is demonstrated to contain mostly the speaker-specific excitation information [12]. The present work also uses the LP residual as a representation of the excitation signal and processes at subsegmental, segmental and suprasegmental levels for speaker-specific excitation information. Hence the thesis is titled as *subsegmental, segmental and suprasegmental processing of LP residual for speaker information*. The work involves developing techniques for extracting features by processing the LP residual, such that when combined will represent the near complete speaker-specific excitation source information. As a result, the thesis proposes a best possible approach for modeling speaker-specific excitation information from the LP residual.

The existing attempts to process the LP residual may be viewed at three levels, namely, subsegmental, segmental and suprasegmental levels of processing. The LP residual is processed in blocks of 3-5, 10-30 and more than 100 msec in case of subsegmental, segmental and suprasegmental levels, respectively [20]. All these attempts process only in one of these levels and then combine such information with that of the vocal tract. However, the speaker-specific information at each of these levels may be different. *Exploring this aspect is the motivation for*

the present work.

The first objective of the work will be to study the amount of speaker information present at each of these three levels. The second objective will be to study how different the speaker information at each of these levels. The third objective will be to explore different approaches for modeling the speaker information at each level and propose the best one. The fourth objective will be to develop a speaker recognition system using the information present in the LP residual at the subsegmental, segmental and suprasegmental levels. A comparison will also be made with the state-of-the art speaker recognition system using the vocal tract information to know how best we can recognize speakers using the excitation source component alone by exploiting the speaker information present at different levels of the LP residual.

1.2 Need for Modeling Speaker-Specific Excitation Information

1.2.1 Speech Production Perspective

The speech production system consists of a resonant structure termed as vocal tract and a source for exciting the resonant structure. A schematic diagram of the human speech production mechanism is shown in Figure 1.1. The major excitation source for speech production is the vocal folds with associated muscle structures in the larynx. The pharynx, oral cavity and nasal cavity constitute the resonant structure for generating different sounds. Both, the resonant structure as well as the major excitation source are unique for each speaker. The way they use them dynamically during speech production is also specific to each speaker. The excitation source and resonant structure participate equally in their role for speech production. Due to the uniqueness and also equal participation, from the speech production perspective, both the components should contribute equally to the speaker-specific information. However, most of the speaker recognition attempts exploit mainly the speaker-specific vocal tract information for speaker recognition. It will therefore be a scientific curiosity to see how well we can model the speaker-specific excitation information.

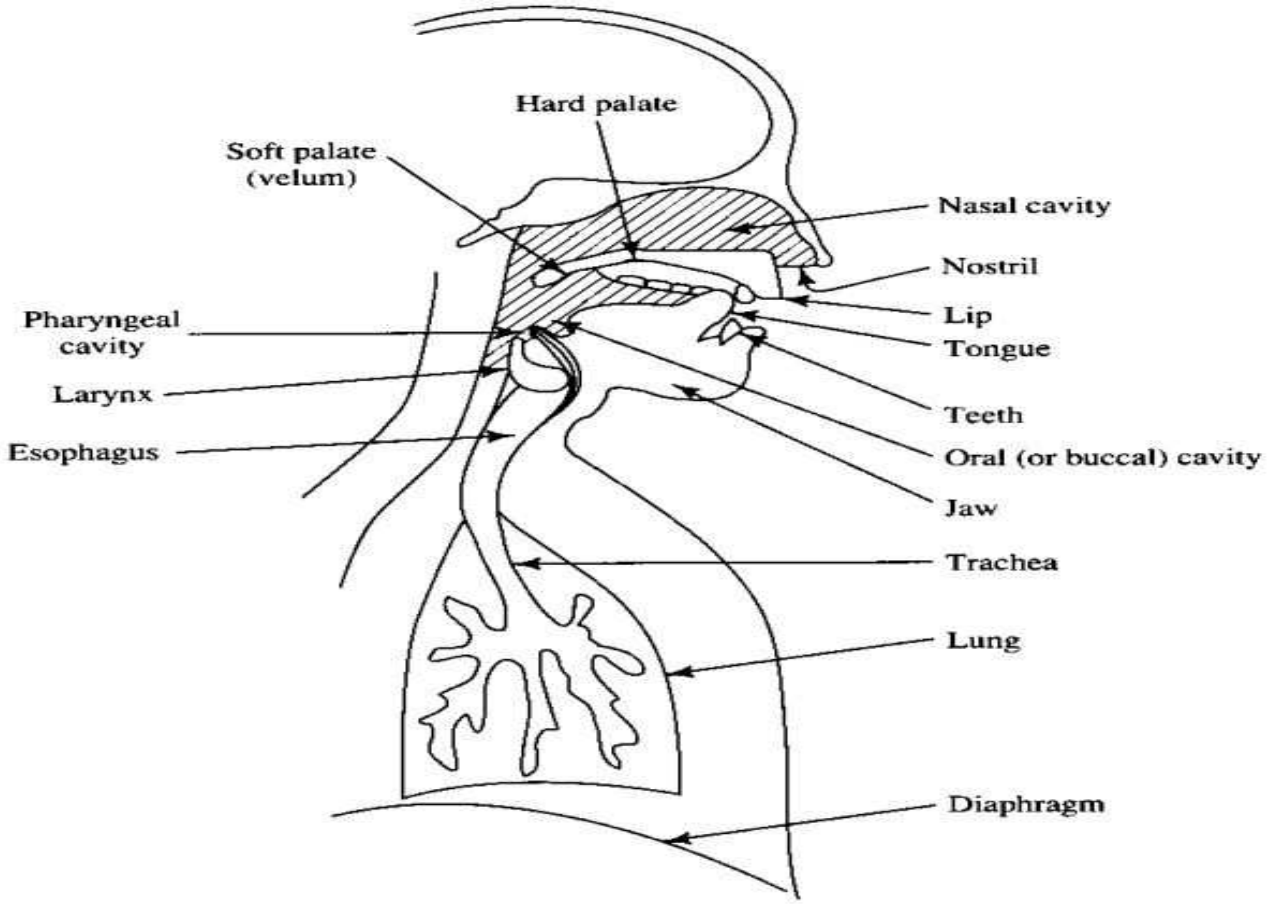

Figure 1.1: A schematic diagram of the human speech production mechanism [21].

1.2.2 Speech Perception Perspective

The first and striking merit of excitation information can be observed by comparing our experience in listening whispered and normal speech signals [22]. During whispering, the major excitation source is not used for speech production. It is done intentionally so that persons at far distance should not be able to make out the message and is meant only for the listeners sitting next. Alternatively, during normal speech production, the major excitation source is used for speech production. In such a case, the listeners sitting next as well as several meters away, as in the case of class room teaching, can indeed make out the message. Further, it is interesting to note that the message comprehension is almost same by the listeners sitting next as well as far away. This brings us to an important aspect of excitation source. That is, the excitation source component adds the required robustness to the speech signal during normal

speech production, so that the listeners sitting at several meters away also can make out the message. Thus from human listening point of view, the excitation source contributes to the robustness. The robustness issue of the existing automatic speaker recognition systems may therefore be addressed using the speaker-specific excitation information.

1.2.3 Speaker-Specific Information Perspective

By origin, the speaker-specific information contributed by the excitation source is different compared to that of the resonant structure [2]. The speaker-specific excitation information is due to the shape, size and dynamics associated with the vocal folds and associated muscle structure. The speaker-specific resonant structure information is due to the shape and size associated with pharynx, oral and nasal cavities and also dynamics associated with the oral cavity. Thus the speaker-specific information from each of them is different. Benefit in terms of recognition performance may be better by combining the speaker-specific information from these two components. The improvement in the performance is directly proportional to the aspect of how well the information from the individual component is modeled. Better modeling of speaker-specific excitation information will in turn improve the overall performance of the speaker recognition system.

1.2.4 Signal Processing Perspective

The mostly used representative for excitation source signal is the LP residual derived using proper LP order. From the theory of LP analysis, the LPCs are estimated by exploiting the second order relations among the speech samples, like autocorrelation analysis [23, 24]. Thus when we do inverse filtering, the relations up to second order of the speech samples are removed in the LP residual. As a result the LP residual looks like a noise sequence. However, the perceptual studies on the LP residual have reported that it is indeed possible to recognize speakers by listening to their speech signal [25, 26]. Therefore it will be a signal processing challenge to process the LP residual for modeling the speaker-specific information present in the noise-like sequence.

1.3 Subsegmental, Segmental and Suprasegmental Processing of LP Residual

As mentioned earlier, the existing methods for processing the LP residual may be broadly grouped under subsegmental, segmental and suprasegmental levels [20]. There are attempts made to model the speaker information present in the LP residual by viewing it in blocks of 3-5 msec. Since the block size is less than 10 msec, the processing is termed as subsegmental processing. The typical attempts include processing of LP residual or its phase using autoassociate neural network (AANN) models [11–13,27,28]. The motivation behind subsegmental level processing is to capture the speaker-specific excitation information present within each glottal cycle or pitch period. This information typically represents excitation source activity during closing, closed, opening and opened phase regions of each glottal cycle. Specifically, the subsegmental processing is meant to capture the speaker-specific excitation information, excluding pitch and its contour. The results from these studies demonstrate that the subsegmental level of processing contains significant speaker-specific excitation information.

Attempts have also been made to capture speaker-specific excitation information present in blocks of 10-30 msec of LP residual and hence viewed under segmental processing. These include modeling residual spectrum by cepstrum, power differences of spectrum in subband (PDSS), wavelet analysis and so on [9, 10, 16, 17, 29]. In all these attempts, the objective is to obtain a compact and enhanced representation of speaker-specific excitation information present at the segmental level. In segmental processing the LP residual information present across 10-30 msec, typically, 2-3 glottal cycles or pitch periods are viewed simultaneously to model the common information present across the cycles. The dominating information present at the segmental level of processing will be periodicity, harmonic structure and segmental energy. These attempts have demonstrated that the segmental level of processing also contains significant speaker-specific excitation information.

At the next level, the information in the LP residual is viewed in blocks of 100 msec or more. Since the block size is more to be treated under segmental, it is termed as suprasegmental

level processing of LP residual. The typical approach for suprasegmental processing is to first process the LP residual by the segmental processing to extract relevant information that can be viewed at the suprasegmental level. For instance, segmental processing is performed on the LP residual initially to extract the pitch and then the pitch values of successive segmental blocks are plotted to obtain the pitch contour. The pitch contour is viewed as the speaker-specific excitation information at the suprasegmental level [8]. In the similar way, several other information can be obtained to view the speaker information at the suprasegmental level [19]. The observations from the existing attempts for suprasegmental processing infer that these attempts also contain good amount of speaker-specific excitation information. However, the suprasegmental processing suffers from large intra-speaker variability.

As briefly reviewed above, the subsegmental, segmental and suprasegmental levels of processing of LP residual model in an independent manner and demonstrate that each level has good or significant amount of speaker-specific excitation information. However, to the best of our knowledge, there are no systematic and visible efforts in the literature to explore whether each level has different information. If they are different, then each of these attempts are modeling only one component of speaker-specific excitation information. If so, can we combine them to achieve further improvement in modelling the speaker-specific excitation information for speaker recognition. Hence the need for such an attempt.

1.4 Speaker Recognition Terminologies

1.4.1 Automatic Speaker Recognition

Automatic speaker recognition (SR) is the task of recognizing people by machine using the information available in their speech [30, 31]. Depending upon the task objective, SR is classified into *speaker verification and identification* [30,31]. In speaker verification, the machine validates the claimed identity of an unknown speaker. It is a process of one to one comparison. In speaker identification, the machine searches the identity of the unknown speaker present in the test speech from the given reference models. Speaker identification is further classified as *open-set and closed-set* [31]. In closed-set identification, it is assumed that the reference

models of all the users having access to that system are available. The test speaker is therefore guaranteed to be any one among the given reference models and the system will give decision accordingly. In an open-set case, there is a possibility that the test speaker may not be from the reference models available in the system. In such case the system is designed to go for a second level of test, that is, how close he/she is, which may give an additional decision like *no match* [31]. Speaker identification is a process of one-to-many comparisons.

Depending on the mode of operation, SR system is also classified as *text-dependant and text-independent* [30]. If speech of same text is used for building the speaker model and later for comparison, then it is called as text-dependant SR system. In text-independent case there is no such constraint. Generally, text-independent speaker recognition is more difficult than text-dependant task, because one has to cope with an additional variability due to differences in the texts of the unknown and reference utterances [23]. *In this thesis all our studies will be on text-independent, closed-set speaker identification and verification tasks.*

1.4.2 Block Diagram of SR system

The basic block diagram of SR system is shown in Figure 1.2. The function of the block diagram may be divided into two phases of operation: *training phase* and *testing phase*. The *training phase* includes *feature extraction* and *modeling* blocks. The *feature extraction* block extracts the speaker-specific features from the speech signal(s) available for training. The *modeling* block uses these extracted features for building the speaker models. The *testing* phase includes *feature extraction*, *comparison* and *decision*. In this phase, the *feature extraction* block uses the similar procedure, as in the case of training phase, to extract features from the test speech signal. Based on the modeling approach and mode of task, a decision is made by comparing test speaker features with reference models in *comparison* and *decision* blocks. Thus, the main functions in SR system may be listed as *feature extraction*, *modeling* and *testing*.

The objective of the feature extraction stage is to extract sufficient and robust speaker information at reduced data rate for effective modeling and later for comparison [32]. The features ultimately determine the separability of the speakers and the recognition performance

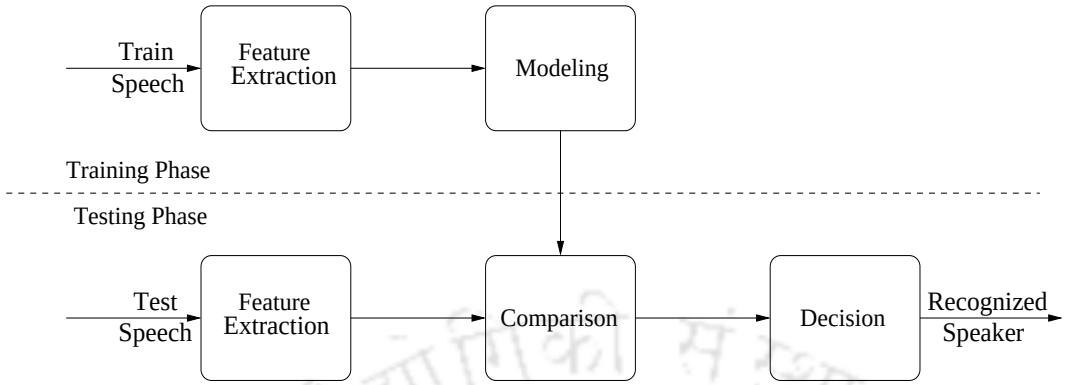

Figure 1.2: Basic block diagram of automatic speaker recognition system.

mostly depends upon the discriminating ability of the features [7, 33]. Moreover, the feature extraction stage is common for both training and testing phases. Thus, the feature extraction stage is an important stage in the SR system [31]. Selection of features having capability of effectively representing the speaker information, its robustness against unfavorable condition(s) and their accurate measurement play an important role for speaker recognition [7].

The features produced from an individual speaker are represented by vectors in the feature space. A good feature should be less variant within speakers and more variant across speakers [2, 23, 31]. But in the feature space, features of different speakers are shared and overlapped with each other. So, a second level of compression among the features of a speaker is made in the modeling stage. A large set of feature vectors of a speaker is grouped into its representative vector by several modeling techniques. The mostly used modeling techniques include, vector quantization (VQ), AANN, Gaussian mixture models (GMM) and GMM-universal back ground model (GMM-UBM) [4, 5, 27, 34–37]. State-of-the-art speaker recognition system mostly use GMM-UBM modeling technique.

The comparison is made at the frame levels and a score is assigned to the reference models. The assigned score depends upon the modeling techniques. For example, Euclidean distance and log likelihood ratio (LLR) are used as the scores for VQ and GMM-UBM modeling techniques, respectively [38, 39]. The frame level scores are accumulated and normalized over the whole utterance of the test speaker. The decision is taken based on the scores assigned to the reference models. Decision involves identifying the unknown speaker in case of identification

and accepting or rejecting the identity claim of the speaker for verification. The model having the best score gives the identity of the test speaker. The identification performance is measured as the ratio of number of correctly identified examples to the total number of examples considered for testing and is expressed in percentage. In case of verification, a threshold is set to accept or reject the claimed identity of the test speaker. In this case, there are two sources of error. The first one is called as the false accept (FA) and occurs when the claimed identity is accepted but in fact is not the claimed speaker. The second is called as the false reject (FR) and occurs when the actual speaker's claimed identity is rejected. The performance of the speaker verification system is evaluated in terms of equal error rate (EER) [40, 41]. EER is defined as the error rate at which false acceptance rate (FAR) is equal to false rejection rate (FRR) [41]. In order to improve the visualization of the speaker verification performance, the detection error tradeoff (DET) curve is also used, where false alarm and miss alarm probabilities are plotted based on the LLR scores assigned to target and imposter models [40, 41]. The false alarm and miss alarm probabilities correspond to FAR and FRR, respectively. As mentioned earlier, the excitation source signal contains speaker-specific evidence at multiple levels. It is expected that by combining evidences from multiple levels, the SR system based on excitation information may be more effective in avoiding the errors in the recognition tasks.

The function of the SR system at different stages shows that the success rate mostly depends upon extracting and then modeling the speaker-dependant characteristics of the speech signal. In this work we assume that the established model like GMM-UBM is general enough for building the speaker models and focus our work on the *feature extraction* stage to develop method for extracting speaker-specific excitation information.

1.5 Organization of the Thesis

The contents of the thesis are organized as follows:

In **chapter 2**, a review of the existing attempts in exploring the excitation information at the subsegmental, segmental and suprasegmental levels for speaker recognition is given. This chapter then compares all these methods. Finally the issues to be addressed as part of this

thesis work are identified and the organization of the work is outlined.

Chapter 3 discusses the subsegmental, segmental and suprasegmental levels processing the LP residual in the time domain for speaker information. Extracting the speaker-specific information using the analytic signal representation of the LP residual is also presented. Motivation for combining the evidences from all these levels are discussed and a method is proposed for the extraction of improved excitation information and used for speaker recognition studies.

Chapter 4 describes a simple and approximate method for the computation of LF model parameters from the LP residual. A method for parameterizing the subsegmental level excitation information by LF parameters is proposed for speaker recognition. A comparative study is made between the proposed approach and the corresponding temporal domain processing of the LP residual at the subsegmental level.

In **Chapter 5** the LP residual is processed in the frequency and cepstral domains for compact modelling of the segmental level excitation information for speaker recognition. The different nature of the speaker information present in the frequency and cepstral domain features are studied and a combined approach is proposed for parameterizing the segmental level excitation information. A comparative study is made between the proposed approach and the corresponding temporal processing of the LP residual at the segmental level.

In **Chapter 6** describes methods for parameterizing the suprasegmental level pitch, epoch strength and cepstral trajectory information from the LP residual. The different nature of suprasegmental pitch, epoch strength and cepstral trajectory information are studied and a combined approach is proposed for parameterizing the suprasegmental excitation information. A comparative study is made with the corresponding temporal domain processing of the LP residual at the suprasegmental level.

In **Chapter 7** a speaker verification system based on excitation information is developed. The proposed techniques are used to model the subsegmental, segmental and suprasegmental excitation information. Evidences from each of these levels are combined to represent the improved excitation information. The effectiveness of the proposed system is demonstrated by the speaker verification study on clean and noisy cases. A comparative study is made between

the proposed system and state-of-the-art vocal tract system.

A summary of the work presented in this thesis is given in **Chapter 8** by listing major contributions of the present work. Some directions for further research in the area of speaker recognition using excitation information are also mentioned.

2

Speaker Information from Excitation Source - A review

Contents

2.1 Introduction

To develop techniques for modeling the speaker-specific excitation information, we need to understand the various methods developed so far and their state-of-the-art performance for speaker recognition. This chapter provides a review of some of the existing approaches for the extraction of speaker-specific excitation information and highlights the issues involved. In particular, we review the methods employed in extracting the speaker-specific excitation information mostly based on processing the LP residual. The chapter is organized as follows: Section 2.2 describes the different excitations for speech production. It also describes the speaker characteristics reflected in each of the excitations. Section 2.3 descries the approach for exploiting the speaker information present in the pitch contours, jitter and shimmer. Section 2.4 describes the use of glottal flow derivative parameters for speaker recognition. LP model of excitation and its use for extracting speaker information are described in Section 2.5. A comparative study of these different approaches are given in Section 2.6. A summary and scope for the present work is given in Section 2.7. This chapter is ended with the organization of the present work.

2.2 Excitation Source Information

In speech production, the constricted airflow forms the excitation and is due to the constriction somewhere along the length of the vocal tract. The constriction can be either total or partial and periodic or aperiodic. The aperiodic, total and impulsive nature constriction is termed as stop or plosive excitation [42–44]. The plosive excitation is used as the excitation signal for producing stop consonant sounds. The aperiodic, partial and random noise nature constriction is termed as fricative excitation [1]. The fricative excitation is used for the producing fricative consonant sounds. Both stop and fricative sounds produced only using stop and fricative excitations, respectively are termed as unvoiced sounds. The total and near periodic constriction at the glottis is termed as glottal vibration [43]. This is due to the vibration of vocal folds present in the glottis. The glottal vibration is used for the production of what are

called voiced sounds that include vowels and some consonants. Depending on the mode of constriction, speech sounds can be classified as voiced or unvoiced [1, 21, 45]. In voiced sound, the constriction takes place at the glottis. The pressure in the air accumulated in the trachea (sub-glottal system) forces the vocal folds to vibrate. This vibration of the vocal folds is nearly periodic, thereby producing quasi-periodic pulses of air flow.

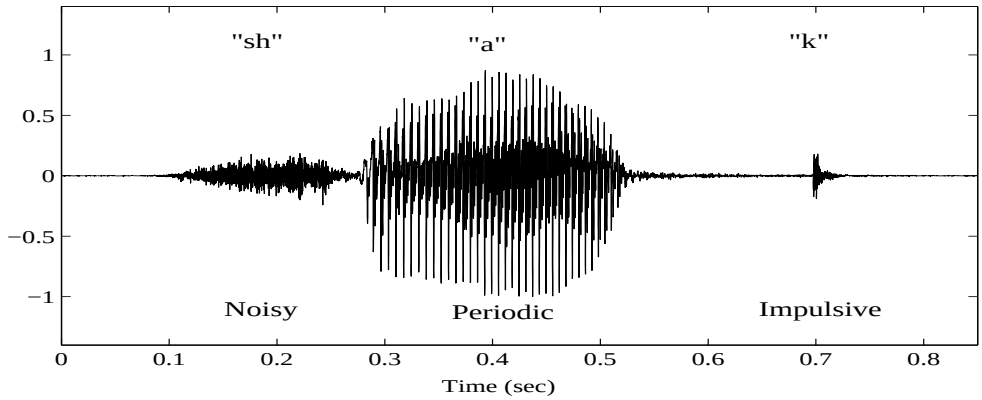

Figure 2.1: Different excitation sources for speech production. There are three excitations, namely, vibration of vocal folds, burst or stop and fricative. Vibration of vocal folds results in periodic or voiced speech (ex. /*a*/), stop excitation results in stop consonants (ex. /*k*/) and fricative results in fricative consonants (ex. /*sh*/), whose waveforms are given in the figure.

Figure 2.1 shows an example speech waveform for the word *shock*. This word has a fricative sound /*sh*/, vowel /*a*/ and plosive sound /*k*/. The nature of excitation may be observed indirectly from this waveform. Speaker characteristics present in speech may be attributed to the strength of excitation indicated by localized energy, duration of excitation signal, aperiodic or periodic and sequence of samples. Since the unvoiced excitations are random noise or impulse like sequences of short duration, they may not provide enough speaker-specific information for modeling and testing. Among the three different excitations present, the major one is the vibration of vocal folds, since majority (about 70%) of speech produced is of voiced type [21]. Most of the speaker-specific excitation information is therefore viewed to be present in the voiced excitation. Existing attempts in developing the excitation based speaker recognition systems are by exploiting information due to the vibration of the vocal folds. The rest of the

chapter discusses methods that exploited the information associated with the vibration of the vocal folds for speaker recognition.

2.3 Speaker Information from Pitch and its Variants

2.3.1 Pitch Contours

Vibration of the vocal folds is nearly periodic and is measured using pitch. Average rate of vocal folds vibration measured in the frequency domain is defined as pitch. The rate of vibration is inversely proportional to the shape and size of the vocal folds. For example, the size of the vocal folds in men is larger than that in women and accordingly pitch of men is lower than that of women [21]. The size of vocal folds is different from speaker to speaker and hence the pitch contains unique information of a speaker. Pitch varies from its average value while speaking, in response to the change in air pressure and vocal folds tension. The dynamics of the vocal folds vibration is also distinct for a speaker. Thus, the change in pitch with respect to time, called as the pitch contour also contains unique information about of a speaker. Since pitch contour contains pitch and its dynamics, it has more speaker information compared to pitch alone.

A study on speaker recognition by exploiting the information from pitch contours was done by Atal [8]. In this work the pitch was computed using autocorrelation analysis. It was reported that the durations of the pitch contours were different from one utterance to another, even for the same speaker. To take care of this, the duration of the speech signals were normalized to the same time interval and using statistical pattern recognition procedure, namely, linear discriminant analysis, speaker recognition study was conducted. In a population of 10 speakers, 97% identification accuracy was achieved. Although, the population was very small, but the number of tests were around 60. As per the available literature, this is the first attempt of using only excitation information for speaker recognition.

2.3.2 Speaker Information from Jitter and Shimmer

During speech production, due to variation in subglottal pressure, the tension and mass lesions in the vocal folds change continuously. As a result, the rate and amplitude of the vocal folds vibration change from cycle to cycle. These differences are significant across different speaking styles [46]. Since the speaking style is different across different speakers, the local variation in the rate and amplitude of the vocal folds vibration may help in quantifying some high level speaker information like speaking style.

In [47], the local variation in rate and amplitude of the vocal folds vibration is used for the speaker recognition task. The variation in the rate and amplitude of the vocal folds vibration is very less around 2-3% of their average value. Therefore the speaker information is extracted using statistical measurements from several glottal cycles. To capture information about the local variation in rate and amplitude of the vocal folds vibration, *Jitter* and *Shimmer* parameters were used. The deviation in the periodicity of vocal folds vibration between pitch periods is measured as *Jitter* [47]. It is measured as the average pitch deviation over successive pitch periods. Usually jitter is measured from three to five successive pitch periods. Mathematically, jitter is defined as,

$$J_T = \frac{\frac{1}{2}[V_{nj} + V_{(n+1)j}]}{\frac{1}{3}[P_{n-1} + P_n + P_{n+1}]} \tag{2.1}$$

where P_n is the pitch period and n is the index of the current frame and V_{nj} is the deviation in the jitteer measurement for the n^{th} frame and is defined as

$$V_{nj} = |P_n - P_{n-1}| \tag{2.2}$$

Similar to intonation, the temporal variation of jitter is also significant for a speaker. Temporal variation of jitter of two different speakers for the text *She had your dark suit in greasy wash water all year* is shown in the Figure 2.2. The temporal variation of jitter is different across different speakers indicating the presence of speaker information in the jitter contour. A speaker recognition study was conducted by extracting information from the jitter [47]. In this study some features derived from the jitter were used as information for speaker recognition. The

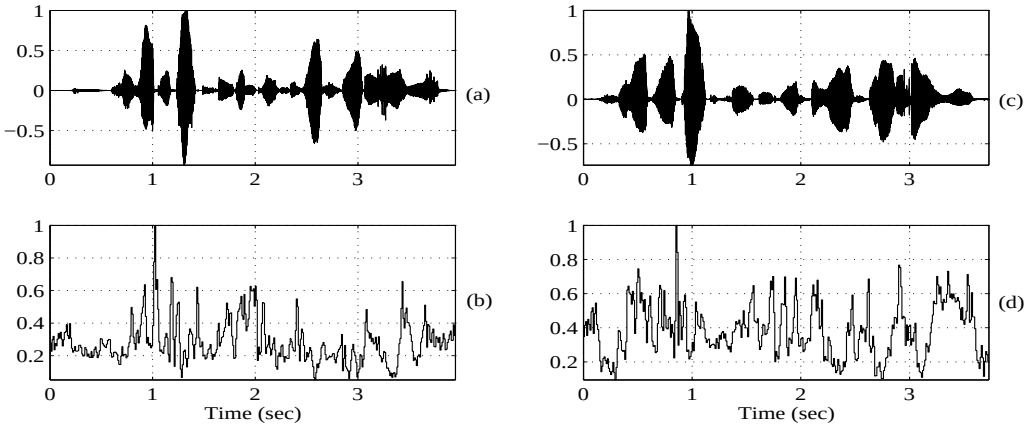

Figure 2.2: Speech signal and temporal variation of *jitter* (normalized) for a speech of text *"She had your dark suit in greasy wash water all year"*. Speech data and corresponding *jitter* contour of speaker 1 in (a) and (b), and speaker 2 in (c) and (d). The *jitter* contours are different from speaker to speaker indicating speaker uniqueness present in them.

derived features are

$$J_T(absolute) = \frac{1}{N_j - 1} \sum_{i=1}^{N_j - 1} |T_i - T_{i+1}| \tag{2.3}$$

$$J_T(average) = \frac{\frac{1}{N_j - 1} \sum_{i=1}^{N-1} |T_i - T_{i+1}|}{< T >} \tag{2.4}$$

$$J_T(relative) = \frac{\frac{1}{N_j - 1} \sum_{i=1}^{N_j - 2} |T_i - T_{i+1} - T_{i+2} - < T >|}{< T >} \tag{2.5}$$

$$J_T(ppq5) = \frac{\frac{1}{N_j - 1} \sum_{i=1}^{N_j - 4} |T_i - T_{i+1} - T_{i+2} - T_{i+3} - T_{i+4} - < T >|}{< T >} \tag{2.6}$$

where $< T >$ is the average time over N_j extracted pitch periods and is equal to

$$< T > = \frac{1}{N_j} \sum_{i=1}^{N_j} T_i \tag{2.7}$$

Speaker recognition was performed on NIST-2001 database [48]. The experimental results obtained are tabulated in the Table 2.1. The following observations were put forward from this study: Absolute jitter measurement is useful for speaker recognition. Fusion of all Jitter measurements does not perform well than individual jitter parameters. Relative measurement is not providing useful information. Fusion of jitter (relative) does not improve the performance (29.3%). Three cycle-to-cycle variation measurement of jitter is more suitable for speaker

Table 2.1: Speaker verification results for *jitter* absolute, *jitter* relative, *jitter* relative average perturbation (rap) and *jitter* five point period perturbation quotient (ppq5) measurements. The verification result in fusion of all these measurements is given in the last row. Results are expressed in terms of EER. The highest performance is obtained from *jitter* absolute measurement. The fusion of information does not improve the verification performance.

Jitter Measurement	$EER(\%)$
J_T(absolute)	26.9
J_T(relative)	36.7
J_T(rap)	34.2
J_T(ppq5)	33.8
Fusion	29.2

recognition.

2.3.3 Speaker Information from Shimmer

Jitter gives information about local variation in the pitch periodicity values. In a similar way, the peak amplitudes of the excitation signal in each pitch period also shows a local variation. Similar to jitter, speaker information is extracted using statistical measurements from several pitch periods. One such measurement is $Shimmer(S_H)$. Shimmer is defined as the deviation in the peak amplitudes of the excitation signal between pitch periods. It is measured as the average peak amplitude deviation over successive pitch periods [47]. Usually shimmer is measured from three to five successive pitch periods. The variation in peak amplitude from period to period is shown in the Figure 2.3. The peak amplitudes over the selected three successive pitch periods vary over 0.19 to 0.21 on the normalized scale. These differences are also significant across different speaking styles in particular, loudness [46]. Since loudness is different across different speakers, the local variation in the peak amplitudes may help in quantifying the high level speaker information like loudness. The temporal variation of shimmer is also distinct for a speaker as in the case of jitter [46]. Temporal variation of shimmer of two different speakers for the text *She had your dark suit in greasy wash water all year* is shown in the Figure 2.4.

The shimmer contours are different across different speakers indicating the presence of speaker information. Mathematically, shimmer is defined as,

$$S_H = \frac{\frac{1}{2}[V_{ns} + V_{(n+1)s}]}{\frac{1}{3}[A_{n-1} + A_n + A_{n+1}]} \tag{2.8}$$

where A_n is the peak amplitude and n is the index of the current frame and V_{ns} is the deviation defined as

$$V_{ns} = |A_n - A_{n-1}| \tag{2.9}$$

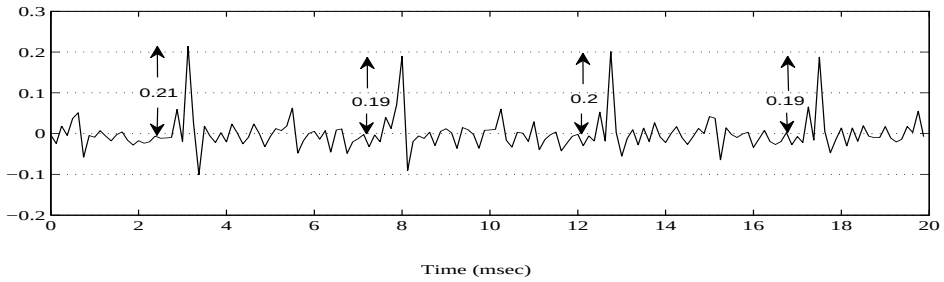

Figure 2.3: Variation of peak amplitude from one pitch period to other in a segment (20 ms) of speech. In this example peak amplitude at one pitch period is 0.21 and changed to 0.19 in the next pitch period and changed to 0.2 in the next pitch period.

A speaker recognition study was conducted by extracting information from the shimmer [47]. In this study some features derived from shimmer were used as information for speaker recognition. The features derived are

$$S_H(absolute) = \frac{1}{N_s - 1} \sum_{i=1}^{N_s-1} |log(A_i - A_{i+1})| \tag{2.10}$$

$$S_H(average) = \frac{\frac{1}{N_s-1} \sum_{i=1}^{N_s-1} |A_i - A_{i+1}|}{< A >} \tag{2.11}$$

$$S_H(apq3) = \frac{\frac{1}{N_s-1} \sum_{i=1}^{N_s-2} |A_i - A_{i+1} - A_{i+2} - < A >|}{< A >} \tag{2.12}$$

$$S_H(apq5) = \frac{\frac{1}{N_s-1} \sum_{i=1}^{N_s-4} |A_i - A_{i+1} - A_{i+2} - A_{i+3} - A_{i+4} - < A >|}{< A >} \tag{2.13}$$

$$S_H(apq11) = \frac{\frac{1}{N_s-1} \sum_{i=1}^{N_s-4} |A_i - \sum_{j=1}^{j=10} A_{i+j} - < A >|}{< A >} \tag{2.14}$$

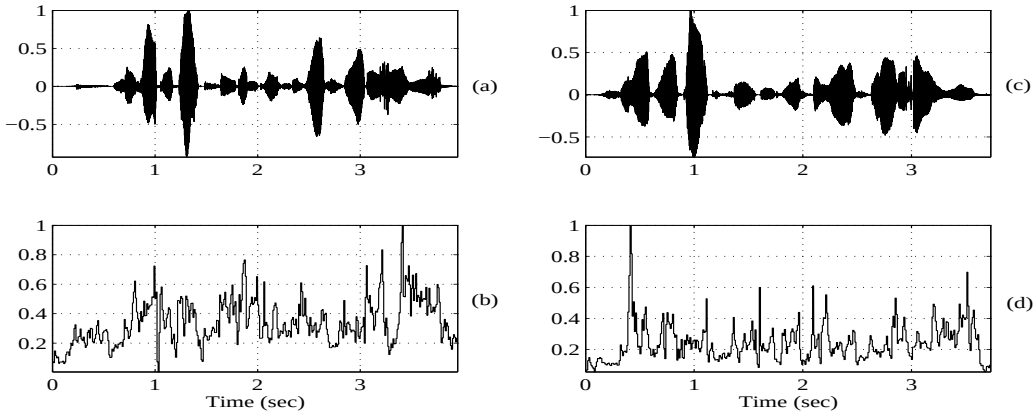

Figure 2.4: Speech signal and temporal variation of *shimmer* (normalized) for a speech of text *"She had your dark suit in greasy wash water all year"*. Speech data and corresponding *shimmer* contour of speaker 1 in (a) and (b), and speaker 2 in (c) and (d). The *shimmer* contours are different from speaker to speaker indicating speaker uniqueness present in them.

where $< A >$ is the average peak amplitude over N_s extracted pitch periods and is equal to

$$< A > = \frac{1}{N_s} \sum_{i=1}^{N_s} A_i \tag{2.15}$$

Speaker recognition was performed using NIST-2001 database [48]. The experimental results obtained for the above parameters are given in the Table 2.2. It was observed that absolute measurement of shimmer is useful for speaker recognition. Relative measurement is providing useful information. Fusion of shimmer (apq3) also provides improved performance (25.5%).

Comparing the performance of jitter and shimmer measurements, in both cases, the absolute measurement seem to provide best performance. Shimmer is more effective (25.5%) than jitter (29.2%) in the fusion of features. In this study the complementary nature of jitter and shimmer was also exploited. It was observed that with the fusion of best jitter (J_T(absolute)) and shimmer (S_H(fusion)) features, a new feature called as *JitShim* improved the performance from 25.5% to 22.1%. The complementary nature of jitter and shimmer were also verified with the conventional spectral (MFCC) and prosodic (pitch contour and duration) features. By combining *JitShim* with MFCC and prosodic features the relative improvements achieved were 15% and 17%, respectively. From this observation, they reported that *JitShim* is more complementary to prosodic features than spectral features.

Table 2.2: Speaker verification results for *shimmer* absolute, *shimmer* relative, *shimmer* three point amplitude perturbation quotient (apq3), *shimmer* five point amplitude perturbation quotient (apq5) and *shimmer* eleven point amplitude perturbation quotient (apq11) measurements. The verification result in fusion of all these measurements is given in the last row. Results are expressed in terms of EER. The highest performance is obtained from *shimmer* absolute measurement. The fusion of information improves the verification performance.

Shimmer Measurement	$EER(\%)$
S_H(absolute)	26.9
S_H(relative)	28.9
S_H(apq3)	28.1
S_H(apq5)	32.9
S_H(apq11)	33.8
S_H(Fusion)	25.5

2.4 Speaker Information from Glottal Flow

The modulated air flow due to the vibration of vocal folds is termed as glottal flow [15, 49]. The manner, speed and change in the rate of vocal folds vibration also changes from speaker to speaker [50]. In some speakers, vocal folds never close completely (soft voice) and in other cases vocal folds close completely and rapidly (hard voice).ation Similarly, duration of opening and closing of vocal folds, the instants of glottal opening and closing and the shape of the glottal flow also vary significantly from speaker to speaker. These differences correspond to the differences in glottis and are reflected in the glottal flow. Hence the glottal flow wave also contains speaker information. The way of characterizing the glottal flow is to measure the volume velocity of air flow through glottis. From the volume/pressure relationship, its derivative represents glottal air pressure flow. In general it is difficult to obtain precise measurement of glottal pressure waveform [21]. One method of modeling the glottal flow is based on direct waveform model. In this technique, a particular shape such as half wave rectified sine pulse is adapted and parameterized from the inverse filtering of the given speech signal. One such approximated model available in literature is Liljencrants-Fant (LF) model [51]. The LF model glottal flow

derivative for one cycle of vocal folds vibration is shown in the Figure 2.5. As marked in the figure, in the LF model, the glottal flow is parameterized by seven LF parameters, defined as

T_o = Time of glottal opening

T_e = Time of maximum negative value of the glottal pulse.

E_e = Value of glottal flow derivative at T_e

α = Ratio of E_e to peak height of positive portion of glottal flow derivative

β = Time constant that determines how quickly glottal flow derivative returns to zero.

T_c = Time of glottal closure.

w_o = Reciprocal of the time that elapses between zero crossing and T_e .

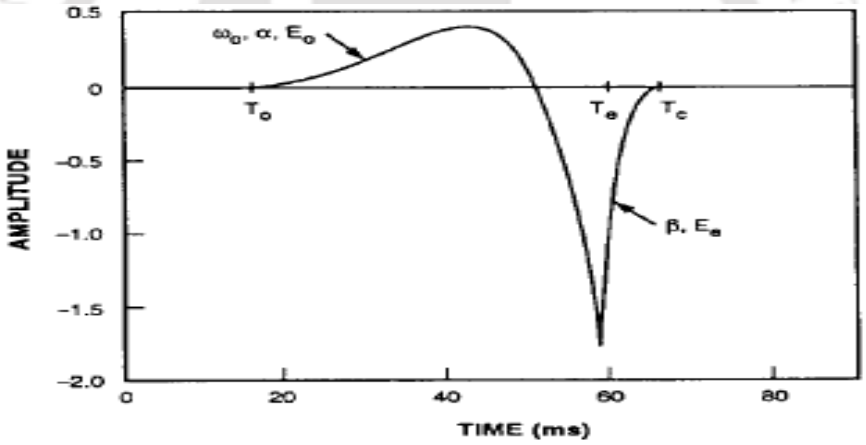

Figure 2.5: LF model of glottal flow derivative waveform for one cycle of vocal folds vibration [15]. The glottal flow derivative is modeled by Seven LF parameters. Three of the parameters (E_0, w_o, and α) describe the shape of the glottal flow during nonzero flow (T_0-T_e). The two parameters (E_e, β) describe the shape of the glottal flow during most negative glottal flow derivative and glottal closure (T_e-T_c). The other two parameters (T_e and E_e) describe the time and amplitude of the most negative peak of the glottal flow derivative.

The LF parameters are derived from the inverse filtering of the speech signal and hence they represent the individual characteristics of the glottal flow of the speaker. The glottal flow and its derivative for a speech utterance is shown in the Figure 2.6. The figure also includes the speech waveform. By comparing all the three plots, it can be observed that the large errors present in the glottal flow derivative correspond to the large error locations in the speech waveform. These large error regions in turn correspond to the regions of instants of glottal closure. Thus it is relatively easy to derive a signal from speech, proportional to the glottal

flow derivative. Once we have an approximate derivative signal, by suitable integration the glottal flow waveform can be obtained. The LF measurements can then be made from the glottal flow waveform and used for speaker recognition.

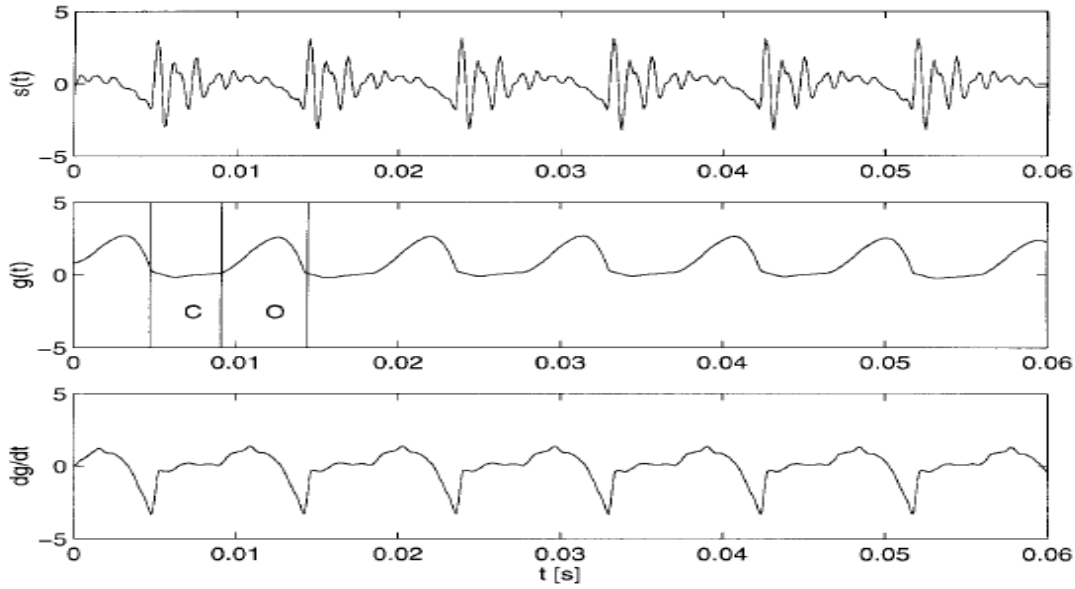

Figure 2.6: Speech and glottal waveforms [52]. Top: speech waveform s(t), Middle: glottal flow waveform g(t), Bottom: glottal flow derivative waveform $\frac{dg}{dt}$. The closed phase (C) and open phase (O) in glottal flow are indicated in glottal flow waveform. The large peak present in glottal flow derivative correspond large peak in the speech signal.

Based on this information a speaker recognition study was conducted on TIMIT and NTIMIT databases by [15]. Seven LF parameters were used as components of feature vectors and Gaussian Mixture Model (GMM) technique was used to build the reference models. An identification accuracy of 60% was achieved in case of TIMIT database and 55% for NTIMIT database. Also, MFCC features of the glottal derivative waveforms obtained from the seven parameters are used for the speaker recognition. The performance significantly improved to 95% in case of TIMIT. It was concluded that MFCC of glottal flow derivative is more compatible to GMM than its feature vector representation.

2.5 Speaker Information from LP Residual

The LP model of excitation is the residual of the speech obtained after suppressing the vocal tract characteristics [12]. In LP model of speech production, each sample of speech is predicted as a linear combination of the past p samples, where p is the order of prediction. The vocal tract and the glottal excitation are represented by

$$\hat{s}(n) = -\sum_{k=1}^{p} a_k s(n-k) \tag{2.16}$$

Since the prediction coefficients mostly represent the vocal tract information, then suppression of this information from the speech signal results in an error or residual signal which grossly approximates the glottal excitation [24]. Therefore the source is approximated by the error signal in the prediction and obtained as [24].

$$r\,(n) = s\,(n) + \sum_{k=1}^{p} a_k s\,(n-k) \tag{2.17}$$

The representation of source information in the LP residual depends upon the order of prediction. It was shown in the literature that the LP order of 10-14 for speech sampled at 8 kHz is appropriate for the proper representation of excitation information [12]. The LP spectrum and LP residual spectrum for different orders of prediction of a voiced speech segment are shown in the Figure 2.7. In the lower order of prediction, LP spectrum contains only the prominent peaks. This means that the residual still contains information about the vocal tract. Therefore the residual spectrum is modulated by formants as shown in the Figure 2.7(b). On the other hand if the prediction order is increased to a very large value, the LP spectrum contains some spurious peaks as shown in Figure 2.7(e). Accordingly, the speech passed through the corresponding inverse filter will affect the residual by attenuating some of the peaks corresponding to pitch and its harmonics as shown in the Figure 2.7(f). With proper choice of LP order, LP spectrum contains mostly vocal tract information as shown in Figure 2.7(c). Formant effects are eliminated and the residual spectrum represents excitation information properly as shown in the Figure 2.7(d). The flat spectrum with certain periodicity implies, LP residual is a noise

like sequence with impulse at regular interval.

A segment of voiced speech and its corresponding LP residual obtained from 10^{th} order LP analysis are shown in the Figure 2.8. The occurrence of peaks in the residual waveform represents the strength and periodicity of the glottal vibrations [24]. As the nature of glottal vibrations are distinct for a given speaker, the LP residuals for different speakers are different [15]. These differences contribute to speaker information. The LP residual for four different speakers (2 Males and 2 Females) for the same segment of speech (40 ms) is shown in the Figure 2.9. It can be seen from the figure that the strength and pattern of the LP residual are different across speakers. These changes in the characteristics of LP residual show the presence of speaker information in them. It is also mentioned in the literature that, humans can recognize speakers by listening to the LP residual [53]. Some attempts have been made for extracting speaker information from the residual such as temporal sequence of residual samples, instantaneous phase of the LP residual derived from the analytic signal representation, cepstral analysis of LP residual and the spectral flatness measurements of LP residual [9–13, 17].

2.5.0.1 Speaker Recognition using LP Residual Samples

A speaker recognition study was conducted for two sets of 20 speakers from NIST-99 database using LP residual samples as the speaker information [11, 54]. In this study small segments of residual samples are used as feature vectors. Blocks of 40 samples with a shift of one sample of residual are used as feature vectors for building the speaker models. The feature vectors are modeled with auto-associative neural network (AANN) [27]. An identification accuracy of 77.5% was achieved. It was reported that, the individual performance of the vocal tract feature is 90%. When the source information is combined with the vocal tract information, the performance improved to 95%. Improvement in the performance due to combination shows the different nature of speaker information present in the LP residual samples.

2.5.0.2 Speaker Recognition using LP Residual Phase

The changes in the sequence of samples of the LP residual are also distinct for a speaker. These changes also contain speaker information. But due to large fluctuations in the amplitudes

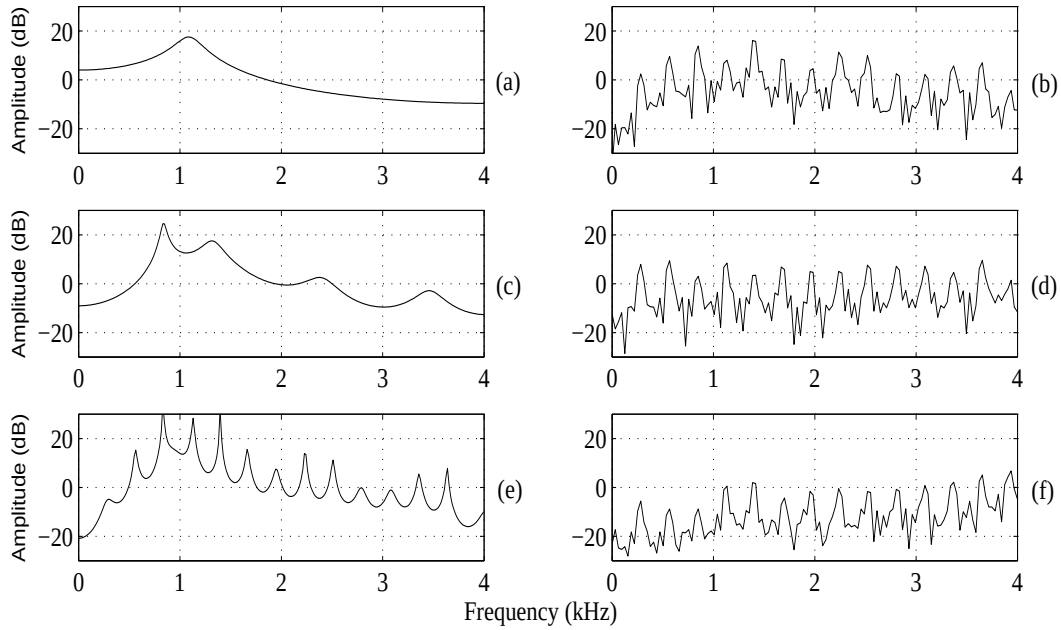

Figure 2.7: LP and LP residual spectra [12]. (a) and (b) LP and LP residual spectrum for order 2. (c) and (d) LP and LP residual spectrum for order 10. (e) and (f) LP and LP residual spectrum for order 30. In lower order of prediction, LP spectrum contain some prominent peaks only, so some formant information still remains in the residual spectrum. In higher order of prediction, LP spectrum contains some spurious peaks, so the corresponding inverse filter affect the residual by attenuating the peaks related to pitch and its harmonics. Proper choice of LP order i.e. 10^{th} LP analysis, eliminate the formant information from the residual spectrum and contain mostly the excitation information.

of the LP residual, it is difficult to extract speaker information present in the sequence from the residual [13]. An alternative way of extracting the sequential information is to obtain the instantaneous phase. Speech is a real signal and hence the instantaneous phase of the LP residual can be obtained from the analytic signal method [55, 56]. The analytic signal of the LP residual $r(n)$ is given by

$$r_a(n) = r(n) + jr_h(n) \tag{2.18}$$

where $r_h(n)$ is the Hilbert transform of $r(n)$ and is given by

$$r_h(n) = IFT[R_h(\omega)] \tag{2.19}$$

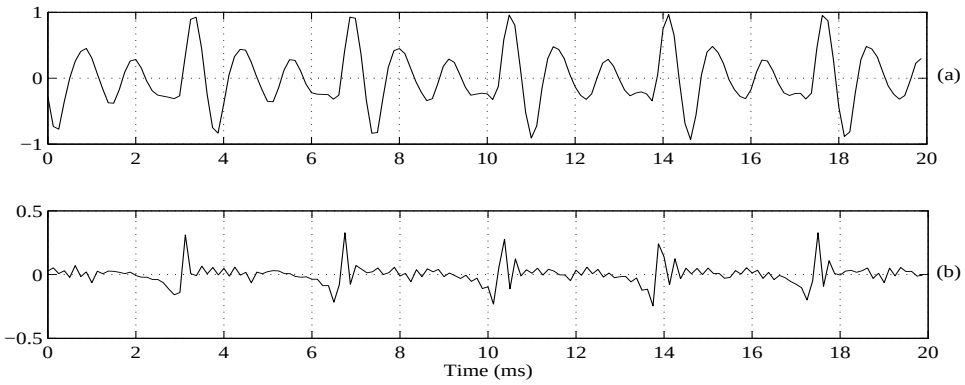

Figure 2.8: Speech and LP residual. (a) Voiced segment of speech (b) Corresponding LP residual obtained from 10 order LP analysis. The occurrence of peaks in the residual waveform represent the strength and periodicity of the glottal vibrations. The sharp negative peaks around the peaks of the residual represent the instants of vocal folds closing.

where

$$R_h(w) = \begin{cases} -jR(w), 0 \leq w < \pi \\ jR(w), 0 > w \geq \pi \end{cases} \tag{2.20}$$

$R(\omega)$ is the Fourier transform of $r(n)$ and IFT denotes the inverse Fourier transform. Then the cosine of the phase of the analytic signal can be obtained as the ratio of real part of the analytic signal and its magnitude. Mathematically,

$$cos(\theta(n)) = \frac{Re(r_a(n))}{|r_a(n)|} \tag{2.21}$$

The steps in the computation of LP residual phase of a voiced segment of speech are shown in the Figure 2.10. In LP residual phase the strength of the excitation present in the LP residual are eliminated. Although the residual phase is a noise like sequence, values of LP residual phase around the instants of glottal closure for voiced speech are significant [13, 57]. The phase information around the glottal closure instants may contain speaker information and can be used for speaker recognition. LP residual and residual phase of a segment of voiced speech signal for two different speakers are shown in the Figure 2.11. The changes in phase around the glottal closure instants are different from one speaker to another. Using these differences as speaker information, speaker recognition study was conducted on NIST-2003 database [58].

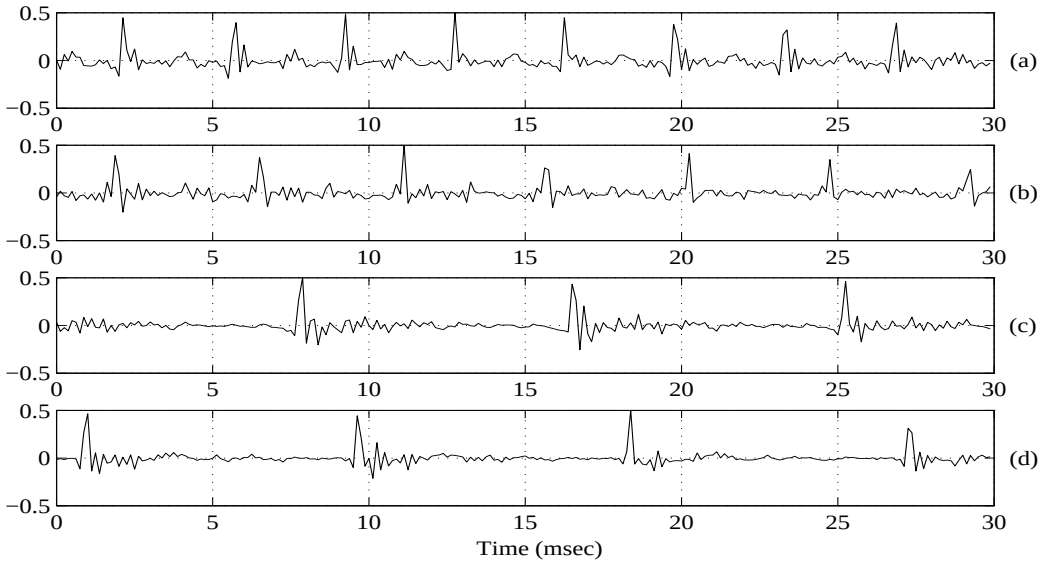

Figure 2.9: LP residual excitation signals of four different speakers for the same speech segment (40 ms). (a), (b) Female speakers and (c), (d) Male speakers. Strength and pattern of LP residuals are different for different speakers indicating speaker uniqueness present in them.

Instantaneous phase was obtained from the Hilbert transform of the LP residual. The EER based on the residual phase was achieved to be 22%. The EER based on MFCC features was 14%. By combining the evidences from both LP residual phase and MFCC features, the EER reduced to 10.5%. Improvements in the performance due to the fusion of excitation information from the residual phase shows the different nature of the speaker information present in the LP residual phase.

2.5.1 Speaker Recognition using Cepstral Analysis of LP Residual

Apart from pitch and phase information, excitation also contains other kinds of speaker information in the residual. Since this information is still less understood, a method was proposed to extract speaker information from the residual after eliminating the pitch and phase information, through cepstral analysis [9, 59]. In this technique, the inverse log magnitude spectrum eliminates the pitch and phase information. The objective was to know the existence of the speaker information in the magnitude spectrum of the residual and its relevance with the cepstral analysis of speech signal that represents the vocal tract information. From the LP

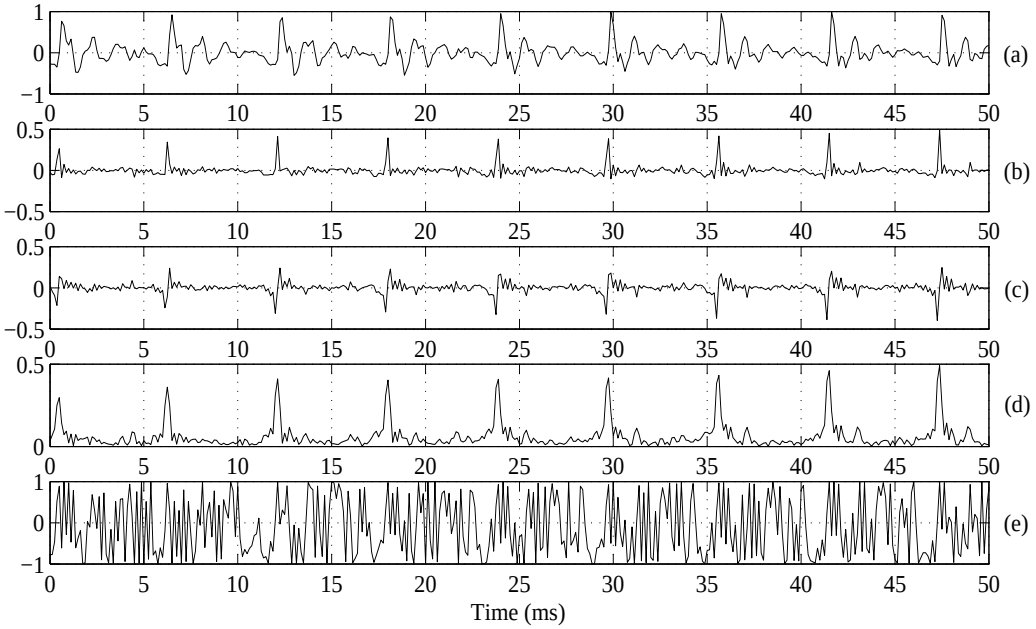

Figure 2.10: Steps in computation of residual phase [13]. (a) Speech signal. (b) LP residual. (c) Hilbert transform of residual. (d) Hilbert envelope of residual. (e) Residual phase. Even though, the residual phase plot looks like a noise, the sequence of phase samples may be unique for each speaker and hence may contain speaker information.

analysis of speech signal, the residual is given by

$$r\left(n\right) = s\left(n\right) + \sum_{k=1}^{p} a_k s\left(n-k\right) \tag{2.22}$$

The magnitude spectrum of the LP residual is given by

$$R\left(k\right) = \left| \sum_{n=0}^{N\_1} r\left(n\right) e^{\frac{-j2\pi nk}{N}} \right| \tag{2.23}$$

Then the residual real cepstrum is obtained from inverse Fourier transform of the log magnitude spectrum as

$$c(n) = \frac{1}{N} \sum_{k=0}^{N-1} \ln \left|R\left(k\right)\right| e^{\frac{j2\pi nk}{N}} \tag{2.24}$$

where, N represents number of points for DFT and IDFT computation.

Prominent information about speaker is present around the cepstral peak occurring at the zeroth coefficient. The authors hypothesized that, apart from peaks in the cepstrum that represents the pitch information, as a whole it carries richer information than the pitch alone. A

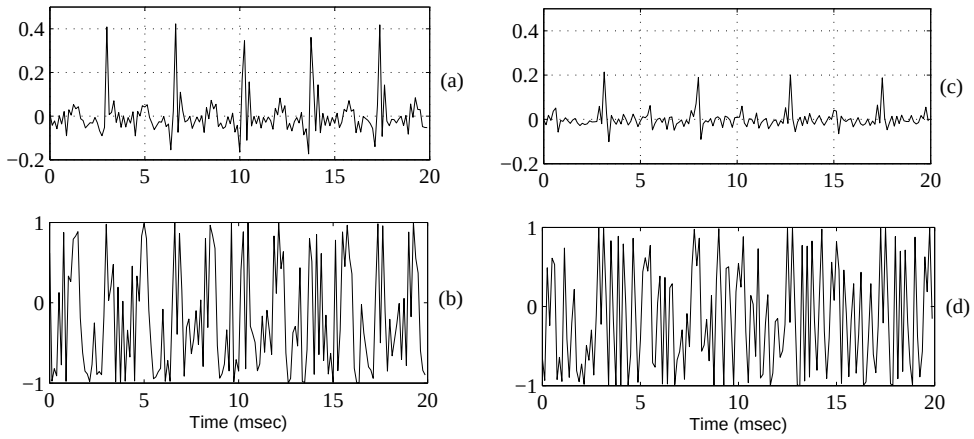

Figure 2.11: Segments of 20 ms duration of voiced excitation signals and their corresponding residual phase signals for two different speakers. (a), (b) speaker 1 and (c), (d) speakers 2. Pattern of LP residual phase signals are different for different speakers indicating speaker uniqueness present in them.

speaker verification study was conducted for 50 speakers set of balanced male and female speakers. The verification accuracy achieved was 12.9%. The verification accuracy achieved with vocal tract information was 5.7%. When both are combined the performance was improved. The combined system provided verification accuracy of 4%.

2.5.2 Speaker Recognition using Harmonic Structure of LP Residual

For proper LP order the spectrum of the LP residual is nearly flat, since the formant information is almost removed. Therefore the LP residual spectrum tends to be independent of phonetic information and it may have only speaker information related to excitation. The LP residual spectrum of four different speakers for the same text segment is shown in the Figure 2.12. It can be seen from the figure that the dynamic range of the harmonic structure of the spectrum vary from speaker to speaker. For instance, the dynamic range is about 35 dB for the speaker shown in the Figure 2.12(a) and 40 dB for the speaker shown in the Figure 2.12(c). This variation in the dynamic range may contain speaker information. The variation in the dynamic range depends on the periodicity of the spectrum. Larger the periodicity, more will be the difference between peaks and valleys of the spectrum. Since the periodicity is unique

for a speaker, it is expected that these differences may be effective for speaker recognition.

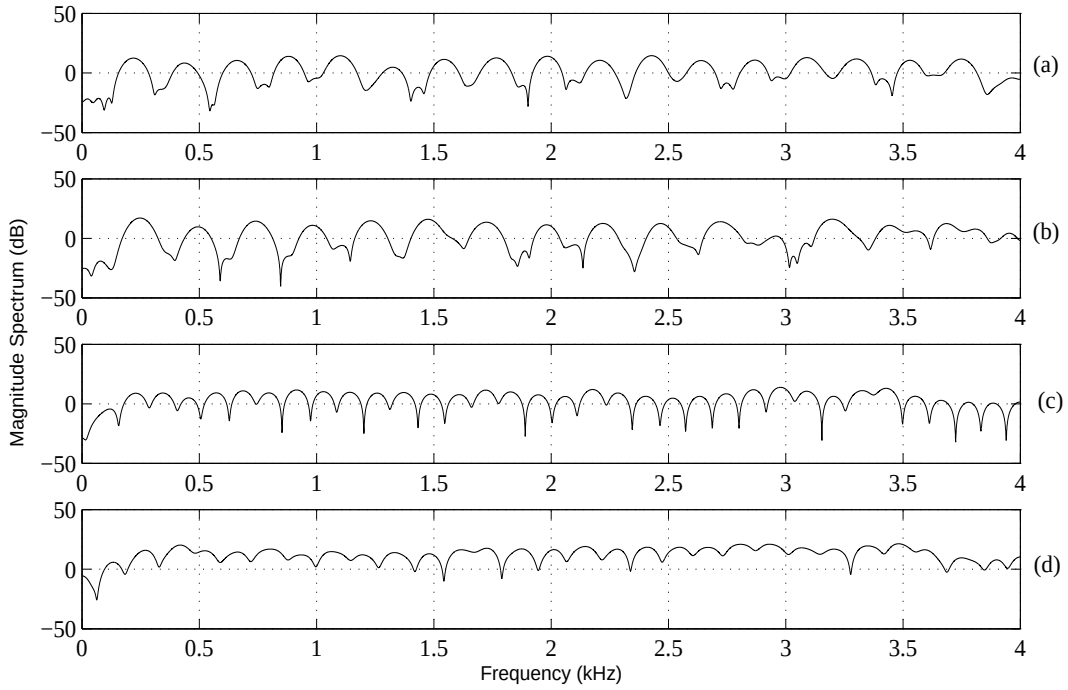

Figure 2.12: LP residual spectra. The harmonic structure and the dynamic range of the spectrum vary from speaker to speaker. (a), (b) Female speakers and (c), (d) Male speakers.

A speaker recognition study was conducted based on the spectral flatness measure that gives the power differences in the subbands [10]. In this experiment, power differences of spectrum in subbands (PDSS) are used as the speaker information. The spectral flatness in the subband is measured as the ratio of geometric to arithmetic mean of the power spectrum in the band,

$$\frac{\left[\prod_{k=L_i}^{H_i} p\left(k\right)\right]^{\frac{1}{N_i}}}{\frac{1}{N_i}\sum_{k=L_i}^{H_i} p\left(k\right)} \tag{2.25}$$

where, $p(k)$ is the power spectrum, $N_i = H_i - L_i + 1$ is the sample number of frequency points in i^{th} subband and L_i, H_i are the lower and upper limits of frequency in the i^{th} subband,

respectively. Then PDSS is computed as

$$V(i) = 1.0 - \frac{\left[\prod_{k=L_i}^{H_i} p(k)\right]^{\frac{1}{N_i}}}{\frac{1}{N_i}\sum_{k=L_i}^{H_i} p(k)}, \tag{2.26}$$

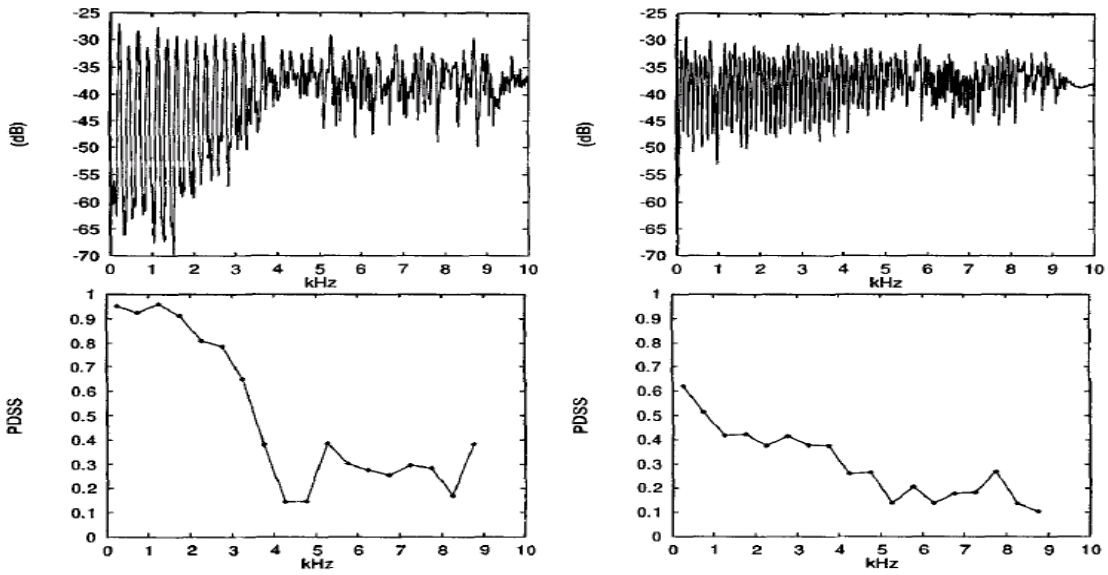

Figure 2.13: LP residual spectra and corresponding PDSS plots of male and female speakers [10]. Higher the periodicity of the LP residual spectra, the corresponding PDSS value is closer to 1 and lower the periodicity, the corresponding PDSS value is closer to 0. The nature of PDSS plots are different indicating the presence of speaker information.

If geometric mean is equal to the arithmetic mean, then the flatness is maximum and equal to 1. In that case the power differences of spectrum in subband i.e., PDSS is zero. Otherwise PDSS value exists between 0 and 1. The PDSS of two different speakers are shown in the Figure 2.13. The nature of PDSS plots are different indicating the presence of speaker information. In the male speaker case, the PDSS varies from about 0.1 to 0.9 and in case of female speaker, the PDSS variation is from 0.1 to 0.6. Thus the male speaker speech has more periodicity and flatness compared to the female speaker. This aspect may be taken as the speaker information for speaker recognition. In a population of 50 speakers, 66.99% of recognition accuracy was achieved. Individually, LPCC is giving 98% accuracy but when combined with PDSS, the performance improved to 99%. This shows the presence of additional

information in the harmonic structure of LP residual spectrum.

2.5.3 Speaker Recognition from Time Frequency Analysis of LP Residual

Although the harmonic structure of the LP spectra reflects the nature of the periodicity as shown in Figure 2.12, but the Fourier transform of the LP residual is nearly flat. A better manifestation of speaker information can be modeled using wavelet transform [29, 60]. The usefulness of wavelet is therefore explored in [29]. Wavelet transform provides multi-resolution analysis with the help of choice of basis function, scaling and translation parameters. Wavelet octave coefficients of residues (WOCOR) are used as the feature parameters to represent speaker information in [29]. These coefficients are computed from the wavelet transform of the windowed residual signal $r_(n)$. The length of the window considered was of two pitch period long around epochs extracted by robust algorithm for pitch tracking (RAPT) [61], and is given by

$$w(a,b) = \frac{1}{\sqrt{|a|}} \sum_{n} r_h(n) \Psi^*(\frac{n-a}{b}) \tag{2.27}$$

where, $a = 2^k | k = 1, 2, ..., K_w$ and $b = 1, 2, ..., N_w$, and N_w is the window length, $\Psi^*(n)$ is the conjugate of the fourth-order Daubechies wavelet basis function $\Psi(n)$ and a and b are the scaling and translation parameters, respectively [29, 60]. In this study, with $K_w = 4$, the signal is decomposed into four subbands and each subband group of coefficients are divided evenly into M_w subgroups to capture the temporal information, where $M_w = 4$ such that the m^{th} coefficient of the k^{th} subgroup is given by

$$W_k^{M_w}(m) = \{w(2^k, b) | b \in (\frac{(m-1)N_w}{M_w}, \frac{mN_w}{M_w}\}, m = 1, .., M_w \tag{2.28}$$

Since $M_w = 4$ and $K_w = 4$, each feature vector consist of 16 components. A speaker verification experiment on text independent mode was conducted on the 74 male speakers taken from NIST-2001 database [48]. GMM-UBM modeling technique is used to build the speaker models [35]. The EER achieved in the proposed method is 21.8%. Performance of the MFCC features is 9.3%. Although individual performance of the WOCOR features is poor compared

to conventional MFCC features, their combination improved EER to 7.67%. This shows the usefulness of wavelet transform for performing time frequency analysis of LP residual.

2.6 Comparison of Speaker Recognition Studies using excitation information

The earlier sections described the important studies that have been made to exploit the excitation information for speaker recognition. This section gives a comparative study of these approaches. The comparison will be made based on the following factors: Nature of task (identification/verification), mode of operation (text dependent/independent), database size, excitation feature(s), nature of speaker information present, approach for modeling, performance, important conclusions and limitations, if any.

The first study was on exploring the pitch information and demonstrated the usefulness of pitch contours for speaker recognition on a database consisting of 10 speakers. The nature of the speaker information present in pitch contours may be attributed to the suprasegmental level. The exploration on very small database and its suitability to only text-dependent speaker recognition are the limitations of this study.

The second study used different measurements of jitter and shimmer for text-independent speaker verification task on NIST 2001 database. This study extracts and models the variations occurring at the segmental level. In terms of EER, absolute measurement of jitter and fusion of all shimmer parameters provide the best performance. By combining their respective evidences (JitShim), the performance further improved. Individually JitShim gives relatively poor performnace, but provides complementary evidence to vocal tract information (MFCC). The poor performance achieved by jitter and shimmer measurements indicates that the speaker information manifested in them may be relatively less. Since, the variations are less, more accurate signal processing methods are required for jitter and shimmer measurements.

The third study used the speaker information manifested in glottal cycle characteristics for text-independent speaker identification study on TIMIT and NTIMIT databases. Since each glottal cycle will be typically less than 10 msec, this study may be treated under subsegmental

processing. The glottal flow derivative parameters were estimated using LF model and used as feature in a GMM based speaker identification system. For TIMIT database 60% and for NTIMIT database 55% accuracy were achieved [15]. In this study, the MFCCs of the seven parameter LF model glottal flow derivative were used as feature for speaker identification task on TIMIT database. The performance significantly improved to 95%. The difficulty with this approach is in the process of deriving the accurate values of the LF parameters. Further, the LF model is one model for modeling the glottal flow derivative and may not completely characterize all the speaker information present in the glottal wave.

The fourth study was on the LP residual of speech. The motivation was that, since the residual of the speech signal is obtained after removing the vocal tract characteristics, the information present in the residual may contain only excitation information. A text-independent speaker identification study was conducted on two independent sets of 20 speakers (each) from NIST 99 database. Blocks of 5 ms with a shift of 125 μs of LP residual are used as feature vectors and hence subsegmental level processing. These feature vectors were modeled using AANN. The average identification performance of 77.5% was achieved [27]. On the same dataset, the average performance of the vocal tract information is 90% and improved to 95% by combining the evidences from the LP residual. The improvement in the performance indicates the different nature of the speaker information present in the LP residual. The structure of AANN model is based on some preliminary experiment. A systemic study is required to determine the suitable structure of the model. Also, since the neural network models are trained with every sample shift, the approach is computationally intensive. The usefulness of this approach require more confirmation and understanding with larger database.

The fifth study used the phase information of the LP residual around glottal closure instants for text-independent speaker verification task. The motivation was that the sequence of samples may also contain speaker information. Blocks of 5 ms with a shift of 125 μs LP residual phase sequences were used used as features. Hence, this approach may be attributed to the subsegmental level processing. The feature vectors were modeled using AANN. The EER was achieved to be 22% [13]. On the same database, MFCC achieved an EER of 14%. The combi-

nation of both the evidences further improved the performance to an EER of 10.5%, indicating the different nature of the speaker information present in them. Although the performance was achieved on a large database, but similar to earlier study the AANN model used were not optimized in terms of the network structure and training. Also the evidences from both the features may be combined in a better way to achieve more improved performance.

The sixth study is the first attempt on extracting the speaker information from the LP residual spectrum by cepstral processing. A text-independent speaker verification study was conducted on 50 speakers database collected in the laboratory. The cepstral representation of the LP residual achieved 12.9% EER [9]. On the same database, the vocal tract information achived 5.7%. By combining both the evidences the performance is further improved to 4.0%. The cepstral representation can be further improved by employing mel-frequency filter banks. The cepstral analysis models only the slowly varying components in the LP residual spectrum at low indices of cepstrum and thus miss the fast varying components. Alternative techniques for complete and compact representation of the LP residual spectrum can be explored.

The seventh study used the speaker information present in the harmonic structure of the LP residual spectrum for text-independent speaker identification task on 50 speakers set. The motivation for this study was that the nature of periodicity and hence the harmonic structure may be unique for each speaker. To demonstrate this, PDSS values are computed using the spectral flatness of the residual spectrum and hence may be attributed to segmental level processing. The features are modeled using VQ modeling technique and achieved 66.9% identification accurac [10]y. By combining with the vocal tract information, its individual performance of 98% improved to 99%. The proposed PDSS measure is one way of modeling the speaker information present in the harmonic structure. It may be possible to develop new methods for modeling speaker information in a better way to improve the performance.

The eighth study was on exploiting the speaker information from time frequency analysis of the LP residual segments using wavelet transform and may be attributed to segmental level processing. The motivation was that the multi-resolution analysis may provide better manifestation of speaker information. The WOCOR values are computed using the wavelet

transform analysis of LP residual and used as features for text-independent speaker verification task on 74 male speakers from NIST-2001 database. The system gave a performance of 21.8% EER [29]. By combining with vocal tract information, its individual performance of 9.3% improved to 7.67%. It may be possible to develop a better way of time frequency analysis of LP residual like matching pursuit for modeling speaker information.

Table 2.3: Summary of speaker recognition studies using source features. In this table, custom database refers to the case where speaker recognition database is collected in their own lab.

Source Feature	**Database (size)**	**Modeling Technique**	**Task (performance)**
Pitch contour [8]	Custom (10)	Template Matching	Identification (97.0%)
Jitter & Shimmer parameters [47]	NIST-2001 (543)	GMM	Verification (EER 26.9%,10%) (EER 26.9%,8.6%)
LP residual [11]	NIST-99 (40)	AANN	Identification (77.5%,90%,95%)
LP residual phase [13]	NIST-2003 (149)	AANN	Verification (EER:22.0%,14%,10.5%)
LP residual cepstrum [9]	Custom (50)	VQ	Verification (EER:12.9%,5.7%,4%)
Harmonic structure of LP residual [10]	Custom (50)	VQ	Identification (66.9%,98%,99%)
Time-frequency analysis of LP residual [29]	NIST-2001 (74 males)	GMM-UBM	Verification (EER:21.8%,9.3%,7.67%)
Glottal flow derivative parameters [15]	TIMIT NTIMIT (168)	GMM	Identification (TIMIT: 60%) (NTIMIT: 55%)

A summary of important facts of this comparative study is given in the Table 2.3. All the studies employed text independent recognition except the first work i.e. pitch contour. In the first case, different utterances of the same text were used for training and testing. In most of the cases standard databases are used and in some cases custom data collected in their own

laboratory environment are used. Number of speakers used for recognition are given inside brackets of the second column. All cases demonstrate the presence of speaker information in the excitation. But in the general speaker recognition context, the stand alone performance of speaker information from excitation is not at par with that of vocal tract information. Therefore, in the state-of-art speaker recognition systems, vocal tract information is used as primary information and excitation information is used as an additional information.

2.7 Summary and Scope for Excitation Source Related Work

The performance of the existing speaker recognition systems using excitation information are evaluated using different databases. Since the databases are different, their performances are not directly comparable. To know their effectiveness, all the excitation features need to be reevaluated on a common speaker recognition database. Some of the studies are conducted in verification mode and others in identification mode. For better comparison, all the studies need to be repeated for the same task. The different excitation features are extracted using the available methods then. These measurements may not be as accurate as with the state-of-the-art methods. For Example, pitch extraction method using epochs is reported to be giving the most accurate values of pitch [62, 63]. This pitch extraction method can therefore be used for pitch and its variants related studies. On the similar lines, benefit may be achieved by extracting each of these source features with state-of-the-art methods. The better the accuracy in the measurement of particular source feature, better may be the speaker recognition performance.

All the existing studies try to model different aspect of excitation information. However, there is no comparison among different excitation features, to know how different they are. If they are different, they may be used in combination to further improve the speaker recognition performance. Based on their effectiveness, a subset of excitation features may be selected for an unified framework of speaker recognition using excitation information. A system may then be developed that uses all these excitation features to improve the performance. The potential of the source features may then be compared with that using vocal tract. The performance of

the existing speaker recognition system based on the vocal tract information may be further improved by combining the evidences from the source and system features.

Based on these observations, the present work proposes the following direction for exploration: All the existing excitation features can be viewed under three analysis levels, namely, subsegmental, segmental and suprasegmental based on the block size at which the information is viewed for further processing. This work defines the block sizes in the range of 3-5 msec, 10-30 msec and more than 100 msec for subsegmental, segmental and suprasegmental analysis, respectively. In the existing literature, processing LP residual or its phase in blocks of 5 msec using AANN models may be viewed under subsegmental processing. Processing pitch, jitter, shimmer, cepstrum of LP residual, PDSS and WOCOR may be viewed under segmental processing. Processing pitch contour may be viewed under suprasegmental processing. Such a view gives us the first hand intuitive feel that the source information available may be different at each level. This is because, the focus at each level is different. In subsegmental level, the interest is on modeling excitation information present in 3-5 msec range which constitutes information within a glottal cycle. On the other hand more than 100 msec at suprasegmental level models the speaker information that may be present in contour manner. Finally 10-30 msec in case of segmental processing, the focus will be to model segmental information like pitch, segmental energy and so on.

Once we agree upon this broad classification of viewing excitation information, the next step will be to develop an unified framework including all of them. The unified framework is such that the incoming excitation signal is subjected to a signal processing operation which views the signal at these three levels and extracts relevant information present at each level. The extracted features may then be further modeled and combined to get as complete source information as possible. Once we have such an unified framework, then the next focus will be to how best we can model the speaker information at each level, in terms of representing speaker and also computations involved.

Finally what will be the best representation for the excitation signal that will be used for further processing. The earlier studies have shown that the LP residual extracted using a

proper LP order best represents the excitation information [12]. Therefore this work uses the LP residual as the excitation signal for extracting speaker-specific excitation information.

2.8 Organization of the Present Work

The LP residual as the representation of excitation signal can be processed in time, frequency, cepstral or any other suitable domain for extracting the relevant information. This work begins with developing approaches for modeling speaker-specific excitation information by processing the LP residual directly in the time domain. Chapter 3 develops techniques that operate in the temporal domain for modeling the speaker information from the LP residual at the subsegmental, segmental and suprasegmental levels. Initial part of the study will be to understand the potential of speaker information present in each of these levels. The later part of the study will be to understand how different the speaker information present at each of these levels. The final part of the chapter will be to develop an unified framework in the time domain for modeling speaker information from the LP residual. Since the LP residual signal itself is modeled without explicit extraction of parameters, the approach is termed as implicit modeling.

In chapter 4, an explicit way of modeling the subsegmental level excitation information is developed by processing the LP residual in time domain. First, a simple and approximate method is proposed for the computation of LF model parameters. The speaker-specific nature of the LF parameters is studied. Then, a feature compromising the LF parameters with their dynamics is proposed to model the subsegmental level excitation information. Finally, the experimental results of the proposed feature is compared with the corresponding subsegmental level processing of the LP residual.

Chapter 5 describes the methods for the explicit way of modeling the speaker-specific excitation energy and periodicity information by processing the LP residual at the segmental level. First, the LP residual is processed in frequency domain to model the energy and periodicity information of the excitation. A more effective way of modeling the speaker-specific excitation energy information is developed from the cepstral domain processing of the LP residual. Fi-

nally, the different nature of the speaker-specific information present in energy and periodicity of the excitation is studied and then an unified representative feature is proposed for modeling the speaker-specific excitation information at the segmental level. The experimental results of the representative feature are compared with the corresponding temporal domain segmental level LP residual vectors.

In chapter 6, a method is developed for the explicit way of modeling the suprasegmental level excitation information by processing the LP residual in time and frequency domains. First, a method is proposed to compute the pitch and epoch strength from LP residual of telephonic speech using instantaneous fundamental frequency approach. Then, with suitable dimension the pitch and epoch strength vectors are used to model the suprasegmental level information. In the later part of this chapter, the LP residual is processed in cepstral domain to model the suprasegmental level excitation trajectory information. Finally, the different nature of the speaker-specific information present in pitch, epoch strength and cepstral trajectories of the excitation is studied and then an unified representative feature is developed for modeling the complete speaker-specific excitation information at the suprasegmental level. The experimental results of the proposed representative feature are compared with the corresponding temporal domain suprasegmental level LP residual vectors.

A speaker verification system based on excitation information is developed in chapter 7. The best possible approach is used to extract the subsegmental, segmental and suprasegmental levels excitation information from the LP residual. Post feature extraction processing like, elimination of channel effect is performed. GMM-UBM modeling technique is used for building the speaker models [35]. Evidences from subsegmental, segmental and suprasegmental levels of the LP residual are combined at the decision level to achieve the maximum possible recognition performance from the excitation information prospective. Different speaker verification experiments on clean and noisy cases are conducted to verify the significance of the developed system. Finally, a comparative study is made with the state-of-the-art vocal tract system to know the potential of the excitation information for speaker recognition.

In chapter 8, first the summary of the different methods developed in this work is given.

Then, the major contributions of the present work in developing the state-of-the-art speaker recognition system based on excitation information from processing the LP residual in time, frequency and cepstral domains are mentioned. Finally, the chapter is concluded with a mention on the possible future direction of the present work.

3

Implicit Subsegmental, Segmental and Suprasegmental Processing of LP Residual for Speaker Information

Contents

In this chapter, the LP residual is processed in the time domain at subsegmental, segmental and suprasegmental levels to extract the speaker-specific excitation information and demonstrates their significance and also different nature for text-independent speaker recognition. The speaker identification and verification studies performed using NIST-99 and NIST-03 databases demonstrate that the segmental information provides the best performance followed by subsegmental analysis. The suprasegmental analysis gives the least performance. However, the evidences from all the three levels of processing seem to be different and combine well to provide improved performance, demonstrating different speaker information captured at each level of processing. The combined evidence from all the three levels of processing together with vocal tract information further improves the speaker recognition performance. The speaker information in the LP residual may be attributed to both the amplitude values and the sequence knowledge. An alternative approach using analytic signal concept is proposed to model the subsegmental, segmental and suprasegmental speaker information by separating the amplitude and sequence information of the LP residual. Experimental results show that the combined evidence from amplitude and sequence information at each level provides relatively better performance than the corresponding LP residual vectors. In all the studies reported in this chapter, the speaker-specific information is implicitly modeled without any explicit parameter extraction and hence the title of the chapter.

3.1 Introduction

As observed from the Chapter 2, majority of the existing studies on exploring the speaker-specific source information use the LP residual as the best approximation of the excitation signal. The LP residual is processed in time, frequency, cepstral and time-frequency domains to extract and model the speaker-specific information [9–12, 17, 29]. Processing the LP residual in time domain has the advantage that the artifacts of digital signal processing like block processing or windowing effect that creep in other domains of processing like frequency or quefrency are negligible [1]. The existing attempts in exploring the speaker-specific excitation information have processed the LP residual in time domain at subsegmental, segmental and

suprasegmental levels and demonstrated the significance. These studies are made independently and used different approaches for the extraction and modeling. An unified framework may be evolved where the LP residual is processed at subsegmental, segmental and suprasegmental levels using a single signal processing approach and use the same to study the level of speaker information present at each level and also their differences. The objective of this chapter is to propose one such approach and hence termed as temporal processing of LP residual for speaker information.

In the present work, the LP residual is processed in blocks of 5 msec with 2.5 msec shift for subsegmental, 20 msec with 2.5 msec shift for segmental and 250 msec with 6.25 msec shift for suprasegmental information. The 5 msec blocks of LP residual sample sequences in the time domain are used as feature vectors for modeling speaker information by the GMM technique to generate the subsegmental speaker models. The 20 msec blocks of LP residual samples are first decimated by a factor of 4 to reduce their dimensionality and also to eliminate the information that has been modeled at the subsegmental level. The decimated LP residual sample sequences are then modeled by the GMM to generate the segmental speaker models. The 250 msec blocks of LP residual samples are first decimated by a factor of 50 to reduce its dimensionality and also to eliminate the information that have been modeled both at the subsegmental and segmental levels. The decimated LP residual sample sequences are modeled by the GMM to generate suprasegmental speaker models. All these models are independently tested using respective blocks of LP residual extracted from the test signals to evaluate the amount of speaker information present at each level. The comparison of evidences from all the three levels is made to observe their different nature of speaker information. The potential of combined source information is demonstrated by combining their evidences. The combined source evidence based system is finally combined with a speaker recognition system using vocal tract information for further improving the performance. The earlier attempts of modeling speaker information from the LP residual in time domain used the AANN models for exploiting sequence information [11, 12]. In the present work an alternative view is taken for the LP residual samples. The LP residual signal is like a random noise sequence, except for the pitch

information. If we treat the residual signal as random noise-like signal, then the distribution of the samples will be Gaussian. Since the LP residual deviates from random noise due to pitch information, to that extent the distribution of the residual samples may be non-Gaussian in nature. However, this can be handled with the help of the GMM. Hence the motivation for using GMM for speaker modeling from LP residual.

The speaker information in the LP residual may be due to the variation in the amplitude and phase. When the LP residual is processed directly, the effectiveness of their individual evidences may not be manifested properly. For example, the amplitude values dominate over the sequence information. It may therefore be better to separate the amplitude and sequence information and then process them independently. In [13], the phase of the analytic signal of the LP residual had been used to model the speaker-specific excitation sequence information at the subsegmental level only. In this work, we use both amplitude and phase of the analytic signal of the LP residual to model the amplitude and sequence information independently. Then, the individual evidences from the amplitude and phase of the LP residual are combined. As in the previous attempt we employ the analytic representation for subsegmental, segmental and suprasegmental levels and a best possible way of combining the evidences from each of these levels is proposed for improved representation of the speaker-specific excitation information. Finally, a comparative study is made with the direct LP residual processing approach.

The rest of the chapter is organized as follows: Section 3.2 explains the experimental setup made to evaluate the effectiveness of the proposed features from time domain processing of the LP residual for speaker recognition. The experimental setup is later used throughout the reminder of this thesis work for evaluating the effectiveness of the proposed features in different chapters. Section 3.3 describes the proposed subsegmental, segmental and suprasegmental processing of LP residual approach for modeling speaker information from the LP residual. This section will also describe the speaker recognition studies that have been performed using the proposed approach. Section 3.5 describes an alternative approach for subsegmental, segmental and suprasegmental analysis using the analytic signal concept and demonstrates its significance in modeling speaker information. Finally, the summary of the work discussed in this chapter

is given in Section 3.6.

3.2 Experimental Setup

We conduct both speaker identification and verification experiments to evaluate the effectiveness of the proposed features. Initial studies are made by conducting the speaker identification experiments on small data sets. The effectiveness of the proposed approaches is further verified on a larger database by speaker verification experiments. The performance achieved from identification and verification experiments by a feature may also enable us to verify its inter and intra-speaker variability nature. For example, a feature with higher identification accuracy may indicate large inter-speaker variability. On the other hand, a feature with higher verification accuracy may indicate less intra-speaker variability. For both identification and verification studies a common modeling technique is employed.

3.2.1 Modeling Technique and Testing

Existing studies on modeling the speaker-specific excitation information by processing the LP residual particularly in time domain have extensively used the AANN modeling technique for building the speaker models [11–13, 64]. A major advantage of using the ANNN technique is that the feature extraction and speaker modeling can be combined into a single network, enabling joint optimization of the (speaker-dependent) feature extractor and the speaker model [18, 27, 65, 66]. However, proper attempt needs to be made to optimize the parameters of the network for feature extraction, and also the decision making stage [12]. Another major drawback of the AANN is that the complete network must be retrained when a new speaker is added to the system and thus is computationally intensive [67]. On the other hand, standard established Gaussian mixture speaker modeling technique is simple and computationally less intensive [67]. One of the disadvantage of the GMM technique is that it requires large amount of speech data for building the speaker models. Since, the speaker recognition system based on excitation information requires less amount of data for capturing the speaker-specific evidence, it is expected that the data amount limitation may be compensated and the GMM technique

can be general enough to model the speaker-specific excitation information from the LP residual. Therefore, in our work we prefer to use probabilistic Gaussian mixtures speaker modeling technique for building the speaker models.

In our experiments, the objective is to verify and compare the effectiveness of the different approaches proposed for modeling the subsegmental, segmental and suprasegmental levels speaker-specific excitation information by processing the LP residual, and then decide the best possible approach to develop a speaker recognition system. Therefore, first the initial study is made using GMM modeling technique [4, 5]. The potential of the final approaches proposed towards the end of this thesis work is then evaluated by GMM-UBM modeling technique [35].

In GMM modeling technique, each model λ is collectively represented by Gaussian mixture density parameters $w_i, \mu_i, \sum_i$, where $i = 1, ..., M_g$ is the number of the Gaussian mixture and w_i, μ_i, ρ_i represent the weight, mean and variance of the i^{th} mixture, respectively. The details of GMM modeling technique is given in *Appendix-C*. For a given test feature x_t, the likelihood of belonging to a model λ_s is given by

$$P(x_t|\lambda_s) = \sum_{i=1}^{M_g} w_i p(x_t|\lambda_s) \tag{3.1}$$

Where, $p(x_t|\lambda_s)$ is the Gaussian mixture density of the feature x_t to a model λ_s. To maximize the likelihood iterative expectation maximization (EM) algorithm, with an initial model trained using the k-means algorithm are popularly used [4, 5, 68, 69]. In our experiment 50 iterations are made for EM and M_g is chosen as 128 for building the speaker models.

The decision is taken based on the log-likelihood score. For a given test data X having $N_F = x_1, ..., x_F$ number of feature vectors, the log-likelihood score given to a model λ_s, $P(X|\lambda_s)$ is obtained by Equation 3.2.

$$P(X|\lambda_s) = \sum_{f=1}^{N_F} logP(x_f|\lambda_s) \tag{3.2}$$

In case of identification task, for a given test speech the model that produced the maximum log-likelihood score is considered as the identified speaker. In case of verification task, an experimental threshold from the log-likelihood scores of the trained models is considered. The

test speaker is accepted, if the log-likelihood score to the claimed model is above the threshold, else the claim is rejected. For comparison, the feature vector with higher identification accuracy and lower EER is considered to be containing more speaker-specific information.

3.2.2 Speaker Recognition Database

In this work the speech data from both NIST-99 and NIST-03 databases are used for experimental study [54, 58]. The speech data present in these databases are sampled at 8 kHz. NIST-99 collected over landline is used as the representation of clean data. NIST-03 collected over mobile phones is used as the representation of relatively noisy data. Two small data sets, called as *Set-1* and *Set-2* are selected for speaker identification study. The speech data of *Set-1* and *Set-2* speakers are collected from NIST-99 and NIST-03 databases, respectively. The speakers having matched condition and testing data of at least 30 sec are considered. Each set consists of 90 speakers that includes 48 males and 42 females. Since, the test data is guaranteed to be from the given reference model, the identification task is a *Closed-set* identification. The speaker verification experiment is conducted on the whole NIST-03 database that consists of 356 target speakers. Each speaker has a training data of about 2 min duration, which is used to build the models. There are totally 2559 test utterances with duration of 15-45 sec. Each test utterance is tested against 11 hypothesized speakers that include the genuine speaker and ten imposters.

3.3 Processing of LP Residual in Time Domain

In LP model of speech production, each sample of speech is predicted as a linear combination of the past p samples, where p represents the order of prediction [12, 24]. If $s(n)$ is the present sample, then it is predicted by the past p samples as

$$\hat{s}(n) = -\sum_{k=1}^{p} a_k s(n-k) \tag{3.3}$$

where, $a_k s$ are the LP coefficients (LPCs) computed by minimizing the mean square prediction error. The procedure to compute LPCs is described in *Appendix-A*. The error between the

actual and the predicted sample value is called as the prediction error or LP residual and is given by

$$r\left(n\right)=s\left(n\right)-\hat{s}(n)=s\left(n\right)+\sum_{k=1}^{p}a_{k}s\left(n-k\right) \tag{3.4}$$

The LP residual $r(n)$ is obtained by passing the speech signal through an inverse filter $A(z)$ given by

$$A(z)=1+\sum_{k=1}^{p}a_{k}z^{-k} \tag{3.5}$$

The predicted samples $\hat{s}(n)$ model the vocal tract information in terms of (LPCs) [24, 70]. The suppression of this information from the speech signal $s(n)$ that results in the LP residual $r(n)$ is therefore mostly contains information about the source. So the source signal can be approximated by the LP residual. The representation of source information in the LP residual depends upon the order of prediction. In [12], it was shown that for a speech signal sampled at 8 kHz, the LP residual extracted using LP order in the range 8-20 best represents the speaker-specific source information. In this study, LP residual computed using 10^{th} order LP analysis followed by inverse filtering the speech signal sampled at 8 kHz is used as the source signal. Example of the speech and LP residual signals are shown in Figure 3.1 (a) and (b), respectively. The instants around the peaks in the LP residual are termed as epochs [63, 71]. Significant speaker-specific excitation information is present around the region of the epochs [13]. These include the strength and rate of occurrence of the epochs, and their temporal variations across several glottal cycles. In this section we describe the methods employed in processing the LP residual to extract speaker information. In extracting such information we consider subsegmental, segmental and suprasegmental level processing of LP residual [20]. In subsegmental processing, features are derived to represent the speaker information present mostly within one glottal cycle. In segmental processing, features are derived to represent the speaker information mostly related to pitch and energy of the excitation present across 2-3 glottal cycles. In suprasegmental level processing, features are derived to represent the prosodic aspects of the speaker present across about 25-50 glottal cycles. The effectiveness of the speaker-specific information present in subsegmental, segmental and suprasegmental levels

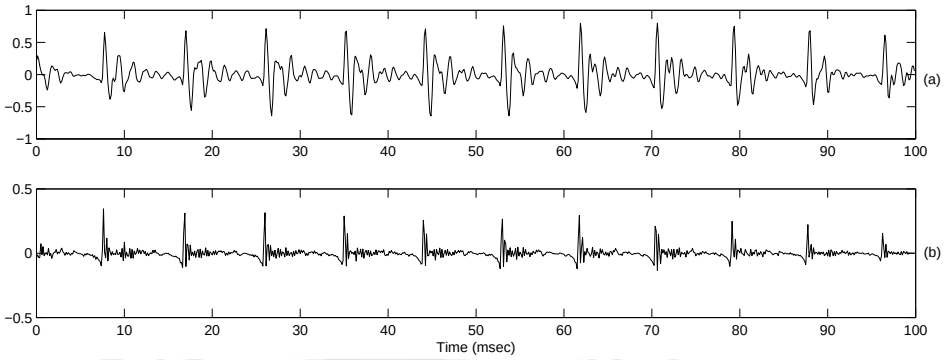

Figure 3.1: Speech and LP residual. (a) Voiced segment of speech (b) Corresponding 10^{th} order LP residual.

is demonstrated by the speaker identification and verification experiments.

3.3.1 Speaker Information from Subsegmental Processing of LP Residual

At the subsegmental level, speaker information present mostly within one glottal cycle is modeled. This information may be attributed to the activity like opening and closing glottal characteristics. To model this information, the LP residual is blocked into frames of 5 msec with a shift of 2.5 msec. For 5 msec at 8 kHz, the frames have 40 samples. One such frame is shown in the Figure 3.2(b) and its spectrum is shown in Figure 3.2 (c). The largest amplitude of the samples of the vector indicate the strength of excitation. The samples in the vector represent the sequence information of glottal cycle. Since these frames are obtained from the LP residual sampled at 8 kHz, they will have excitation information present as the fine variations represented by frequency components up to 4 kHz. These frames of LP residual samples in the time domain are used as the feature vectors to represent the speaker information at the subsegmental level and used for speaker recognition experiments. The nature of the LP residual signal that will be processed at the subsegmental level is the one shown in Figure 3.2(a). This is nothing but the original LP residual.

The results of identification and verification experiments are given in the first row of the Table 3.1 and Table 3.2, respectively. In these tables the performance of the vocal tract based

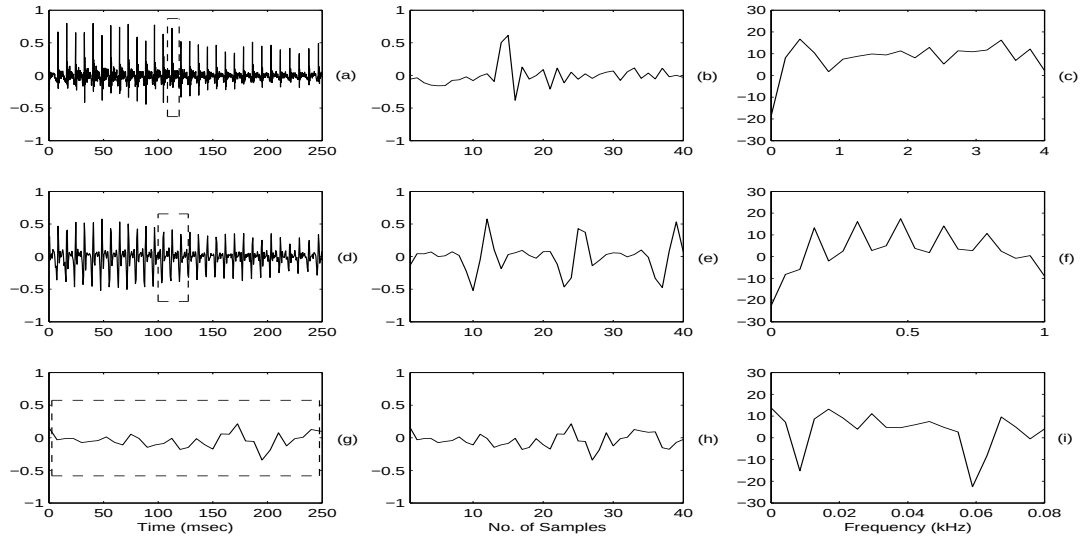

Figure 3.2: Temporal sequences and their spectra from subsegmental, segmental and suprasegmental processing of LP residual. (a) LP residual. (b)-(c) Subsegmental sequence and its spectrum, respectively. (d) LP residual decimated by a factor 4. (e)-(f) Segmental sequence and its spectrum, respectively. (g) LP residual decimated by a factor 50. (i)-(j) Suprasegmental sequence and its spectrum, respectively. The dotted box in (a), (d) and (g) represents the nature of the LP residual that will be processed at subsegmental, segmental and suprasegmental levels, respectively.

Table 3.1: Speaker identification performance (in %) of subsegmental (Sub), segmental (Seg), suprasegmental ($Supra$) and spectral ($MFCC$) information for two subsets of 90 speakers. Src_1 represents $Sub + Seg$. Src_2 represents $Sub + Seg + Supra$. $Comb_1$ and $Comb_2$ represent linear score level and logical OR combination schemes.

$Feature$		$Performance$		$Relative$
		$Set\text{-}1$	$Set\text{-}2$	$Degradation$
Sub		64	57	11
Seg		60	58	3
$Supra$		31	13	58
Src_1	$Comb_1$	64	60	6
Src_2	$Comb_1$	68	60	12
$Src_2 + MFCC$	$Comb_1$	84	70	17
Src_1	$Comb_2$	71	67	6
Src_2	$Comb_2$	76	67	12
$Src_2 + MFCC$	$Comb_2$	96	79	18
$MFCC$		87	66	24

features namely, MFCC is also given. The computational procedure of MFCC is described in *Appendix-B*. It is to be cautioned at this stage that the speaker verification system using MFCC and GMM is a baseline system without any normalization technique. Hence the performance itself is poor compared to the state-of-the-art on NIST-03 [58]. Since our objective is only relative comparison among source and vocal tract features, we have settled to the baseline system. The results show that subsegmental features provide good performance and hence contain speaker information. However the performance is comparatively poorer than the vocal tract features. The reason may be that the subsegmental features contain only one aspect of source information. The performance can be improved by using additional information from segmental and suprasegmental levels.

It is interesting to observe that the performance of both subsegmental source information and vocal tract features degrade in case of NIST-03, as expected. However, the amount of degradation in the performance is relatively less in case of subsegmental source information, about 11%, as against to 24% in case of vocal tract features. This demonstrates the relative robustness of excitation information present at the subsegmental level.

Table 3.2: Speaker verification performance (in EER) of subsegmental (Sub), segmental (Seg), suprasegmental ($Supra$) and spectral ($MFCC$) information for whole NIST-03 database. Src_1 represents $Sub + Seg$. Src_2 represents $Sub + Seg + Supra$. $Comb_1$ and $Comb_2$ represent linear score level and logical OR combination schemes.

$Feature$		$Performance$
Sub		41.01
Seg		26.96
$Supra$		44.49
Src_1	$Comb_1$	32.02
Src_2	$Comb_1$	32.25
$Src_2 + MFCC$	$Comb_1$	27.78
Src_1	$Comb_2$	23.21
Src_2	$Comb_2$	21.22
$Src_2 + MFCC$	$Comb_2$	17.43
$MFCC$		22.94

3.3.2 Speaker Information from Segmental Processing of LP Residual

At the segmental level, speaker information present in two to three glottal cycles is modeled. This information may be attributed mostly to pitch and energy. Speaker information represented by variations within a glottal cycle have already been modeled by subsegmental analysis. In segmental level processing of LP residual, other information that can be observed at the segmental level needs to be emphasized. For this we propose to decimate the LP residual by a factor 4 so that the sampling rate becomes 2 kHz and we may have source information up to 1 kHz. The decimated LP residual is shown in Figure 3.2(d). Even after decimation, the dominant speaker information at the segmental level, that is, pitch and energy information, still can be preserved. Moreover, in segmental level processing, LP residual frames of 20 msec duration are used as the feature vectors. For 20 msec at 8 kHz, the feature vectors with 160 samples is of very large dimension for building the models. By decimating the LP residual by a factor 4, the dimension of the feature vectors is reduced to 40 samples per vector which is equal to the subsegmental feature vector length. Since the LP residual is decimated by a factor 4, we prefer to compute the feature vectors for every 2.5 msec frame shift so that the

number of feature vectors will remain same as the subsegmental feature. One such feature vector derived from the decimated LP residual is shown in Figure 3.2(e). It contains mainly the pitch and energy information. The fine variations within the glottal cycle are suppressed by smoothing. Similar observation can also be made from the spectrum of the feature vector shown Figure 3.2(f). The periodicity and the amplitude of the spectrum clearly represent the pitch and energy information. This observation indicate that segmental feature vectors reflect different aspect of source information compared to subsegmental feature vectors. This will also be confirmed from the comparison study in Section 3.4.

The effectiveness of these features are evaluated from the identification and verification experiments. The results are given in the second row of the Table 3.1 and Table 3.2, respectively. The high performance show that segmental features contain good speaker information, even better than those contained at the subsegmental level. This shows that the pitch and energy may be dominating speaker-specific source information. The recognition performance is comparatively poor than vocal tract features. The same reason of incomplete representation of speaker information may be attributed. The segmental source features are relatively more robust compared to both vocal tract as well as subsegmental features, since it shows only about 3% relative degradation in the performance from *Set-1* and *Set-2* database.

3.3.3 Speaker Information from Suprasegmental Processing of LP Residual

Subsegmental processing models speaker information up to 4 kHz. Segmental processing models speaker information up to 1 kHz. Beyond that LP residual also contains some speaker information at very low frequency range, that is, may be less than 100 Hz. For example the variation in pitch and energy across several glottal cycles [8, 72]. In capturing such information, we need to process the LP residual at the suprasegmental level, for example, with frames of 100-300 msec range. For the LP residual sampled at 8 kHz, the feature vectors from such frames will be of very large dimension for building models. We prefer to decimate the LP residual by a factor 50 so that the sampling rate becomes 160 Hz and we may have the source information

up to 80 Hz. The dimension of the feature vector is also reduced by 50 factor. Further, the high frequency information that is already modeled by subsegmental and segmental level processing will be smoothed out. Therefore in suprasegmental level processing of LP residual, we decimate the LP residual by a factor of 50 and process in frames of 250 msec with shift of 6.25 msec. The frame size is decided so that the dimension of the feature vectors will remain same as in subsegmental and segmental processing. However, the minimum possible frame shift in this case is 6.25 msec which corresponds to one sample shift. Figure 3.2(h) shows a suprasegmental feature derived from the decimated residual shown in Figure 3.2(g). The fast varying components of the original LP residual are eliminated and it mostly represent the long term variations. This can also be observed from the spectrum of the shown feature vector from Figure 3.2(i). Information present in the smoothed spectrum is up to 80 Hz. The periodicity and other high frequency related information are absent.

The speaker information present in these features is verified from the recognition experiments as performed earlier. The results of the identification and verification experiments are given in the third row of the Table 3.1 and Table 3.2, respectively. Results show that suprasegmental level features contain some speaker information. The recognition performance is significantly poor compared to subsegmental, segmental and vocal tract information. The poor result indicates that the suprasegmental features may have large intra-speaker variability. The other major factor is text-independent mode of operation. However, it may contain different aspect of speaker information and hence combine well with other features.

3.4 Combining Evidences from Subsegmental, Segmental and Suprasegmental Levels of LP Residual

As described in the previous section, by the way of deriving each feature, the information present at subsegmental, segmental and suprasegmental levels are different and hence may reflect different aspect of speaker-specific excitation information. By comparing their recognition performance it can be observed that the segmental features provide best performance. Thus the segmental features may have more speaker-specific evidence compared to other level fea-

tures. The different performances in the recognition experiments indicate the different nature of speaker information present. In this section we use confusion patterns and scatter diagrams to further explain the different nature of the speaker information present in the proposed features and their usefulness for combined use in speaker recognition.

In case of identification, the confusion pattern of features is considered as an indication of the different nature of information present [73]. In the confusion pattern, principal diagonal represents correct identification and the rest represents miss classification. Figures 3.3 and 3.4 show the confusion patterns of the identification results conducted for all the proposed features using *Set-1* and *Set-2* databases, respectively. In each case, the confusion pattern is entirely different. The decisions for both true and false identification are different. This indicates that they reflect different aspect of source information. This may help in combining the evidences to further improve the recognition performance from the source perspective.

For combination we use score level fusion and logical *OR* combination scheme [74]. In this work the score level and logical *OR* combinations are abbreviated as $Comb_1$ and $Comb_2$, respectively. In the score level fusion, the respective scores are weighted by their performances and linearly combined. For example, the log-likelihood ratio (LLR) of the combined system, LLR_c, is given by the following relation:

$$LLR_c = \sum_{i=1}^{C_j} \frac{P_i}{\sum_{i=1}^{C_j} P_i} \times LLR_i \tag{3.6}$$

where, C_j is the number of systems combined, LLR_i and P_i are the LLR and identification performance of the i^{th} system, respectively. In case of verification mode, the P_i in the above equation is replaced by the reciprocal of the respective EER and then the scores of the combined system is computed accordingly. The performance of the linearly combined systems are given in Table 3.1. In all the cases, the performance is improved compared to their respective individual performance. In case of *Set-1* database, the performance is improved from 64% to 68% and in case of *Set-2* database from 57% to 60%. It should be noted here that the small improvement in the performance should not be confused with the worth of combined use of all the features as a

best representation of the source information. It is because the performance of the combination system also depends on the combination scheme employed. It is well known that, simple linear

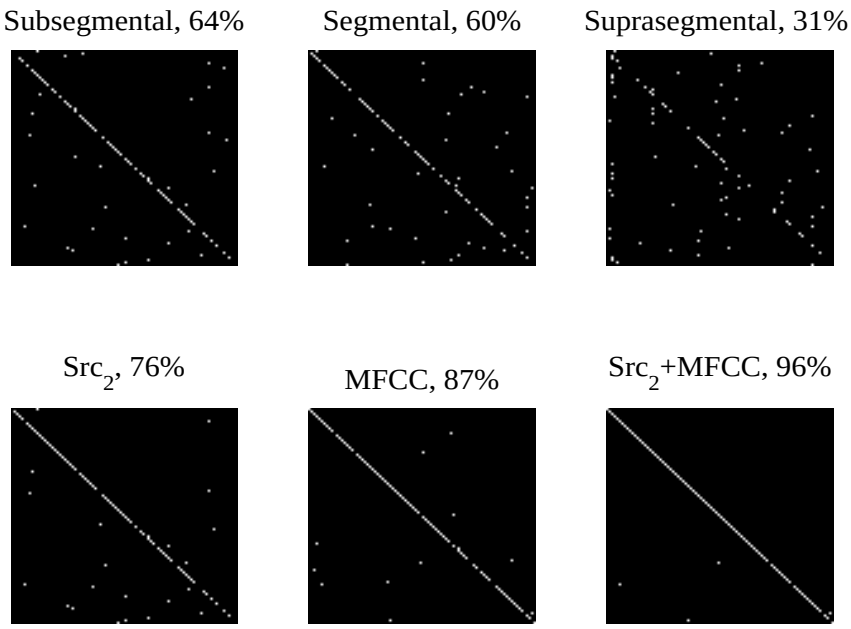

Figure 3.3: Confusion patterns of subsegmental (Sub), segmental (Seg), suprasegmental ($Supra$) and spectral ($MFCC$) information from identification results of *Set-1* database.

combination with predefined weights may not necessarily provide the best result [29]. This is because, fusion of scores may result in a wrong decision. To get a feel of the potential of the combined use of the features in representing the source information, we use logical OR combination. In this combination, if any one system is giving correct decision, we consider it as a correct decision. The performance of the OR combined systems are also given in the Table 3.1. The results show that the maximum benefit we can achieve from the proposed features for the *Set-1* and *Set-2* databases are 76% and 67%, respectively. This result shows that if we have a suitable combination scheme, we will benefit by the proposed features. In comparison with the vocal tract information, the confusion patterns of the combined system is different from the vocal tract system. By combining evidences from both the features ($Src_2 + MFCC$), the respective performances given in the Table 3.1 are improved. This indicates that the proposed

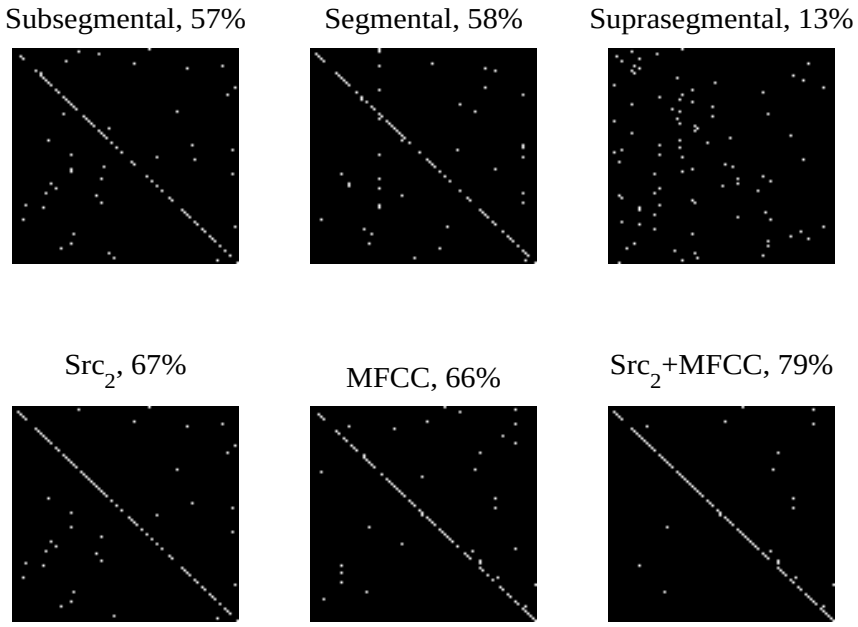

Figure 3.4: Confusion patterns of subsegmental (*Sub*), segmental (*Seg*), suprasegmental (*Supra*) and spectral (*MFCC*) information from identification results of *Set-2* database.

feature is well combined with the vocal tract information.

In case of verification, as suggested in [29], the different aspect of speaker-specific information in the three features are verified from their distribution of scores for imposter and genuine trails. Distribution of two dimensional (2-D) LLR scores for genuine and imposter trials among subsegmental, segmental and suprasegmental features are shown in Figures 3.5(a) to (c), respectively. In these figures o represent genuine and x represent imposter speaker. In the regions marked as marked as I and II, the respective features give different decision. For example, in region I, feature represented by x-axis rejects, but the other one accepts. Similarly in region II, feature represented by x-axis accepts but the other one rejects. Further, in these regions, some genuine rejected and imposters accepted by one feature are corrected by other. These observations indicate the different nature of speaker information present in these features. In combining the evidences, we use two combination techniques such as linear and logical *OR*

combination scheme. In linear combination, weighted scores are combined linearly. In logical *OR* combination, the true scores around the mean provided by the good system are modified based on the information provided by the poor system. In case of linear combination, the performance given in Table 3.2 is decreased. The reason may be as mentioned earlier. In case of logical *OR* combination, the performance achieved for the combined system as shown in Table 3.2 is 21.22% which is even better than the MFCC features. This shows that it is indeed possible to get better performance from the source than vocal tract information, provided we have suitable combination technique. In case of combining the evidences from the proposed feature with MFCC, performance is further improved by the logical *OR* combination scheme. From this section we observe that the combined use of subsegmental, segmental and suprasegmental features provide useful speaker-specific excitation information. This information is also well combined with the vocal tract information to improve the recognition accuracy. Individually the subsegmental, segmental and suprasegmental features are not providing recognition performance at par with vocal tract information. The reason may be that each of them represent one aspect of speaker information due to source. The results given in Table 3.1 and Table 3.2 show that, the combination of the subsegmental, segmental and suprasegmental level information performs slightly better compared to vocal tract information. These results are interesting because they demonstrate as a proof of concept that it is indeed possible to achieve

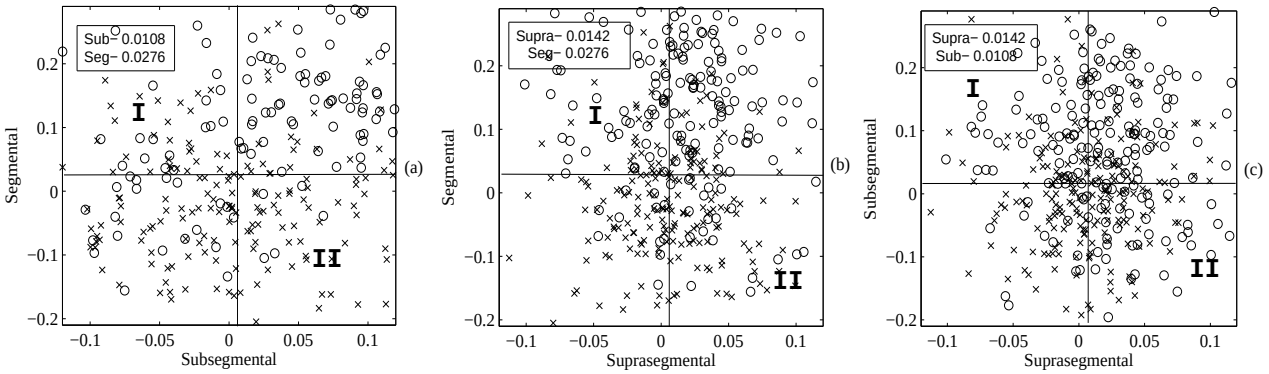

Figure 3.5: Distribution of 2-D LLR scores of, (a) Subsegmental and segmental information, (b) Suprasegmental and segmental information, (c) Suprasegmental and subsegmental information.

speaker recognition performance using only excitation information, which is either comparable or even better compared to the vocal tract information.

The speaker information in the LP residual may be attributed to both the amplitude values and the sequence knowledge. In the next section we describe a method in extracting the subsegmental, segmental and suprasegmental speaker information by separating the amplitude and sequence information of the LP residual. The method involves analytic signal representation of the LP residual. Since the amplitude and sequence information are two different aspects of speaker information, their combined effect may provide improved performance.

3.5 Speaker Information using Analytic Signal Representation of LP Residual

In the previous section, speaker information from the LP residual was derived by direct processing of the LP residual at the subsegmental, segmental and suprasegmental levels. The dominant speaker information present in these three levels of processing mostly represents the amplitude and sequence information of the source. When the LP residual is processed directly, the effect of amplitude values dominate over the sequence information, especially, around the instants of glottal closure [13]. It may therefore be better to separate the amplitude and sequence information and then process them independently. One approach to achieve this is with the use of analytic signal representation of the LP residual [55, 56]. In this representation, the magnitude of the analytic signal of LP residual represents the amplitude values of the LP residual and the cosine of the phase of the analytic signal represents the sequence information. Thus the analytic signal representation of the LP residual may help in exploiting the amplitude and sequence information separately. We propose to derive the subsegmental, segmental and suprasegmental features from the analytic signal representation of the LP residual.

The analytic signal of the LP residual $r_a(n)$ corresponding to the LP residual $r(n)$ is given by [56]

$$r_a(n) = r(n) + jr_h(n) \tag{3.7}$$

where $r_h(n)$ is the Hilbert transform of $r(n)$ and is given by

$$r_h(n) = IFT[R_h(\omega)] \tag{3.8}$$

where

$$R_h(w) = \begin{cases} -jR(w), & 0 \leq w < \pi \\ jR(w), & 0 > w \geq -\pi \end{cases} \tag{3.9}$$

$R(\omega)$ is the fourier transform of $r(n)$ and IFT denotes the inverse fourier transform. The magnitude of the analytic signal, called as the Hilbert envelope (HE) of the LP residual is given by [13]

$$|r_a(n)| = \sqrt{r^2(n) + r_h^2(n)} \tag{3.10}$$

and the cosine of the phase, called as the residual phase (RP) is given by [13]

$$cos(\theta(n)) = \frac{Re(r_a(n))}{|r_a(n)|} = \frac{r(n)}{|r_a(n)|} \tag{3.11}$$

The procedure to compute the subsegmental, segmental and the suprasegmental feature vec-

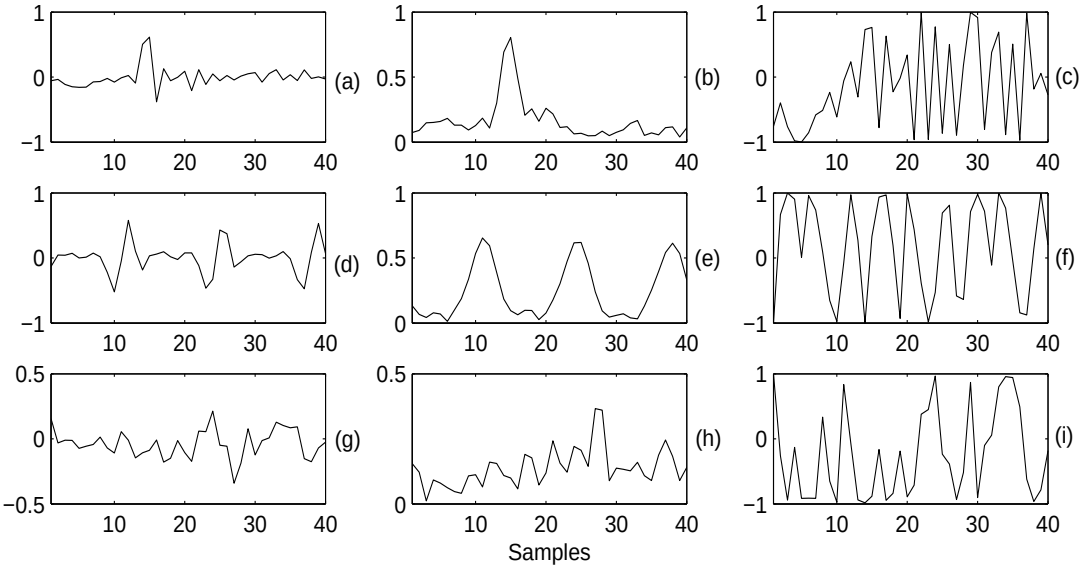

Figure 3.6: Decomposition of subsegmental, segmental and suprasegmental feature vectors using analytic signal representation. (a) Subsegmental feature vector. (b)-(c) HE and RP of subsegmental feature vectors, respectively. (d) Segmental feature vector. (e)-(f) HE and RP of segmental feature vectors, respectively. (g) Suprasegmental feature vector. (h)-(i) HE and RP of suprasegmental feature vectors, respectively.

tors from HE and RP of the LP residual is same as described earlier except the input sequence. In one case the input will be HE and the other case it will be RP. Example of subsegmental information derived from the LP residual and HE of the LP residual are shown in Figures 3.6(a) and (b), respectively. The unipolar nature of the HE helps in suppressing the bipolar variations representing sequence information and emphasizing only the amplitude values. As a result, the amplitude information in the subsegmental sequence of the LP residual is further emphasized by its HE counterpart. Similar observation can also be made in case of segmental and suprasegmental levels processing as shown in Figures 3.6(d) and (e), and Figures 3.6(g)and (h), respectively. On the other hand, the residual phase represents the sequence information of the residual samples. Figures 3.6(c), (f) and (i) show the residual phase of the subsegmental, segmental and suprasegmental processing, respectively. In all these cases, the amplitude information is absent. Hence analytic signal representation provides amplitude and sequence information of the LP residual samples independently. In [13], it was shown that information present in the residual phase significantly contributes to the speaker recognition. We propose that, the information present in the HE may also contribute well to speaker recognition. As they reflect different aspect of the source information, the combined representation of both the evidences may be more effective for speaker recognition. We conduct different experiments to verify this proposal. The observation from all these experiments are described next.

Subsegmental, segmental and suprasegmental sequences are derived from the HE and RP of the LP residual. In this study subsegmental, segmental and suprasegmental sequences derived from the LP residual, HE of the LP residual and phase of the LP residual are called as the residual features, HE features and RP features, respectively. The potential of the HE and RP features are verified from different recognition experiments. For fair comparison with the residual features, the experimental conditions remain same as mentioned earlier, except for the use of the HE and RP features.

The speaker identification performances of these features for both the datasets are given in Tables 3.3 and 3.4 and the verification performances for whole NIST-03 database is given in Table 3.5. In these tables the performance of the residual features are also given for comparison

purpose. For both the tasks, the performance of individual HE and RP features is comparatively poorer than their corresponding residual features. Because, as mentioned earlier, HE and RP features independently represent different aspects of the information that is present in the residual features. The different nature of the information present in the HE and RP features can also be observed from their confusion patterns obtained from the identification tasks. Figure 3.7 shows the confusion patterns of the identification results conducted for HE and RP features using NIST-99 database. At each level, the confusion patterns of the HE and RP features are different. Their decisions for both true and false identification are different. This indicates that the information present in HE features is different from that of RP features. By combining individual evidences, the respective performances may be further improved.

There are two approaches that can be used for combining evidences from HE and RP. In one approach, at each level, HE and RP can be combined independently (vertically) and this evidence at each level can be further combined to obtain overall source information. Alternatively, the HE and RP from all the three levels can be combined first (horizontally) and then these combined HE and RP evidences are further combined to obtain complete source information. From the experimental results we observed that the later approach seem to give better performance. The reason may be that HE and RP information from all the three levels together may combine well to become more speaker-specific, because their origin is same.

In combining the evidences we employ both $Comb_1$ and $Comb_2$ combination schemes described earlier. The identification performance of the various combinations for NIST-99 and NIST-03 databases are given in Table 3.3 and Table 3.4, respectively and the verification performance for the whole NIST-03 database is given in Table 3.5. The results show that for less noisy data (i.e. *Set-1*), the performance achieved from combined HE and RP features is better than the residual feature. For noisy data (i.e., *Set-2*), for both the tasks, the performance is slightly poor than the residual feature. The reason may be the quality of the data and the combination technique employed. For example in case of combination scheme $Comb_2$, the recognition performance is improved. In noisy condition, with MFCC features, the combined representation of the HE and RP features is providing better performance than the residual

feature. This shows the robustness of the combined HE and RP representation of the source in providing the additional information to the MFCC feature. From this observation we conclude that combined representation of HE and RP features may be better than the residual feature alone.

Table 3.3: Speaker identification performance (in %) of residual, Hilbert envelop (HE), residual phase (RP) and HE+RP features for *Set-1*. *Sub*, *Seg* and *Supra* represent subsegmental, segmental and suprasegmental LP residual sequence, respectively. $Comb_1$ and $Comb_2$ represent linear score level and logical *OR* combination schemes.

Feature		*Sub*	*Seg*	*Supra*	Src_2		*MFCC*	$Src_2 + MFCC$	
					$Comb_1$	$Comb_2$		$Comb_1$	$Comb_2$
Residual		64	60	31	68	76	87	84	96
HE		44	56	8	66	71		88	94
RP		49	69	17	69	73		86	93
HE+RP	$Comb_1$	57	69	13	74	88		87	98
	$Comb_2$	64	78	22					

Table 3.4: Speaker identification performance (in %) of residual, Hilbert envelop (HE), residual phase (RP) and HE+RP features for *Set-2*. *Sub*, *Seg* and *Supra* represent subsegmental, segmental and suprasegmental LP residual sequence, respectively. $Comb_1$ and $Comb_2$ represent linear score level and logical *OR* combination schemes.

Feature		*Sub*	*Seg*	*Supra*	Src_2		*MFCC*	$Src_2 + MFCC$	
					$Comb_1$	$Comb_2$		$Comb_1$	$Comb_2$
Residual		57	58	13	60	67	66	70	79
HE		32	39	7	47	54		70	76
RP		23	51	14	48	56		69	77
HE+RP	$Comb_1$	40	54	12	58	72		70	83
	$Comb_2$	48	59	17					

The above observations indicate that complete information present in the excitation can be represented by the combined representation of the HE and RP features. To achieve maximum benefit, it may be better to first combine the HE and RP at subsegmental, segmental and suprasegmental levels separately and then combine them. The speaker recognition performance of the information present in the segmental level is comparatively better than the other two

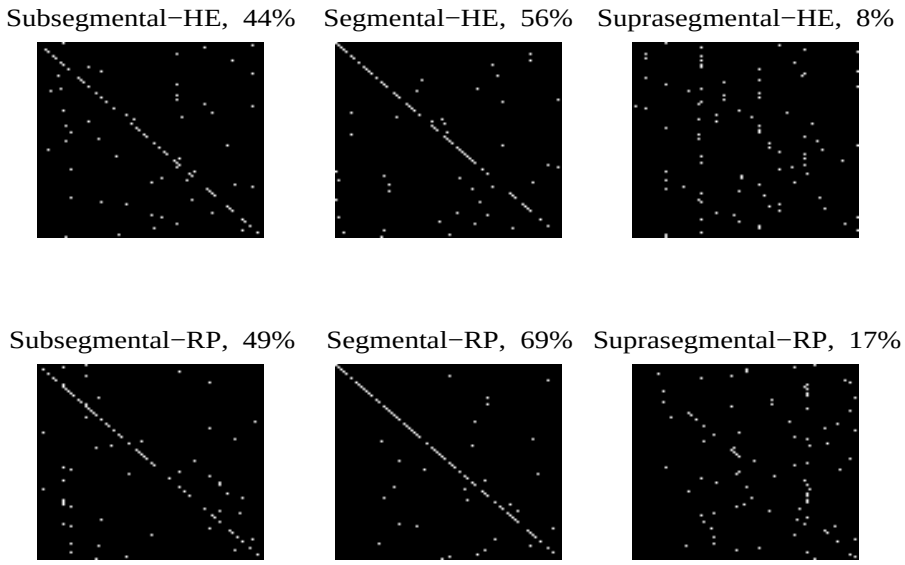

Figure 3.7: Confusion patterns of Hilbert envelop (HE) and residual phase (RP) features from identification results of *Set-1* database.

levels. The segmental level features namely, pitch and energy seem to be more speaker-specific. The recognition performance of the information present in the suprasegmental level is poor compared to the other levels. The suprasegmental level information may have large intra-speaker variability and also due to the text-independence. In Chapter 6, we have made an alternative approach for extracting the suprasegmental level speaker information by using pitch, epoch strength and cepstral trajectory vectors. This approach enables us to understand whether poor performance is due to the level of processing or the method employed.

3.6 Summary

In this chapter an unified framework is proposed for the extraction of improved source information by the time domain analysis of the LP residual. Speaker-specific information in the LP residual includes those within one glottal cycle, pitch and energy across two to three glottal cycles, and variation of the pitch and energy across several glottal cycles. In the proposed method, speaker information within one glottal cycle is extracted by the subsegmental processing of the

Table 3.5: Speaker verification performance (in EER) of residual, Hilbert envelop (HE), residual phase (RP) and HE+RP features for whole NIST-03 database. *Sub*, *Seg* and *Supra* represent subsegmental, segmental and suprasegmental LP residual sequence, respectively. $Comb_1$ and $Comb_2$ represent linear score level and logical *OR* combination schemes.

Feature		*Sub*	*Seg*	*Supra*	Src_2		*MFCC*	$Src_2 + MFCC$	
					$Comb_1$	$Comb_2$		$Comb_1$	$Comb_2$
Residual		41.01	26.96	44.49	32.25	21.22	22.94	27.78	17.43
HE		45.52	32.92	45.66	36.27	22.31		26.92	21.01
RP		41.73	26.83	45.84	31.39	22.13		20.01	20.46
HE+RP	$Comb_1$	43.90	27.19	44.94	33.28	20.41		22.99	16.67
	$Comb_2$	30.12	21.36	32.83					

LP residual. The pitch and energy information is extracted by the segmental processing of the LP residual. To model the speaker information effectively using GMM, the segmental and suprasegmental level information are decimated by a factor of 4 and 50, respectively. Experimental results show that subsegmental, segmental and suprasegmental levels contain speaker information. Combining the evidences from each level, the performance improvement indicates the different nature of speaker information at each level. In direct processing of the LP residual the effect of the amplitude dominate the sequence information. To minimize this, the amplitude and sequence information is captured independently using the analytic signal representation of the LP residual. The combination of amplitude and sequence information seem to be a better choice. At the individual level, information provided by segmental level of the LP residual is most effective compared to the other two levels. The information provided at the suprasegmental level processing of the LP residual is poor due to intra-speaker variability and text-independence.

In this chapter the excitation information is extracted by processing the LP residual in the time domain without any explicit extraction of speaker-specific parameters. An alternative is to develop methods to extract parameters at subsegmental, segmental and suprasegmental levels. The following three chapters will be dedicated for that. In the next chapter method is explored for explicit extraction of speaker-specific parameters at the subsegmental level for modelling.

4

Explicit Subsegmental Processing of LP Residual for Speaker Information

Contents

In this chapter, a method is proposed for explicit modeling of the speaker-specific subsegmental information from LP residual. The significance of the proposed approach is demonstrated by a comparative study with the corresponding implicit approach. For explicit modeling, the standard Liljencrants-Fant (LF) parameters that model the glottal flow derivative (GFD) are used. A simplified technique is proposed for approximate estimation of the LF parameters from the LP residual blocks. These blocks are identified by locating the glottal closing and opening instants. The GCIs are located by using the modified *zero-frequency filtering* method. The glottal opening instants (GOIs) are computed as the fixed fraction of the close-phase intervals [75]. Initially, the LF parameters are computed approximately from the assumptions that, the instants of the first zero-crossing and the slope of the return phase are 50% and 10% of the GFD cycle interval, respectively. Then, these parameters are optimized by using the constraint that the flow return to zero after the end of each glottal cycle. This constraint forces all the parameters adjustment to be made concurrently. The proposed approach significantly reduces the computation needed to implement the LF model. The static and dynamic values of the LF parameters are used as features for explicit modeling of the subsegmental information. For implicit modeling, the LP residual is processed in the time domain as described in Chapter 3, Section 3.3.1. The results from different speaker recognition studies show that, in case of speaker identification, the implicit modeling provides significantly better performance compared to explicit modeling. Alternatively, the explicit modeling seem to be providing better performance in case of speaker verification. This indicates that explicit modeling seem to have relatively less intra and inter-speaker variability. The implicit modeling on the other hand has more intra and inter-speaker variability. What is desirable is less intra and more inter-speaker variability. Therefore, for speaker verification task explicit modeling may be used and for speaker identification task implicit modeling may be used. However, for both speaker identification and verification tasks the explicit modeling is found to be providing relatively more different information to other levels of excitation and vocal tract features. The contribution of the explicit features is relatively more robust against noise. Thus, we suggest that the proposed explicit approach can be used to model the subsegmental level information

for speaker recognition.

4.1 Introduction

Explicit approach of modeling provides compact representation of the excitation information. Independent modeling of subsegmental, segmental and suprasegmental information by explicit approach and combining them may provide better performance and also may compensate for the lossy representation. Due to the modeling at multiple levels, computationally simplified methods should be available. There is no simple method for explicit modeling of the subsegmental level information. For example, the GFD estimation and then the computation of the LF parameters for explicit modeling of the subsegmental level information is one of the popular approaches followed in the literature [15]. This involves closed-phase covariance analysis and solution of non-linear equations [15]. The closed-phase covariance analysis approach requires accurate detection of the closed phase which is difficult [52, 65]. The solution of the non-linear equations through iterative algorithms is computationally more intensive [49, 76]. A more effective and simplified method needs to be developed for explicit modeling of the subsegmental level information and then its usefulness can be verified from the corresponding implicit approach for the subsegmental source information.

Speaker-specific information present at the subsegmental level of the excitation signal can be modeled by characterizing the glottal flow activities. The modulated air flow through the vocal folds due to their vibration is termed as glottal flow [15, 71]. The area between the vocal folds is called as the *glottis* and hence the name glottal flow [49]. Due to differences in the manner, speed and change in the rate of vocal folds vibration, the nature of the glottal flow varies from speaker to speaker. In some speakers, vocal folds never close completely (soft voice) and in other cases vocal folds close completely and rapidly (hard voice) [15]. Similarly, duration of opening and closing of vocal folds, the instants of glottal opening, and closing and the shape of the glottal flow also vary from speaker to speaker. Thus, the nature of the glottal flow contains speaker information and can be modeled by characterizing the glottal flow. Since, this information corresponds to one pitch period or glottal cycle, it may be viewed as subsegmental

excitation information.

The way of characterizing the glottal flow is to measure the volume velocity of air flow through the glottis [15]. From the volume/pressure relationship, what we observe in the speech signal is the glottal air pressure that represents its derivative. Figure 4.1 shows the electro glotto gram (EGG) waveforms and their respective derivative of two male speakers, *MS-1* and *MS-2*, collected from arctic database [77]. The text of the speech signal remains same for fair comparison. EGG represents the actual glottal waveform that is produced by closing and opening of the vocal folds during the speech production. It can be observed that the EGG waveforms are different across speakers. The shape and duration of the waveforms are different across speakers. For example, the amplitude of the peaks in the glottal cycles in case of speaker *MS-2* are larger than *MS-1*. This shows that the amount of air flow in producing the speech for speaker *MS-2* is relatively more. Further, the duration of the glottal cycles in case of speaker *MS-1* is more than *MS-2*. As a result, the opening, closing and location of the peaks in each glottal cycle are different across speakers. This can also be clearly observed from the respective derivatives of the EGG waveforms. The first negative zero-crossing in each cycle of the EGG derivative represents the instants of maximum air flow and the location of the sharp peaks represents the instants of rapid closing of the vocal folds. They are different across the speakers. Thus, GFD can also be used for explicit modeling of the subsegmental excitation information. For this, method for accurate estimation of the GFD from the speech signal should be available. This is because, the EGG may not be available in many applications, since only speech is recorded.

As mentioned in the introduction chapter, due to dynamic nature of the excitation, it is difficult to obtain a precise measurement of the GFD [21]. Therefore, analytical methods have been employed to model the GFD. Several methods have been proposed in the literature for analytic modeling of the GFD [78]. A comparative study on different methods for parameterizations of the GFD has shown that, (in terms of smallest error) the LF model performs best [78]. The LF model has also been successfully used for speech synthesis and speaker recognition tasks [15,79]. This shows that the LF model and its behavior are well known and hence

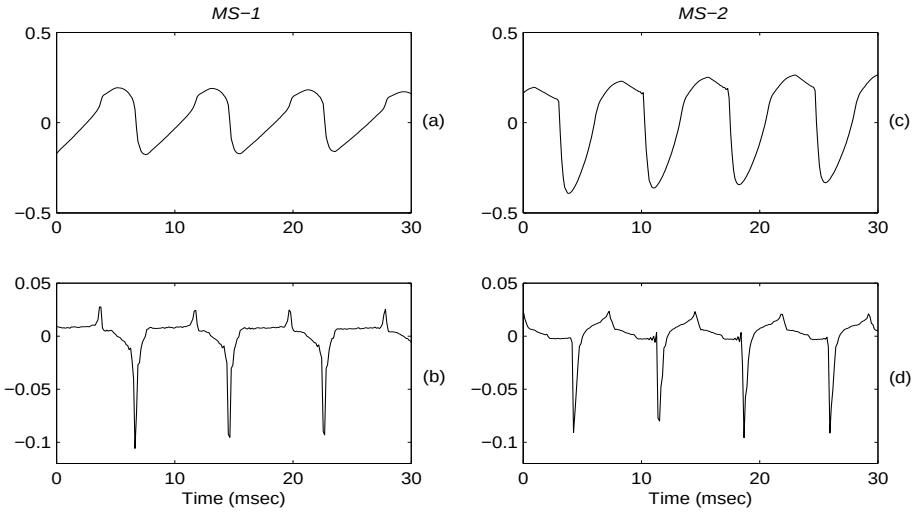

Figure 4.1: Examples of EGG and its derivative. (a), (c) EGG waveforms and their respective derivatives in (b) and (d) for speakers *MS-1* and *MS-2*, respectively.

may be useful for characterizing the GFD for explicit modeling of the subsegmental excitation information. In [15], it was shown that the parameterizations of the GFD by LF parameters contain significant speaker information and also helps *MFCC* features based speaker recognition system to further improve the performance. Thus, we prefer to use the LF model of the GFD cycle for modeling the subsegmental excitation information [51]. Unfortunately, the computational complexity involved in the LF parameters limits its use for modeling the glottal flow [49]. It is expected that if a simplified algorithm is available to compute the LF parameters, then subsegmental excitation information can be modeled explicitly from the LF parameters with reduced computational complexity.

In this work, a simplified and approximate estimation method is proposed to compute the LF parameters from the LP residual and used it to model the subsegmental excitation information. In the proposed approach, the LF parameters are computed from blocks of LP residual samples of one glottal cycle duration. The blocks of LP residual samples are identified by locating the glottal closure and opening instants. The modified *zero-frequency filtering approach* is used to locate the GCIs. Glottal opening instants (GOIs) are obtained from GCIs as the fixed fraction of the close-phase intervals as proposed in [75]. First, the LF parameters computation is made with an assumption that the instants of the first zero crossing and the slope of the return phase

are 50% and 10% of the GFD cycle interval, respectively. This assumption is made based on the concept that glottal closing is faster than its opening activities. The initial estimation of the LF parameters are optimized by using the constraint that the flow return to zero at the end of each glottal cycle. This constraint forces all parameters adjustment to be made concurrently [80]. Then, the subsegmental level information is modeled explicitly by using the LF parameters and their dynamics. The effectiveness of the proposed approach is demonstrated by conducting different speaker recognition experiments and a comparative study with the corresponding implicit approach. The method described in Chapter 3 is used for implicit modeling of the subsegmental excitation information. The comparison is made based on the compact representation, computational complexity involved, recognition performance and their usefulness in providing the additional information to other levels of excitation and the vocal tract information for speaker recognition. Since, the LF parameters are computed directly from the LP residual, the approach is also a temporal processing one, but gives explicit modeling of subsegmental excitation information, as opposed to the implicit modeling in Chapter 3.

The rest of this chapter is organized as follows: Section 4.2 describes the nature of the glottal flow and its derivative. Section 4.3 gives a brief review of the LF model of the GFD. Section 4.4 describes the proposed approach for the computation of LF parameters from the LP residual. In Section 4.5, the speaker-specific features are extracted using the LF parameters for explicit modeling of the subsegmental excitation information. In this section different speaker recognition studies are made to demonstrate the significance of the proposed approach in modeling the subsegmental excitation information explicitly. In Section 4.6, a comparative study is made on explicit and implicit modeling approaches of subsegmental excitation information. In this section the potential of the proposed explicit modeling approach is also compared with other levels of source and vocal tract related information and finally a combined feature is proposed for the complete representation of the excitation information. The last section summarizes the present work presented in this chapter.

4.2 Glottal Flow Derivative

The glottal flow acts as the major excitation to the vocal tract for the production of speech. In voiced speech, the glottal flow is periodic and one period of the glottal flow is called as the glottal pulse. As mentioned in the introduction section, the subsegmental excitation information is modeled explicitly from the GFD. The relation between the glottal flow and its derivative for an ideal case is shown in Figure 4.2 (a) and (b), respectively. The glottal flow cycle is

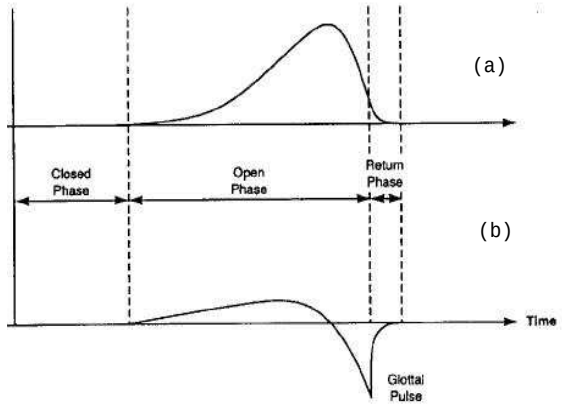

Figure 4.2: Relation between glottal flow and its derivative [15]. (a) Glottal flow. (b) GFD.

divided into three phases, called as, *closed-phase*, *open-phase* and *return-phase*. The interval during which the folds are closed and no flow occurs is called as the *closed-phase*. Practically, the vocal folds may not fully close and some air flow may always be present [15]. This may be due to very quick and incomplete closure of the vocal folds [52]. In this phase the air flow is nearly constant and the folds are loosely coupled to the vocal tract. Due to differentiation, the constant air flow has negligible effect on the GFD and may be assumed to be zero [49]. Thus, the closed-phase of the GFD may not provide any useful speaker-specific information.

The interval during which the vocal folds are open and there is an air flow through glottis is called as the *open-phase*. When the folds start to open, the interaction between the vocal folds and the vocal tract increases until a constant flow has been achieved. The variation in the shape of the glottal flow is therefore mostly due to the manner in which the glottis changes and the

loading by the vocal tract [15]. Thus, variations in the GFD during the *open-phase* may contain significant speaker information. This may be attributed to the rate of increase of the flow, the maximum flow and the maximum rate of decrease of the flow and their corresponding instants. This can be easily observed from the GFD. For example, the shape of the GFD represents the rate of increase or decrease of the air flow, the zero-crossing represents the instants of the maximum air flow and its negative peak represents the maximum rate of decrease of the air flow. Since, the physical structure of the vocal folds differs from speaker to speaker, it is expected that the nature of the GFD cycle during *open-phase* may contribute to significant speaker information.

The interval from the instant of the maximum rate of decrease of the air flow to the instant of the zero air flow is called as the *return-phase*. This phase is particulary important, as it determines the amount of high frequency energy present in the glottal cycle [15]. The more rapidly the vocal folds close, the shorter the duration of the *return-phase* resulting in more high frequency energy. This can be observed from the exponential nature of the GFD during the return phase. The air flow in the *return-phase* is generally considered to be of perceptual importance, because it determines the spectral slope [49]. Thus, the nature of the glottal cycle during the *return-phase* is also expected to contribute significant speaker information.

4.3 LF Model of Glottal Flow Derivative

The LF model of the glottal flow describes the GFD waveform in terms of exponentially growing sinusoid in the *open-phase* and a decaying exponential in the *return-phase* [15, 21, 51]. The shape of the model waveform is controlled by a set of analysis and synthesis parameters. The parameters derived from the inverse filtering of the speech signal is called as the analysis parameters. These include, the time of glottal opening and closing (T_o, T_e), the maximum rate of the flow decrease and its location (E_e,T_e). The parameters derived from the complex relationship of the GFD model are called as the synthesis parameters. These include, growth factor (α), flow decrease curvature and its nature (w_z, β). The description of the seven LF parameters are given in Table 4.1, [15].

Table 4.1: Description of the seven parameters of the LF model of the GFD [15, 49, 80].

T_o	The time of glottal opening.
T_c	Time of glottal closure.
E_e	Absolute value of the maximum rate of glottal flow decrease.
T_e	The time of maximum rate of glottal flow decrease.
α	The growth factor defined as the ratio of E_e to maximum rate of glottal flow increase.
w_z	Frequency that determines flow derivative curvature to the left of the first GFD zero crossing (T_z), $w_z = \frac{\pi}{T_z - T_o}$ [15, 51].
β	Exponential time constant that determines how quickly glottal flow derivative returns to zero after time T_e.

The mathematical expression for closed, open and return phases of one cycle of the GFD, $e_{LF}(t)$, from the synthetic LF model is given by the following equation [15, 80].

$$\begin{aligned} e_{LF}(t) &= 0, && 0 \leq t < T_o \\ &= E_o e^{\alpha(t-T_o)} \sin[w_z(t-T_o)], && T_o \leq t < T_e \\ &= -\frac{E_e}{\beta T_a}[e^{-\beta(t-T_e)} - e^{-\beta(T_c-T_e)}], && T_e \leq t < T_c \end{aligned} \tag{4.1}$$

In Eqn. 4.1, E_o is an arbitrary gain constant and T_a (time constant of the return phase) is the the projection of the modeled GFD at $t = T_e$ in the time axis [49, 80]. In the LF model of the GFD, the closed phase is assumed to be zero. This will not affect in modeling the GFD. Because, as mentioned earlier, the small air flow during closed phase have less effect on the GFD. A synthetic glottal flow and its derivative modeled by the LF parameters are shown in Figure 4.3 (a) and (b), respectively. It can be observed from this figure that, T_o, T_e, E_e, α, w_z and E_0 characterize the *open-phase* and E_e, β, and T_c characterize the *return-phase*. It should be noted here that the LF model also includes the parameter T_a that determines spectral tilt which is perceptually important [49]. All these parameters needs to be supplied explicitly for modeling the GFD. Some parameters like T_o, T_c, T_e and E_e can be obtained from the LP residual by locating the glottal opening and closing instants using recently proposed event based approach of fundamental frequency estimation [62]. However, the computation of the

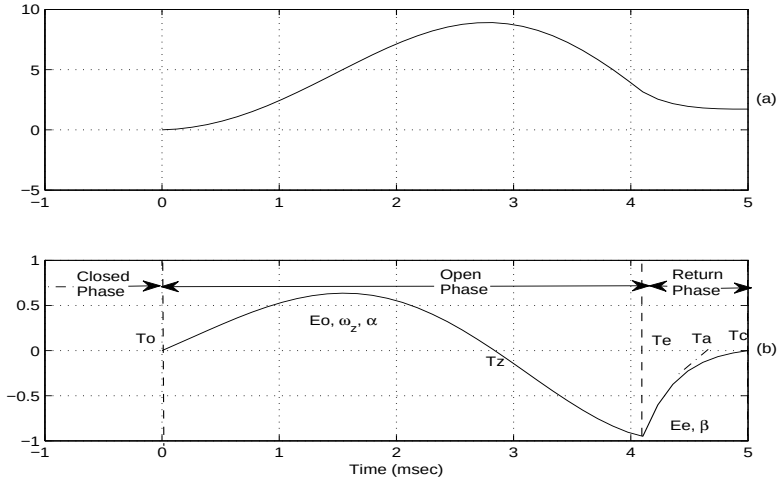

Figure 4.3: Typical glottal flow and GFD waveforms with the parameters of the LF model. (a) glottal flow. (b) GFD of the flow in (a). The GFD is modeled by Seven LF parameters. Three of the parameters (E_0, w_z, and α) describe the shape of the glottal flow during nonzero flow (T_o-T_e). The two parameters (E_e, β) describe the shape of the glottal flow during most negative glottal flow derivative and glottal closure (T_e-T_c). The other two parameters (T_e and E_e) describe the time and amplitude of the most negative peak of the glottal flow derivative.

other parameters are computationally tedious due to noise like nature of the LP residual and the complex relationship among the LF parameters. In the next section, a simplified method for the approximate computation of these parameters from the LP residual is described.

4.4 Computation of LF parameters

The LF parameters are computed for each glottal cycle individually. In Section 4.3, we observe that amplitude of the GFD during the closed phase is zero and the parameters used to describe the LF model of the GFD are associated with the open and return phases. Thus, the effective GFD cycle can be assumed to be starting at the instant of the glottal opening and ends at the instant of the glottal closing. By considering every glottal cycle starts at $t = 0$ and ends at $t = T_c$, the mathematical expression of the LF model for an individual GFD cycle

given in Eqn. 4.1 can be modified as,

$$e_{LF}(t) = E_o e^{\alpha(t)} \sin[w_z t], \qquad 0 \leq t < T_e \tag{4.2}$$

$$= -\frac{E_e}{\beta T_a}[e^{-\beta(t-T_e)} - e^{-\beta(T_c-T_e)}], \qquad T_e \leq t < T_c \tag{4.3}$$

Further, at $t = T_e$, $e_{LF}(t) = -E_e$. Thus, we have from Eqn. 4.2

$$E_o e^{\alpha T_e} \sin(\omega_z T_e) = -E_e$$

$$\Rightarrow \alpha = \frac{1}{T_e} \ln[-\frac{E_e}{E_o \sin(\omega_z T_e)}] \tag{4.4}$$

and from Eqn. 4.3

$$-\frac{E_e}{\beta T_a}[1 - e^{-\beta(T_c-T_e)}] = -E_e$$

$$\Rightarrow 1 - e^{-\beta(T_c-T_e)} = \beta T_a \tag{4.5}$$

Eqns. 4.4 and 4.5 indicate that the relationship among the LF parameters have no close form solution and may be solved iteratively using nonlinear least squares algorithms like Gauss-Newton or Newton-Raphson algorithms [15, 76]. But, these algorithms are computationally intensive and are not adequate when the minimum error is large [15]. To avoid this difficulty, adaptive nonlinear least squares regression techniques have been proposed [15]. The difficulty in this approach is that, the estimated parameters may be too close to their bound that leads to physically unrealistic conditions. For example, E_o taking a negative value or a value near to zero. Thus, iterative approach of estimating the parameters from the actual LF model may not seem to be useful. To avoid the risk of unrealistic conditions and also to reduce the computational complexity, we prefer to estimate the parameters from a simplified LF approximation model. In [80], one such model is proposed, where it is assumed that the return flow is relatively faster, for example, $\beta(T_c - T_e) \gg 1$. This assumption helps us in reducing the nonlinear Eqn. 4.5 to a simplified form as given below.

$$\beta(T_c - T_e) \gg 1$$

$$\Rightarrow e^{-\beta(T_c-T_e)} \simeq 0 \tag{4.6}$$

Thus, from Eqn. 4.5 and Eqn. 4.6,

$$\beta T_a = 1 \tag{4.7}$$

In the above assumption a constraint is imposed that the GFD returns to zero at the end of each cycle. Such that,

$$\int_0^t e_{LF}(t)dt = 0 \tag{4.8}$$

The above constraint forces all the parameters adjustment to be made concurrently by providing another form of relation among the parameters. For example, integrating the Eqn. 4.8 by parts (from $t = 0 \rightarrow T_e$ then $t = T_e \rightarrow T_c$) and using Eqn. 4.4, one can easily find out the value of the β as given below.

$$\beta = \frac{E_e(\alpha^2 + \omega_z^2)}{E_o\{e^{\alpha T_e}[\alpha sin(\omega_z T_e) - \omega z cos(\omega_z T_e)] + \omega_z\}} \tag{4.9}$$

The above assumptions considered in the approximate LF model are relevant to the present work in the sense that, it provides a simple method to compute the parameters for characterizing the GFD that is comparable to that produced by the actual LF model [80]. The assumptions taken do not affect the speaker information present in the estimated parameters much. For example, the first assumption considers that $\beta T_a = 1$, so that T_a is relatively less (10% of the glottal cycle duration) compared to T_e and T_c. Since, T_a is the derivative of the glottal flow at minimum of its first derivative, it does not have any apparent physical correspondence with human voicing events [80]. Similarly, in the second assumption the minimum air flow during the end of each glottal cycle may not affect much because of the differentiation.

The parameters like T_o, T_e, T_c, E_e are computed from the LP residual by locating the glottal cycles. The parameters like ωz are E_o are initially approximated and then α and β are computed using Eqns. 4.4 and 4.9. These parameters are modified iteratively until the Eqn. 4.7 is satisfied. To limit the computation, the process is bounded by 10 iterations. The number of iterations is chosen based on the observations from the experimental studies done in this work. It should be noted here that, unlike the earlier approach, this iteration process involves only computation of two parameters and adjust all parameters concurrently. Thus, the

computational complexity and risk of unrealistic conditions are relatively less in the proposed approach.

4.4.1 Computation of T_o, T_e, E_e and T_c

Due to weak excitation, direct estimation of the glottal opening instant is a difficult task [75]. Thus, first we compute the closing instant and then the corresponding opening instant is identified as the fixed fraction of the glottal cycle as suggested in [75]. Each glottal pulse is considered as the segment of the LP residual from respective glottal opening to closing instant. T_e and E_e are calculated by identifying the peak in the glottal pulse. The accuracy in the computation of all these parameters depends upon how accurate the GCIs are estimated. There are several methods that have been proposed in the literature for computing the glottal closure instants and the most recently proposed *zero-frequency filtering* approach is found to be more accurate [62, 63]. The advantage of using this method is that it computes the glottal closure instants directly from the speech signal and does not require finding the closed phase region. A brief description of this method is given below [62, 63].

4.4.1.1 Zero-frequency Filtering Method for Computation of T_c

The *zero-frequency filtering* method locates the GCIs by passing the speech signal through a zero-frequency resonator twice. The zero-frequency resonator is a second order infinite impulse response (IIR) filter located at 0 Hz [81]. The purpose of passing the speech signal twice is to reduce the effects of all (high frequency) resonances [63]. Passing the speech signal twice through a *zero-frequency resonator* is equivalent to four times successive integration. This will result in a filtered output that grows/decays as a polynomial function of time. The trend in the filtered signal is removed by subtracting the local mean computed over an interval corresponding to the average pitch period. The resulting mean subtracted signal is called as zero-frequency filtered signal (ZFFS). The positive zero crossings in the ZFFS correspond to the locations of GCIs [63]. The steps involved in processing the speech signal to derive the ZFFS are given below.

(i) Difference the speech signal $s(n)$

$$x(n) = s(n) - s(n-1) \tag{4.10}$$

(ii) Pass the difference speech signal $x(n)$ twice through zero-frequency resonator

$$y_1(n) = -\sum_{k=1}^{2} a_k y_1(n-k) + x(n) \tag{4.11}$$

and,

$$y_2(n) = -\sum_{k=1}^{2} a_k y_2(n-k) + y_1(n) \tag{4.12}$$

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

(iii) Compute the average pitch period using the autocorrelation over a 20 msec speech segment

(iv) Remove the trend in $y_2(n)$ by subtracting the mean computed over average pitch period. The resulting signal

$$y(n) = y_2(n) - \frac{1}{2N_a + 1} \sum_{m=-N_a}^{N_a} y_2(n+m) \tag{4.13}$$

is the ZFFS. Here, $2N_a + 1$ corresponds to the number of samples in the window used for mean substraction.

4.4.1.2 ZFFS of telephone speech

The zero-frequency resonator filters out a monocomponent centered around the zero frequency from the speech signal. However, in case of the telephonic speech, the frequency components below 300 Hz are heavily damped. The output of the zero-frequency resonator obtained from processing the telephonic speech may not therefore give correct estimation of the pitch and epoch strength values. To avoid this difficulty, we propose to use the HE of the LP residual of telephone speech for zero-frequency filtering to compute the pitch and epoch strength values. HE is defined as the magnitude of complex time function of the LP residual [82]. Due to impulse-like nature of the LP residual, the information about the fundamental frequency will spread across all the frequencies including the zero frequency. The purpose of using the HE is

to emphasize the peaks around the GCIs in each glottal cycle and hence the reinforcement of energy around the impulse at zero frequency [71, 82].

To verify the effectiveness of the proposed approach, we compute the epochs from a telephonic speech and compare them with the estimated epochs from the ZFFS of the corresponding clean speech. For this, we collect the speech data of a speaker from TIMIT and NTIMIT databases [83, 84]. For both the cases the text of the speech remains same. The speech data collected from TIMIT database represents the clean speech and from the NTIMIT database represents the corresponding telephonic speech. Figures 4.4(a) and (b) show a segment of clean speech and the corresponding ZFFS derived from the clean speech, respectively. The arrows in the ZFFS indicate the location of the positive zero-crossings. It can be observed that the instants of the positive zero-crossings in the ZFFS clearly indicate the location of the epochs. Figures 4.4 (c), (d) and (e) show the segment of telephonic speech of the same text as in case of clean speech, HE of the LP residual of the telephonic speech and ZFFS derived from the HE of the LP residual of the telephonic speech, respectively. It can be observed that the time instants of the zero-crossings indicated by arrows in the ZFFS correspond to the original epochs shown in Figure 4.4(b). From this observation we may conclude that in case of telephone speech, the ZFFS derived from the HE of the LP residual can be used to compute the pitch and epoch strength values.

4.4.1.3 Computation of T_o, T_e and E_e

Once the GCIs are computed, then the GOIs are computed as the following GCIs plus a fixed duration of the larynx cycle [75]. The larynx cycle is considered as the minimum of difference between preceding and following GCI and the average pitch period. Closed-phase intervals have been reported to be 30% to 45% of the larynx cycle for normal speech. In this work the opening instant of a glottal cycle is computed as the closing instant of the just preceding glottal cycle plus 30% (minimum range) of the larynx cycle [75]. It should be noted here that, since in the computation of the opening instant, both preceding and following glottal cycles closing instants are involved, the first and the last glottal cycles are ignored. Further, the

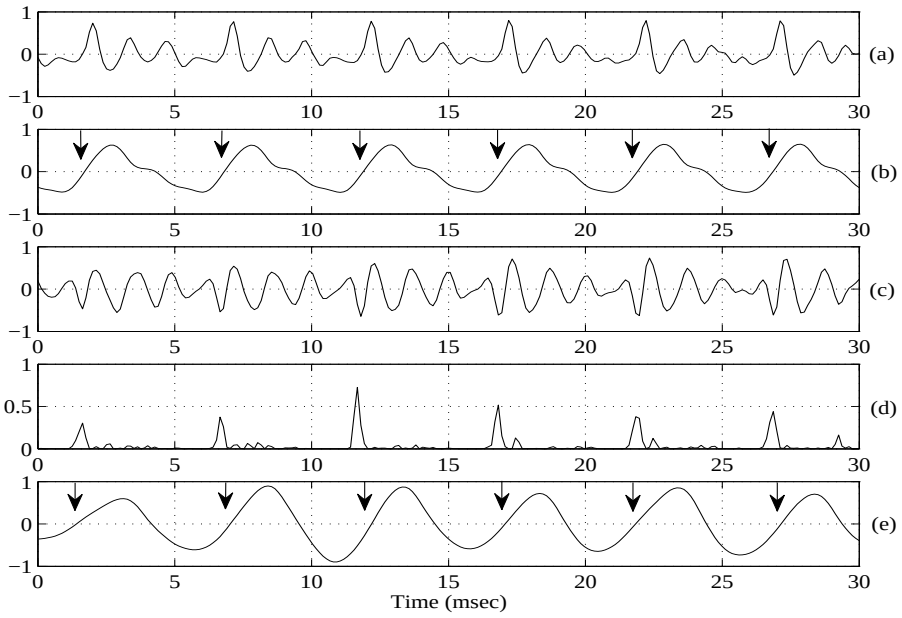

Figure 4.4: Estimation of pitch period from clean and telephonic speech signal. (a) Clean speech. (b) *Zero-frequency filtered signal* (ZFFS) derived from the speech signal in (a). (c) Speech signal of the same text as in (a) collected over telephone channel. (d) The Hilbert envelop (HE) of the LP residual of the speech signal in (c). (e) ZFFS derived from the signal in (d). The location of the positive zero-crossings in the filtered signal (b) and (e) are shown by arrows.

glottal cycles duration may be beyond the practical range, that is of duration 5 to 20 msec [21]. This is because the successive glottal cycles may not be continuous. We consider only those glottal cycles that have 5 to 20 msec duration. The detailed procedure to compute the GOIs from GCIs are given below [75].

(i) Compute P_g, that is the pitch period of the g^{th} glottal cycle as, $P_g = T_{c(g)} - T_{c(g-1)}$, where, $T_{c(g)}$ and $T_{c(g-1)}$ are the closing instants of the g^{th} and its just previous glottal cycle.

(ii) Compute the average pitch period $\widehat{P}$, that is the maximum of seventh order median filtering of P_g [75].

(iii) Compute $T_{o(g+1)}$, that is the opening instant of the $(g+1)^{th}$ glottal cycle using Eqn. 4.14. The opening instants are considered as the 30% of the larynx cycles [75].

$$T_{o(g+1)} = T_{cg} + 0.3 \times \min[T_{c(g+1)} - T_{c(g)}, \widehat{P}] \tag{4.14}$$

(iv) Consider only corresponding opening and closing instants those have the difference within 5 to 20 msec range.

A segment of voiced speech, ZFFS and the LP residual are shown in Figure 4.5 (a), (b) and (c), respectively. The positive zero-crossings in the ZFFS indicate the GCIs. The corresponding instants in the LP residual marked as 'x' are considered as the end points of the LP residual blocks. The computed GOIs marked as 'o' are considered as the starting points of the LP residual blocks. The locations and peaks of the LP residual blocks indicate their corresponding T_e and E_e, respectively. It should be noted here that, the LP residual is an error signal results from the prediction of the speech signal [24]. The peaks in the LP residual corresponds to large error which may either be positive or negative. Thus, in finding the T_e and E_e, we prefer to consider the absolute peaks in the LP residual blocks.

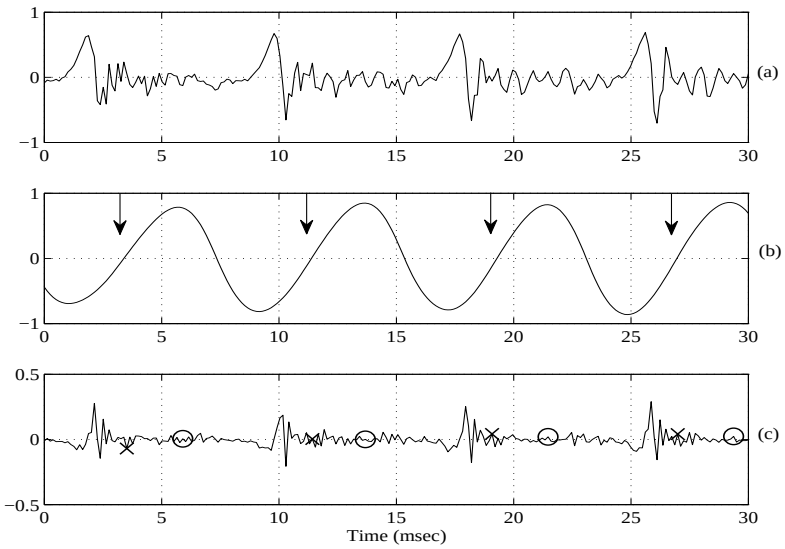

Figure 4.5: Estimation of the glottal cycles from the LP residual of the speech signal. (a) Speech signal. (b) *Zero-frequency filtered signal* derived from the Hilbert envelop (HE) of the LP residual shown in (c). The location of the positive zero-crossings in the filtered signal (b) are shown by arrows. The 'xs' and 'os' in the LP residual (c) represent glottal closing and opening instants, respectively.

4.4.2 Computation of T_z, E_o, α and β

Once the values of T_e and E_e are computed as described in Section 4.4.1, one can observe from the Eqns. 4.4 and 4.9 that for computation of the α and β, the only unknown parameters

left are T_z and E_o. Since the LP residual is an error signal it is difficult to find the first zero-crossings T_z and then in turn the E_o also, accurately. Thus, we propose to estimate these parameters through an iterative process that involves initial approximation of the parameters followed by modifications. The detailed procedure of the proposed approach is described below.

The parameters T_z and E_o are associated with the open-phase of the glottal cycle which is generally larger than the return phase. Thus, initially we assume that T_z is 50% of the glottal cycle duration. With this assumption, E_o is measured as the absolute maximum of the glottal cycle up to T_z. The reason for using the absolute maximum value is as mentioned earlier. Now, one can easily compute the value of α and β from Eqns. 4.4 and 4.9, respectively. By observing these equations one can found that the parameter β depends upon α and ω_z. Thus to verify the accuracy of the initial estimation of the parameters, we prefer to use the constraint imposed by Eqn.4.7. Any modification in the value of T_z will concurrently adjust all the remaining parameters. In every modification, the T_z value is increased by 5% of the glottal cycle. The reason for increasing the T_z value is due to the larger duration of the open phase. The steps involved for computing the LF parameters for each glottal cycle are summarized below.

(i) Initially, assume that $T_z = 0.5T_c$

(ii) Compute, $w_z = \frac{\pi}{T_z}$

(iii) Compute, E_o as the maximum absolute of the sample values up to T_z

(iv) Compute, α using Eqn. 4.4

(v) Compute, β using Eqn. 4.9

(vi) Compute, the value of βT_a

(vii) If $\beta T_a \neq 1$, then replace $T_z = T_z + 0.05T_c$ and repeat from step 1

(viii) Continue step 7 until $\beta T_a \simeq 1$ or the number of repetitions is 10.

It should be noted here that an iterative method was previously applied to estimate the LF model parameters in [15, 76]. In this method, the iteration process minimizes the error

between the estimated glottal flow and the LF model. The error in the estimated glottal flow mostly due to the requirement of the closed-phase region estimation which may affect the computational accuracy of the parameters. The proposed approach does not require the estimation of the closed-phase region. This approach also adjusts all parameters concurrently and hence provides global modification in estimating the parameters. Thus, it is expected that the proposed approach may be effective in computation of the LF parameters for explicit modeling of the subsegmental excitation information for speaker recognition.

4.5 Speaker-specific Information from LF Parameters

In this section we use the proposed approach to compute the seven LF parameters described in the Table 4.1 and extract the speaker-specific features for explicit modeling of the subsegmental excitation information. The significance of the extracted features is demonstrated by speaker identification and verification studies. To improve the recognition performance further, the speaker-specific information associated with the temporal dynamic nature of the LF parameters are incorporated by standard delta (Δ) and delta delta ($\Delta\Delta$) measures [3, 4].

4.5.1 Speaker-specific Feature from LF Parameters

The LF parameters are derived from the respective LP residual of the speech signal and hence they represent the GFD characteristics of the individual speaker. The seven LF parameters described in the Table 4.1 may be classified into two groups as: *wave shaping* and *timing* parameters. The *wave shaping* parameters that include ωz, α and β characterize the shape of the GFD. The *timing* parameters that include T_o, T_e and T_c characterize the glottal timing information. The *wave shaping* parameters are directly used for feature representation. The *timing* parameters are first normalized with respect to the length of the glottal cycle and then used for feature representation. The normalization is performed to limit the values of the absolute times. Otherwise, these absolute time values will increase beyond limit with increase in the number of the glottal cycles [15].

The discrete domain representation of the four *wave shaping* parameters are given by

$$\Omega_z = \frac{\pi}{N_z} \tag{4.15}$$

$$\alpha = \frac{1}{N_e}\ln[-\frac{E_e}{E_o \sin(\Omega_z N_e)}] \tag{4.16}$$

$$\beta = \frac{E_e(\alpha^2+\Omega_z^2)}{E_o\{e^{\alpha N_e}[\alpha sin(\Omega_z N_e) - \omega z cos(\Omega_z N_e)] + \Omega_z\}} \tag{4.17}$$

where, N_z and Ω_z are sample-time counterparts to their corresponding continuous time variables T_z and ωo, respectively.

The discrete domain representation of the three normalized *timing* parameters are given by [15]

$$\text{Close quotient (CQ)} = \frac{N_o - N_{c-1}}{N_e - N_{e-1}} \tag{4.18}$$

$$\text{Open quotient (OQ)} = \frac{N_e - N_o}{N_e - N_{e-1}} \tag{4.19}$$

$$\text{Return quotient (RQ)} = \frac{N_c - N_e}{N_e - N_{e-1}} \tag{4.20}$$

where, N_o, N_e and N_c are sample-time counterparts to their corresponding continuous time variables T_o, T_e and T_c, respectively.

The speaker-specific nature of the *wave shaping parameters* and the normalized *timing parameters* is demonstrated by comparing their histograms from two different female speakers (*FS-1* and *FS-2*) shown in Figure 4.6. In these figures the parameters are computed for two examples of different texts per speaker are shown to demonstrate the inter and intra variation nature of the LF parameters. The histograms are obtained from about 30 sec of data of each speaker. The text of the speech per example and the gender of the speakers remain same for fair comparison. It can be observed from the histograms that the LF parameters are less variant within speaker and more variant across speakers. In general, the *wave shaping parameters* are less speaker-specific than the *timing parameters*. In particular, the distribution of the (OQ) and (RQ) values are significantly different. This may be due to the perceptual importance of the open and return phase that determines the spectral tilt [49, 78, 80].

The *wave shaping* and *timing* parameters computed from a glottal cycle are concatenated to

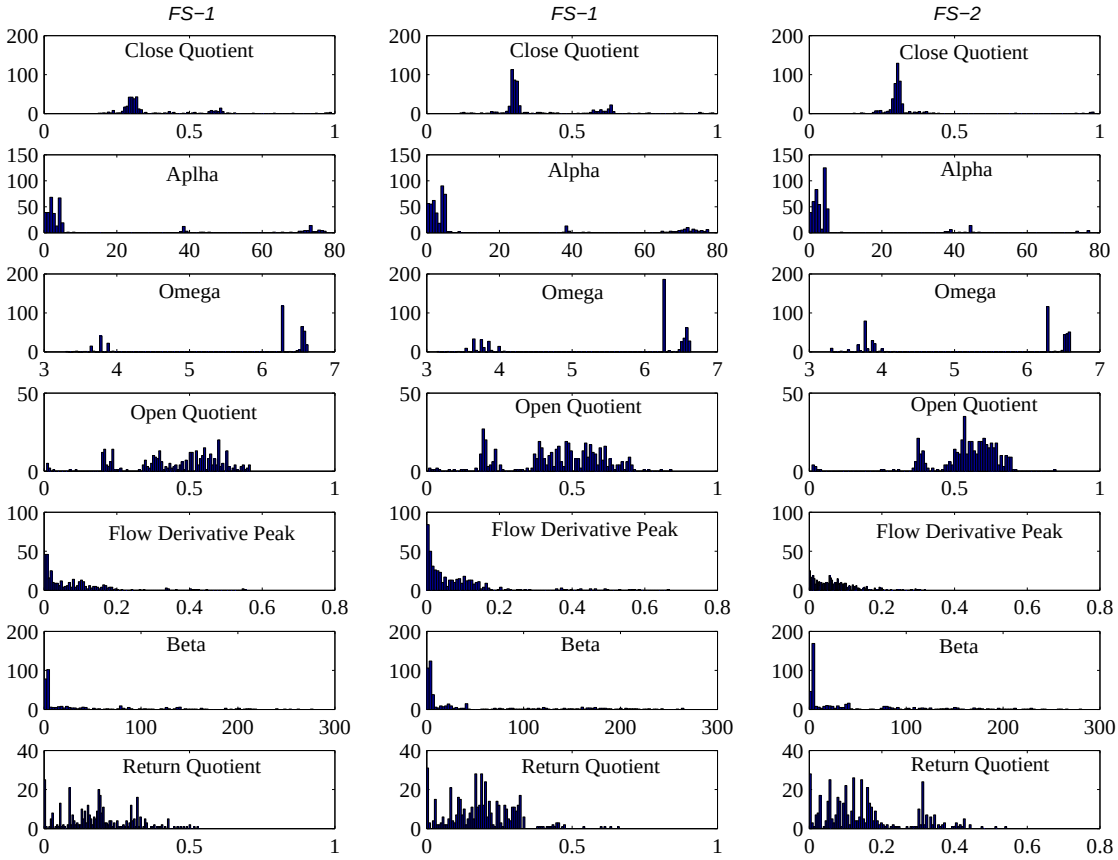

Figure 4.6: Comparison of the histograms of seven components of the glottal flow derivative (GFD) feature for two female speakers (*FS-1* and *FS-2*). The feature component values are divided across 100 histogram bins. The first two columns represent two examples of the speaker *FS-1*.

form the feature vector. Further, the value of the *wave shaping* parameters are varying across the different range. For example, the timing parameters are within the range of 0 to 1, where as, the value of α, β are in the range of the 0-20 and the value of the Ω_z in the range of 3.5-7. Therefore, to avoid the large variations, the computed parameters need to be normalized to a common range. In this work we normalize these parameters by zero-mean unit variance and then used to represent the features for building the speaker models and later for testing.

4.5.2 Speaker-specific Feature from Dynamics of LF Parameters

The fine variations of the GFD like aspiration and ripple also provide useful speaker information [15]. This information was captured through formant modulation in the closed phase region and used for speaker identification study [15]. It was shown that, the recognition performance of the features associated with the fine variations of the glottal flow is relatively poor but provide additional information to the *GFD* feature. Motivated by this we try to use this information in our study. However, due to the difficulty in finding the closed phase region, we exploit the dynamic nature of the LF parameters to capture the fine variations. The hypothesis is that, the variation in the LF parameters from one glottal cycle to the other may be attributed to the fine variations in the glottal cycles. Figure 4.7 shows the contours of the LF parameters obtained from same segment of the speech for *FS-1* and *FS-2*. Due to difference in the pitch, the number of LF parameters obtained are different across speakers. Therefore, cycle-to-cycle comparison of the LF parameters may not be fair. The variation in the parameters value from one cycle to the next cycle is different across speakers. These differences may be due to the different speaking style that affects the glottal cycles.

In this work, the speaker-specific information associated with the dynamics of the LF parameters are represented by standard (Δ) and ($\Delta\Delta$) measures [3, 4]. The commonly used equations for Δ and $\Delta\Delta$ measures are given by [3, 85]

$$\Delta X^n(m) = \frac{\sum_{i=-2}^{2} iX^n(m)}{\sum_{i=-2}^{2} |i|} \tag{4.21}$$

$$\Delta\Delta X^n(m) = \frac{\sum_{i=-2}^{2} i\Delta X^n(m)}{\sum_{i=-2}^{2} |i|} \tag{4.22}$$

where $X^n(m)$ represents the m^{th} component of the feature X for the n^{th} frame [7, 85]. The Δ and $\Delta\Delta$ measures gives the rate of change of the LF parameters values from cycle to cycle. Since, the LF parameters are unique, it is expected that their Δ and $\Delta\Delta$ measures may also

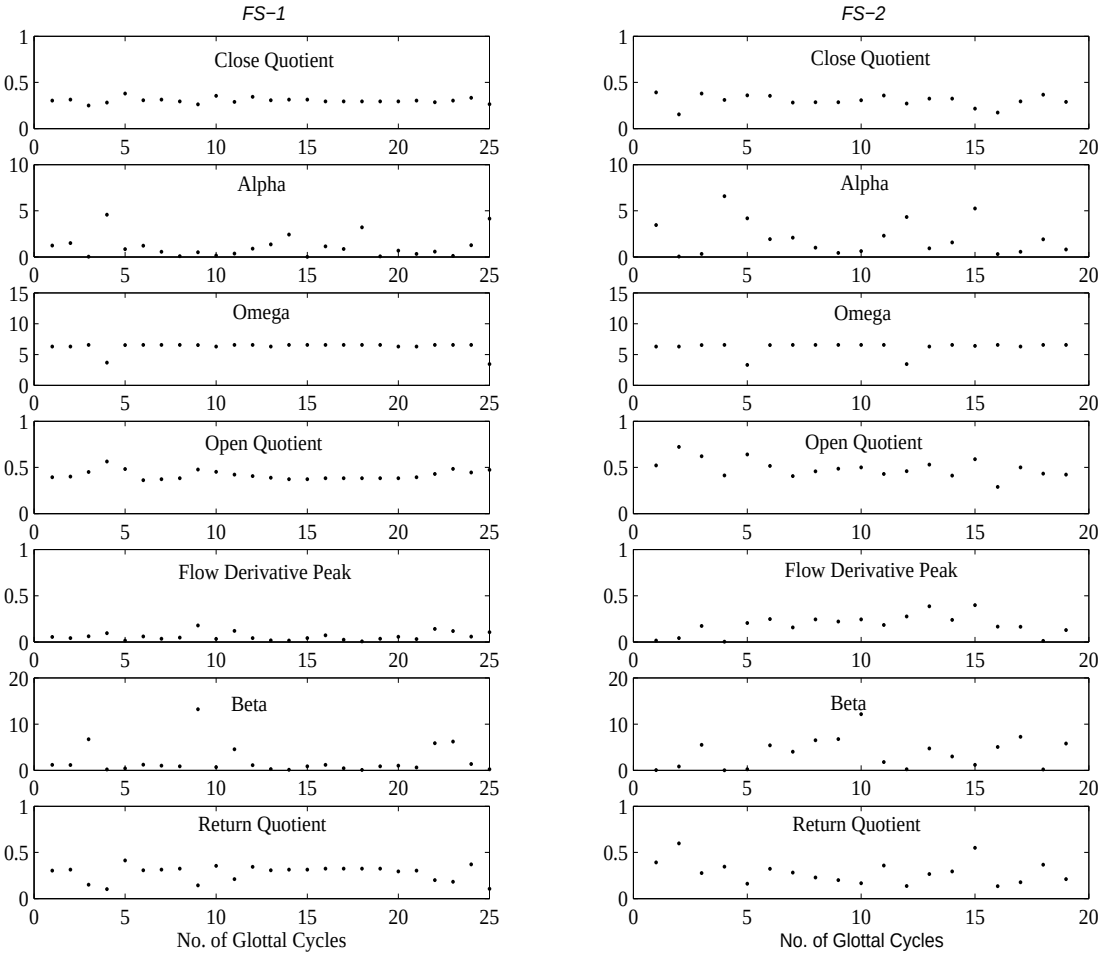

Figure 4.7: Example of the contours of seven components of the glottal flow derivative (GFD) feature from 0.5 sec duration of speech for two female speakers (*FS-1* and *FS-2*).

contribute speaker information.

4.5.3 Speaker Recognition Study using LF Parameter Information

To evaluate the effectiveness of the features extracted from the proposed LF parameters estimation approach, we consider the *Divergence* measure and the recognition performance. *Divergence* measure which is a generalization concept of the F-ratio to the multidimensional case [23, 86–88]. *Divergence* is a statistical measure of effectiveness of a set of distributions for discriminating between categories [23]. It is measured from two covariance matrices: *Between class* and *Within class* covariance matrices [87]. For our application, the *Divergence* of a feature

as a measure of speaker discriminating ability can be measured from intra-speaker (W) and inter-speaker (B) covariance matrices [23,88]. W represents the variation of the feature vectors within an individual speaker and B represents the variation of the feature vectors across different speakers. A feature is said to be more discriminating, if it is less variant within an individual speaker and more variant across different speakers [2]. The scalar measure of this property is the trace of the matrix $W^{-1}B$ called as *Divergence* [23, 88]. The trace of a matrix is the sum of its diagonal elements. Let us consider a distribution χ consists of set of feature vectors of S speakers. Each s^{th} speaker, where $s = 1, 2, ..., S$, in the distribution has n_s number of $d-$ dimensional feature vectors. If x_{ψ_s} is the ψ_{th} feature vector of the s^{th} speaker then its mean feature vector is given by

$$m_s = \frac{1}{n_s} \sum_{\psi_s=1}^{n_s} x_{\psi_s} \tag{4.23}$$

There is a total of $N_S = \sum_{s=1}^{S} n_s$ number of such mean feature vectors in the distribution χ. The total mean feature vector of the distribution χ is given by [87, 88],

$$m = \frac{1}{N_S} \sum_{s=1}^{S} n_s m_s \tag{4.24}$$

then the covariance matrices W and B are given by [88],

$$W = \sum_{s=1}^{S} \sum_{\psi_s=1}^{n_s} (x_{\psi_s} - m_s)(x_{\psi_s} - m_s)' \tag{4.25}$$

$$B = \sum_{s=1}^{S} n_s (m_s - m)(m_s - m)' \tag{4.26}$$

A feature with higher *Divergence* value usually provides more discriminatory information. A good feature should have higher *Divergence* and better recognition accuracy. It should be noted here that *Divergence* has no simple relation with recognition performance [8]. *Divergence* takes into account the inter and intra parameter correlations. However, a small correlation between parameters does not imply lack of any relationship between them [8]. The redundancy in the information may degrade the recognition performance. So, a feature having higher divergence value but lower recognition performance shows that it contains redundant information [2,8,88].

Thus, in this work, *divergence* is used as a first assessment measure to verify the quality of a feature without conducting recognition experiments. Final assessment is made based on the recognition performance. Further, if two features are providing similar recognition performance, then the feature having higher *divergence* value may be preferred.

In this work features from the LF parameters are called as GFD features. To incorporate the dynamic information, the Δ and $\Delta\Delta$ values are concatenated with LF parameters and together used as the features. We call them as $GFD + \Delta + \Delta\Delta$ feature. The *Divergence* and the speaker recognition results of GFD and $GFD + \Delta + \Delta\Delta$ are given in the Table 4.2. The *Divergence* values of GFD and $GFD + \Delta + \Delta\Delta$ features indicate that the LF parameters computed by the proposed approach have the speaker discriminating ability. By comparing their respective *Divergence* values it can be observed that the $GFD + \Delta + \Delta\Delta$ features have significantly better speaker discriminating ability. Due to noise, the relative degradation in the discriminating ability of the $GFD + \Delta + \Delta\Delta$ feature is relatively less. For example, in case of GFD feature, the relative degradation in the *Divergence* value is 29.67%, as against 19.89% for $GFD + \Delta + \Delta\Delta$ feature. These observations indicate that the $GFD + \Delta + \Delta\Delta$ features have relatively more speaker discriminating ability and robust against noise.

Table 4.2: Speaker identification (in %) and verification performance (in EER) of GFD and $GFD + \Delta + \Delta\Delta$ features. Div and $Perf$ represent *Divergence* and performance, respectively.

Task	Feature Database	GFD		$GFD + \Delta + \Delta\Delta$	
		Div	$Perf(\%)$	Div	$Perf(\%)$
Identification	*Set-1*	48.74	26	111.87	30
	Set-2	34.28	20	89.58	25
	Relative Degradation	29.67	23	19.89	17
Verification	NIST-03	40.93	42.24	101.57	39.79

For both the data sets, the identification performance of $GFD + \Delta + \Delta\Delta$ feature is relatively better than the GFD feature. For the verification task also, the EER of 39.79% achieved by $GFD + \Delta + \Delta\Delta$ feature is better than the GFD feature of around 42.24%. Further, the identification performance of both GFD and $GFD + \Delta + \Delta\Delta$ features degrades for noisy speech.

The relative degradation in case of $GFD + \Delta + \Delta\Delta$ is 17%, that is less against 24% in case of the GFD feature. This demonstrates the relative robustness of the $GFD + \Delta + \Delta\Delta$ features against noise. From these observations, we propose that the proposed approach of computing the LF parameters together with their dynamics may be the possible way of modeling the subsegmental excitation information explicitly.

4.6 Comparison of Explicit and Implicit Modeling of Subsegmental Excitation Information

In the previous section, a method is proposed to model the subsegmental excitation information by estimating the LF parameters from the LP residual. Since the speaker-specific information is modeled by parameterizing the LP residual, the proposed approach may be viewed as the explicit modeling. On the other hand, in Chapter 3 the subsegmental excitation information is modeled directly by processing the LP residual and hence may be viewed as the implicit approach. In this section a comparative study is made between the explicit and implicit approaches of modeling the subsegmental excitation information for speaker recognition. The comparison is made based on computational complexity, speaker discriminating ability, recognition accuracy and in providing the different evidence to other levels of excitation and vocal tract information.

4.6.1 Nature of Speaker-specific Evidence in Explicit and Implicit Modeling of Subsegmental Information

In implicit modeling, the 5 msec LP residual blocks are directly processed to model the subsegmental excitation information. On the other hand the characteristics of the GFD is parameterized from the LP residual blocks. Figure 4.8 shows the example of GFD cycles and the corresponding LP residual of two Male (*MS-1* and *MS-2*) and one female (*FS-2*) speakers for a common utterance taken from arctic database [77]. The GFD cycles give a clear view of the glottal timing and the shape at different phases as well. These characteristics are parameterized by the *wave shaping* and *timing* LF parameters and hence corresponds to explicit modeling.

Since, these parameters explicitly characterize the glottal flow, it is expected that the proposed approach may provide better recognition accuracy. On the other hand, due to noise-like nature, the glottal timing and shape at different phases are not clearly visible from the LP residual. Similar to glottal waves, the shape of LP residuals is also different across speakers and hence may be speaker-specific. The large peaks approximately correspond to the negative peaks of the respective GFD cycles. The strength and location of these peaks are different across speakers. The variation in the sequence of samples around the epochs mostly represent the glottal activity information. Thus the direct use of 5 msec (40 samples for 8 kHz speech signal) blocks is found to be useful to capture the glottal activity information. It is not clearly understood what part of the subsegmental excitation information is captured in the LP residual blocks and hence corresponds to implicit modeling. The proposed explicit approach of modeling the subsegmental excitation information is relatively more compact and involves less computational complexity. For example, for a given two minute training data, the number of feature vectors available for implicit modeling is around 40000 as against to 10000 for explicit modeling. Computational complexity is less in the sense that, the zero-frequency approach used in the proposed approach for locating the GCIs can also be used for modeling the suprasegmental pitch and epoch strength contours information. It is expected that the proposed explicit approach of modeling the subsegmental excitation information may be more effective.

4.6.2 Discriminating ability of Explicit and Implicit Subsegmental Features

The *Divergence* takes into account the inter and intra variances and hence has no simple relation with recognition performance [8]. For example, higher *Divergence* value may be achieved either with low intra-variation or high inter-variation. In case of verification task for a genuine trail, where one-to-one comparison is made, the low intra-variation than more inter-variation may be preferred. In this case a feature with higher *Divergence* value due to high inter-variation may not be useful. On the other hand, in case of text independent identification task, where one-to-many comparisons are made, the more inter-variation than low intra-variation may be

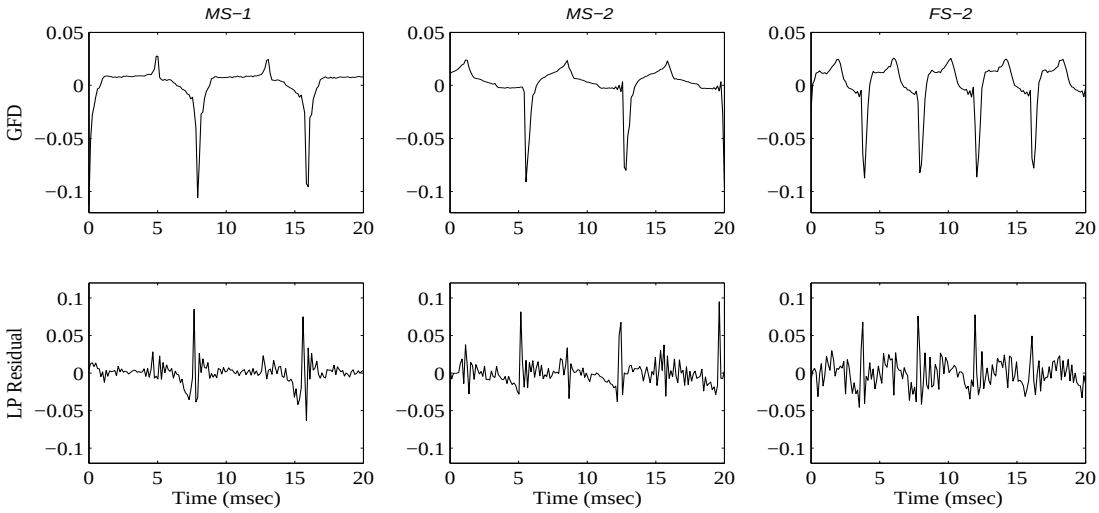

Figure 4.8: Example of GFD and LP residual segments of two males and one female speakers from a common utterance.

preferred. In this case a feature with higher *Divergence* value due to low intra-variation may not be useful. Since, B and W represent the inter and intra covariance matrices, the trace of these matrices can be used as the independent scalar measures (proportional) of the inter and intra variances, respectively. The measurement of the inter and intra variations depends upon the dimension of the feature vectors. The inter and intra variances need to be normalized for fair comparison. In this work, we normalize the inter and intra variations by their *Divergence* value and used as the assessment measure to verify the usefulness of the explicit and implicit subsegmental excitation features.

The *Divergence* value of the *Sub* feature is computed as described in Section 4.5.3, and is found to be 13.03% and 9.75% for *Set-1* and *Set-2* data sets, respectively. By comparing these values with the *Divergence* value of GFD and $GFD + \Delta + \Delta\Delta$ features given in the Table 4.2, it can be observed that the subsegmental explicit features have significantly less discriminating ability. The *Divergence* value of both implicit features is also decreased for more noisy database. The relative degradation in the *Divergence* value from *Set-1* to *Set-2* of *Sub* feature is around 27%, which is more than the GFD and $GFD + \Delta + \Delta\Delta$ features. This shows that the discriminatory information provided by the explicit features is relatively more robust against noise.

As mentioned earlier, the higher *Divergence* value in case of explicit features does not indicate that they will be more effective than *Sub* feature for both speaker identification and verification studies. Instead, they may be effective for a particular recognition task. For example, by comparing the values of respective *intra-variances* from both sets given in Table 4.3, it can be observed that the explicit features have less *intra-variance* as compared to *Sub* features. Thus, the explicit features may be relatively more effective than *Sub* features for speaker verification task. On the other hand, the explicit features have significantly less inter-variance value. Thus, even with higher *Divergence* value but due to less *inter-variance*, the explicit feature may be relatively less effective for speaker identification task. This is in fact we can observe by comparing the speaker identification and verification results of implicit and explicit features from the Tables 4.2, 3.2 and 3.1. In case of identification task, the *Sub* features provide significantly better performance. On the other hand, the GFD and $GFD + \Delta + \Delta\Delta$ features provide the better performance for speaker verification tasks. Thus, it is suggested that to gain maximum benefit from the subsegmental excitation information, the *Sub* and $GFD + \Delta + \Delta\Delta$ features can be used for speaker identification and verification tasks, respectively.

Table 4.3: *Inter-variance* and *Intra-variance* values of *Sub*, GFD and $GFD + \Delta + \Delta\Delta$ features for *Set-1* and *Set-2* data sets.

Parameters	*Set-1*			*Set-2*		
	Sub	GFD	$GFD + \Delta + \Delta\Delta$	*Sub*	GFD	$GFD + \Delta + \Delta\Delta$
Inter-variance	5.68	2.05	1.75	3.68	1.57	1.53
Intra-variance	40.83	0.54	0.26	52.51	0.38	0.27

4.6.3 Complementary Speaker Information from Explicit and Implicit Modeling

So far we observe that, both implicit and explicit features can be used to model the subsegmental level information. The implicit and explicit features seem to be relatively more effective for identification and verification tasks, respectively. However, as mentioned earlier, the excitation information present at subsegmental, segmental and suprasegmental levels are different

from each other and can be combined to improve the recognition performance. The combined evidence from the excitation information is also different from the vocal tract information. The improvement in the performance by combining the evidences mostly depends upon how much they provide different aspect of speaker-specific evidences to each other. To verify the effectiveness of the different information present in the subsegmental implicit and explicit features, we combine the evidence from Sub and $GFD + \Delta + \Delta\Delta$ with segmental, suprasegmental and vocal tract information and a comparison is made.

The speaker identification and verification results are given in Table 4.4. The results show that both Sub and $GFD + \Delta + \Delta\Delta$ are providing different speaker-specific evidences to other levels of the excitation and vocal tract features. For the identification task, due to large inter-speaker variability, the Sub feature is providing relatively more speaker-specific evidence than $GFD + \Delta + \Delta\Delta$ for improved representation of the excitation information. For example, the identification performance of the combined evidence from segmental and suprasegmental levels (Src_3) is 61% and 58% for *Set-1* and *Set-2* datasets, respectively. The benefit we can achieve by combining the evidence from Sub is 74% and 58%, as against 64% and 58% from $GFD + \Delta + \Delta\Delta$ feature for *Set-1* and *Set-2* data sets, respectively. On the other hand, for verification task, due to less intra-speaker variability, the $GFD + \Delta + \Delta\Delta$ is providing relatively more speaker-specific evidence than Sub feature for improved representation of the excitation information. For example, the verification performance of the Src_3 is 33.73%. The benefit we can achieve by combining the evidence from $GFD + \Delta + \Delta\Delta$ is 31.39%, as against 33.28% from Sub feature. However, for both identification and verification tasks the speaker-specific evidence from $GFD + \Delta + \Delta\Delta$ together with the other levels source and vocal tract information provides improved performance. In case of identification task, the benefit we can achieve by $GFD + \Delta + \Delta\Delta + Src_3 + MFCC$ is 89% and 72%, as against 87% and 70% from $Src_2 + MFCC$ for *Set-1* and *Set-2* data sets, respectively. Similarly, for speaker verification task, the benefit we can achieve by $GFD + \Delta + \Delta\Delta + Src_3 + MFCC$ is 22.76% as against 22.99% from $Src_2 + MFCC$. Also, due to noise the degradation in the identification accuracy in case of $GFD + \Delta + \Delta\Delta + Src_3$ is 9%, that is less than Src_2 of around 21%. This shows

that combining evidences from $GFD + \Delta + \Delta\Delta$ with the other levels excitation information is relatively robust against noise.

Table 4.4: Comparison of speaker identification and verification performances of implicit and explicit subsegmental (Sub) features combined with segmental (Seg) and suprasegmental ($Supra$) excitation and vocal tract ($MFCC$) features. Src_3 represents combination of Seg and $Supra$ source information. Src_2 represents combination of Sub, Seg and $Supra$ source information. $Comb_1$ represents linear score level combination schemes.

Feature		$Performance$		
		$Identification(\%)$		$Verification$
		$Set\text{-}1$	$Set\text{-}2$	(EER)
Src_3	$Comb_1$	61	58	33.73
Src_2	$Comb_1$	74	58	33.28
$GFD + \Delta + \Delta\Delta + Src_3$	$Comb_1$	64	58	31.39
$Src_2 + MFCC$	$Comb_1$	87	70	22.99
$GFD + \Delta + \Delta\Delta + Src_3 + MFCC$	$Comb_1$	89	72	22.76
$MFCC$		87	66	22.94

The above observations indicate that the proposed approach of computing the LF parameters for explicit modeling of the subsegmental excitation information is relatively more compact, provide additional information to other levels of the excitation and vocal tract features for speaker recognition and robust against noise. Thus we conclude that the proposed approach may be the best possible way of modelling the subsegmental excitation information explicitly.

4.7 Summary

In this chapter, the speaker-specific information present at the subsegmental level is modeled explicitly by the LF parameters of the GFD. A simplified method is proposed for approximate measurement of the LF parameters from the LP residual. The proposed approach significantly reduces the computational complexity involved in computing the LF parameters. The statistical distribution of these parameter values from different speakers demonstrate their speaker-specific nature. The feature representation of the LF parameters together with their dynamics is found to effective for explicit modeling of the subsegmental excitation information for speaker recognition. A comparative study made with the implicit modeling of the subsegmental excitation

information by direct processing of the LP residual revealed that, due to small intra-speaker variation, explicit approach is useful for speaker verification task. On the other hand, due to large inter-speaker variation, implicit approach is useful for speaker identification task. The proposed explicit feature is more compact, speaker discriminating and provides robust information for speaker recognition. For both identification and verification tasks, the proposed explicit features provide relatively more different and robust information to segmental and suprasegmental excitation and vocal tract features to further improve the recognition accuracy. Hence, we concluded that, explicit approach of modeling the subsegmental excitation by the LF parameters computed using the proposed approach may be the best possible way of representing the subsegmental excitation information for both speaker identification and verification tasks.

The next chapter will develop methods for the explicit modeling of speaker-specific information at the segmental level of the LP residual.

5

Explicit Segmental Processing of LP Residual for Speaker Information

Contents

5. Explicit Segmental Processing of LP Residual for Speaker Information

This chapter explores methods for explicit modeling of speaker-specific information present at the segmental level. The proposed methods process LP residual in spectral and cepstral domains and the speaker-specific excitation information is extracted by parameterizing the Fourier magnitude spectrum of the LP residual. Fourier representation often provides certain evidences of the signal that may be implicit or less evident in the time domain [1]. This helps us in modeling the speaker-specific excitation information from spectral and cepstral domains processing of the LP residual, that is relatively more compact and effective.

The LP residual magnitude spectrum mostly represents the energy and harmonic information associated with the excitation. In this chapter, methods are proposed to model the excitation energy and periodicity information and demonstrate their significance and different nature for speaker recognition. The excitation energy and periodicity information are modeled by subband energies and spectral flatness measure of the LP residual spectrum, respectively. First, we use the existing approaches to evaluate the excitation energy and periodicity information from the LP residual spectrum. Later, some modified techniques are proposed for effective representation of the excitation energy and periodicity information. In particular, by exploiting the nature of the LP residual spectrum, some refinements in the existing methods for extracting the speaker-specific excitation energy and periodicity information are proposed. The effectiveness and the different nature of speaker-specific information associated with the energy and periodicity of the excitation are demonstrated by speaker identification and verification experiments. The speaker-specific evidences from both mentioned aspects of the excitation are combined and an unified representative feature is proposed for modeling the excitation information. A comparative study is made on processing the LP residual in temporal, spectral and cepstral domains for compact and efficient representation of the speaker-specific excitation information. This chapter is concluded by proposing the best possible approach for processing the LP residual in spectral and cepstral domains for the extraction of speaker-specific excitation information at the segmental level.

5.1 Introduction

In Chapter 3, the LP residual is processed in the time domain to extract the subsegmental, segmental and suprasegmental levels speaker-specific excitation information. We observed that significant speaker-specific information is present at each of these levels and they together provide improved recognition accuracy. However, processing the LP residual in the temporal domain is computationally intensive. Because, in temporal domain processing of the LP residual the waveform itself is directly processed to model the speaker-specific excitation information. For example, for a given two minute training data sampled at 8 kHz, the number of LP residual feature vectors available for building the speaker model is around 40000 and of dimension 40. In case of spectral or cepstral domain approach the speaker-specific information is mostly extracted by parameterizing the short-term LP residual magnitude spectrum. This reduces the data rate. For example, a 20 msec segment of speech sampled at 8 kHz consists of 160 samples. By processing the short-term magnitude spectrum, the energy information associated with the segment can be represented by about 12 cepstral coefficients [59]. For this case the data reduction achieved is $160/12 \simeq 14$. Further, the computation of the short-term spectrum using fast fourier transform (fft) algorithm is simple and efficient [89]. Therefore spectral and cepstral domains approach are more compact and computationally efficient. Thus processing LP residual in spectral and cepstral domains is expected to model the speaker information in a more compact manner.

The existing attempts for processing the LP residual in spectral and cepstral domains for the extraction of speaker-specific excitation information mostly use short-time magnitude spectrum. Information present in the short-time magnitude residual spectrum may be broadly classified into two aspects, namely, the spectral energies and the nature of the harmonic structure. The spectral energies give information about the energy associated with the excitation [9, 59]. The harmonic structure of the spectrum gives information about the periodicity nature of the excitation [10, 15]. As mentioned in Chapter 2, these information have earlier been studied independently for speaker recognition and demonstrated their significance. However, these

approaches suffered from certain drawbacks. In [9], the cepstral analysis has been performed directly on the short-term LP residual magnitude spectrum which is nearly flat and hence may not seem to be a good choice. It may be better if the magnitude spectrum is passed through a bank of filters and then perform the cepstral analysis. In [10], the PDSS has been measured and used as feature to represent the excitation information. These PDSS values measure essentially represent the periodicity nature of the excitation [90]. In this study the features are extracted from multiple subbands. This is better because obtaining a global value from the spectrum may not likely to show good speaker-dependant characteristics. In this case the subbands are linearly spaced and may benefit by using non-linearly spaced subbands. The nonlinearity nature of human auditory perception system is expected to be an important factor for deciding the spacing in the subbands. All these studies demonstrate that the performance of the features derived from the source spectrum is not at par with the vocal tract information. This is because all the previous studies are independent and reflect different aspect of the source information and hence may not provide the complete source information. For example, cepstral analysis do not reflect the periodicity information and the $PDSS$ feature do not provide the energy information. The objective of this work is to develop a signal processing method for compact and efficient way of representing the speaker information associated with both energy and periodicity nature of the excitation by processing the LP residual in the spectral and cepstral domains.

The present work carries out some refinements in the methods employed earlier for extracting the energy and harmonic information associated with the source spectrum and analyze their different nature. While computing the features, the short-time residual spectrum is multiplied by the bank of filters. We call it as residual subband spectrum. In representing the energy information, the subband energies (SBE) are used. Then cepstral analysis is performed on the residual subband spectrum to capture more effective energy information. To capture the harmonic information, PDSS values are computed from the subband spectrum. The effectiveness of all these features are demonstrated by performing speaker recognition experiments. To verify the effect of the shape of the filters on the performance, rectangular, triangular and

mel filters are employed to compute the features. We analyze the nature of all these features and conduct both identification and verification experiments on relatively more noisy and large database to verify the effectiveness of the shape of the filter on the performance. The effective filter is chosen based on the recognition performance. The different nature of the energy and harmonic information is observed from the computational procedure and detailed recognition performance. The potential of the combined energy and harmonic information is demonstrated by comparing the performance of the combined information with individual information and also with vocal tract information. Finally, a comparative study is made between time and proposed spectral and cepstral domains approaches of processing the LP residual. The effectiveness of the proposed feature is verified by comparing the respective computational complexity and recognition performance.

The rest of the chapter is organized as follows: Section 5.2 describes the methods employed in processing the LP residual in the spectral domain for the extraction of features to model energy and periodicity information of the excitation. This section will also describe the speaker recognition studies that have been performed using these features. Section 5.3 describes cepstral domain processing of the LP residual to extract energy information and demonstrates its significance. In Section 5.4, the different nature of the information associated with energy and periodicity nature of the excitation will also be described and finally a combined feature will be proposed for complete representation of the excitation information. In Section 5.5, a comparative study of processing the LP residual in temporal, spectral and cepstral domains for the extraction of the excitation information is made and finally a method is proposed for the extraction of the complete excitation information. The last section summarizes the present work discussed in this chapter.

5.2 Processing of LP Residual in Spectral Domain

As mentioned in the introduction section, in spectral domain, the speaker-specific excitation information is extracted from the Fourier representation of the LP residual. The Fourier representation is a superposition of complex exponentials [1, 21]. Thus, unlike in time domain,

the Fourier representation contains both the amplitude and phase information. However, in the digital domain processing, the entire phase information of the original signal is confined to -π to +π. Hence, the phase component of the Fourier representation of the LP residual may not reflect significant speaker-specific information. Therefore, like processing the speech signal, majority of the existing attempts use Fourier magnitude spectrum of the LP residual (commonly termed as LP residual spectrum) for the extraction of speaker-specific excitation information.

In digital domain the fourier representation of a signal is commonly computed from discrete fourier transform (DFT). Mathematically, the N point DFT magnitude spectrum of the LP residual is given by Equation (5.1). In case of speech signal DFT computation, the $e(n)$ and $R(k)$ in Equation (5.1) are replaced by $s(n)$ and $S(k)$, respectively.

$$R(k) = \sum_{n=0}^{N-1} r(n) e^{-j\frac{2\pi nk}{N}} \tag{5.1}$$

The examples of speech, residual signals and their spectra for two different speakers, (*FS-1* and *FS-2*), are shown in Figure 5.1. Speech signals of the two speakers are different. This can also be observed from the modulation of the respective speech spectrum. Modulation in speech spectrum mostly represents the speaker-specific vocal tract information. There are also noticeable differences between the residual signals of the two speakers. In addition to the significant difference between their sample values, the residual signal of *FS-2* shows much stronger periodicity than that of *FS-1*. Due to impulse like nature, the residual spectra of both speakers are nearly flat. It therefore seems that they do not provide any speaker-specific information. However, one can observe from the residual spectrum that there are local variations in the spectral amplitudes and also dynamic ranges of the harmonics. For example, the distribution of the spectral amplitude within a frequency range is different from one speaker to the other. *FS-1* has significant energy around 1 kHz or at 2 kHz where as *FS-2* has around 500 Hz. This indicates that the nature of the distribution of the excitation energies across different ranges of frequencies is expected to be speaker-specific. Similarly, the harmonic structure of the residual spectrum within a frequency range is also different. For example, the local dynamic range

(defined as the difference between peak and dip of the spectrum within a frequency range) of the residual spectrum is different. For instance, the dynamic range of the residual spectrum of *FS-1* around 2 kHz Hz is 40 dB and 30 dB for *FS-2*, although both have overall dynamic range of 30 dB. The variation in the dynamic range depends on the periodicity of the spectrum. Larger the periodicity, more will be the difference between peaks and dips [10]. Since the periodicity is unique for a speaker, the nature of the harmonic structure of the residual spectra across different range of frequencies is expected to be containing speaker-specific information. All these variations described above are less variant for the same speaker as illustrated for the case of speaker *FS-1* in Figure 5.1 (j) and (k).

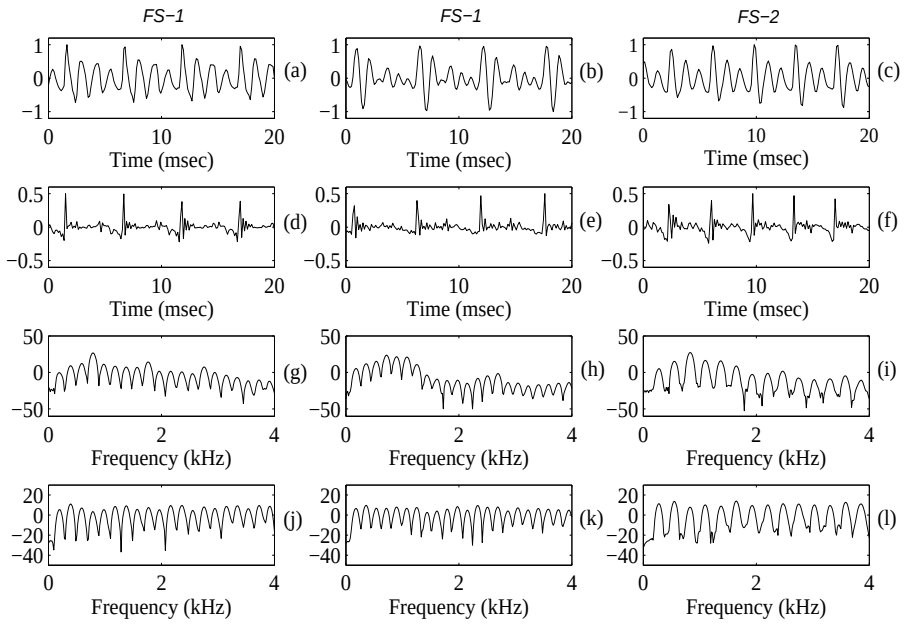

Figure 5.1: Vocal tract and excitation information of *FS-1* and *FS-2*. (a)-(b) Speech signals of *FS-1*. (c) Speech signal of *FS-2*. (d)-(e) LP residuals of *FS-1*. (f) LP residual of *FS-2*. (g)-(h) Magnitude spectra of speech signals of *FS-1*. (i) Magnitude spectrum of speech signal of *FS-2* . (j)-(k) Magnitude spectra of LP residuals of *FS-1*. (l) Magnitude spectrum of LP residual of *FS-2*. The first two columns represent two examples of the speaker *FS-1*.

In this section we describe the methods employed in processing the LP residual to extract speaker-specific information associated with energy and harmonic structure of the source spectrum. This information is extracted from the subbands of the source spectrum. The information associated with the energy is extracted in terms of subband energies (SBE) of the source

spectrum. The information associated with the harmonic structure is extracted by the power difference in subband spectrum (PDSS) measure. In the computation of SBE and PDSS features, the effect of filter shape on performance is verified by employing rectangular, triangular and mel filters.

5.2.1 Speaker Information from SBE of LP Residual Spectrum

SBE of the excitation represents the information about the distribution of the excitation energy in producing speech in different bands of frequencies. To model this information, SBE are computed from the LP residual magnitude spectrum. First subband spectra are obtained by multiplying the residual spectrum with a filterbank and then SBE $R_b(m)$ are computed by using Equation (5.2) given by

$$R_b(m) = \sum_{k=l_m}^{h_m} [|R(k)|H_m(k)]^2, \quad m = 1, 2, \ldots, M_b \tag{5.2}$$

where, l_m, h_m are the lower and upper limits of the sample frequency points of the m^{th} filter $H_m(k)$. We set M_b to be 24 filters with half window overlapping for computation of SBE. The purpose of choosing 24 is for later comparison with $MFCC$ features computed from the speech spectrum using 24 mel filters.

First we use rectangular filters to compute SBE and use them as features. We call them as residual rectangular subband energy (RRSE) features. Since 24 filters are used, the dimension of each $RRSE$ feature is 24. The training speech of each speaker is processed in blocks of 20 msec with a shift of 10 msec duration to extract $RRSE$ features. Components of each $RRSE$ feature is normalized by its maximum component value so that they remain between 0 to 1. The normalization is needed to avoid large fluctuations in the signal energy at different regions, such as in the weak and strong voiced sounds. Examples of residual energy spectra and corresponding $RRSE$ features of *FS-1* and *FS-2* are shown in Figure 5.2(e) and (f), respectively. It can be observed that $RRSE$ features of two speakers are different. This indicates that $RRSE$ features are speaker dependant and hence may contain speaker information. The RRSE feature also varies for the same speaker as illustrated for *FS-1* in Figure 5.2 (d) and (e). Thus RRSE

may also have large intra-speaker variability. To measure qualitatively the potential of *RRSE* features we compute its *Divergence* as described in Section 4.5.3. The divergence of *RRSE* features computed for both identification and verification experiments are given in the third column of the Table 5.1. The scalar divergence values indicate that *RRSE* features have the ability for speaker discrimination. The *RRSE* features from identification task of *Set-2* database have relatively lower divergence values than *Set-1* database. This indicates that *RRSE* features have less discriminating ability for more noisy speech. This is also observed from the recognition experiments conducted using these features. The results of speaker identification and verification experiments are given in the fourth column of the Table 5.1. The results show that *RRSE* features provide good recognition performance and hence contain speaker information. Also, the recognition performance is degraded for noisy database. The relative degradations in divergence and performance are 21.19% and 62%, respectively.

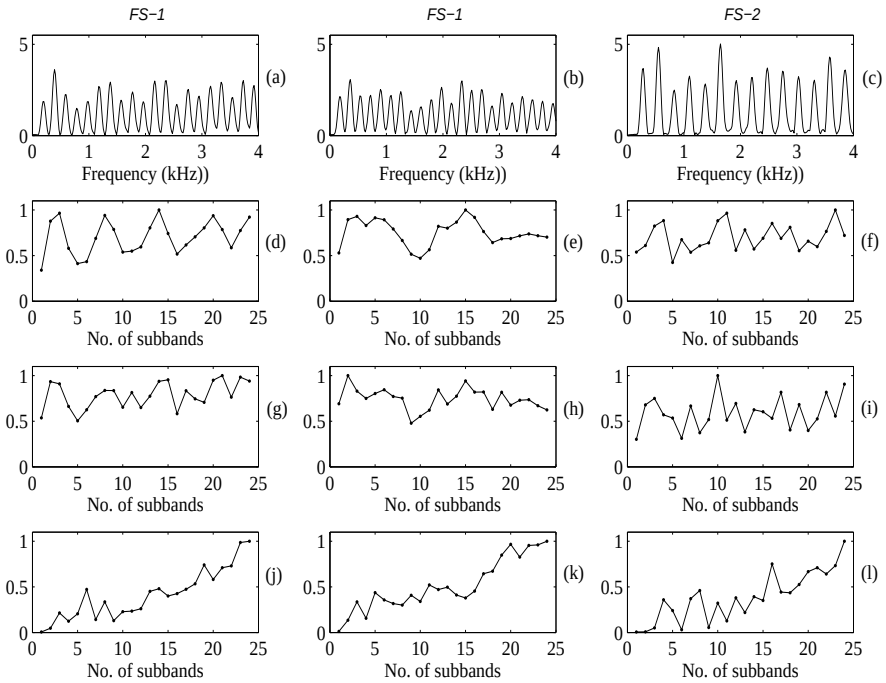

Figure 5.2: Subband energies computed from LP residual spectrum of *FS-1* and *FS-2*. (a)-(b) Residual spectra of *FS-1*. (c) Residual spectrum of *FS-2*. (d)-(e) *RRSE* features of *FS-1*. (f) *RSE* feature of *FS-2*. (g)-(h) *RTSE* features of *FS-1*. (i) *RTSE* feature of *FS-2*. (j)-(k) *RMSE* features of *FS-1*. (l) *RMSE* feature of *FS-2*. The first two columns represent two examples of the speaker *FS-1*.

Table 5.1: Speaker identification performance (in %) and verification performance (in EER) of $RRSE$, $RTSE$ and $RMSE$ features. (DIV) represents the *Divergence*. ($Perf$) represents the performance.

Task	Feature Database	$RRSE$		$RTSE$		$RMSE$	
		DIV	$Perf$	DIV	$Perf$	DIV	$Perf$
Identification	*Set-1*	21.97	53	23.11	59	24.98	71
	Set-2	17.31	20	19.50	26	21.01	37
	Relative Degradation	21.21	62	15.62	56	15.89	49
Verification	NIST-03	19.02	40.57	21.00	39.18	22.40	38.20

To verify the effect of the filter shape on performance, first we employ triangular filters to compute SBE features. We call them as residual triangular subband energies ($RTSE$). The computational procedure of $RTSE$ features is similar to $RRSE$ features except that the use of triangular filters. The comparison is made based on the *Divergence* measure and the recognition performance. Examples of $RTSE$ features for *FS-1* and *FS-2* are shown in Figure 5.2(h) and (i), respectively. RTSE also shows large intra-speaker variability (Figure 5.2(g) and (h))The effect of $RTSE$ features can be observed in the higher energy region. For example for *FS-1*, in the higher energy region around 1 kHz, the $RTSE$ features concentrate more compared the $RRSE$ features. This is due to the distribution nature of the triangular filters that provide more weightage to the central frequency component. Similar observation can also be made for *FS-2* around the 500 Hz or 2 kHz. Since significant speaker-specific information is present around higher energy regions of the excitation, $RTSE$ features may be more speaker discriminating in nature. This is also observed from the *Divergence* measure. The computed divergence values for $RTSE$ features are given in the fifth column of the Table 5.1. The *Divergence* values of the $RTSE$ features are better than the $RRSE$ features. Similar to $RRSE$ features, the *Divergence* of $RTSE$ features decrease with noisy speech. However, the relative degradation in the divergence of $RTSE$ features is less around 15.62% as compared to 21.19% in the case of $RRSE$ features. This indicates that $RTSE$ features are more robust against noise. Since $RTSE$ features emphasize more higher energy regions, the effect of noise may be reduced. This is also

observed from the experimental results given in the sixth column of the Table 5.1. Performance achieved by $RTSE$ features for both identification and verification task are better than $RRSE$ features. The relative degradation in the recognition performance for noisy database in case of $RTSE$ features is less around 56% as compared to 62% in case of $RRSE$ features. This indicates that triangular filters may be a better choice for computation of SBE features.

Motivated by this, we also employ non-uniformly placed triangular filters based on the mel scale and computed SBE features. We call them as residual mel subband energies ($RMSE$) features. The computational procedure of $RMSE$ features is similar to the $RTSE$ features except that the use of the mel filters. Examples of $RMSE$ features for *FS-1* and *FS-2* are shown in Figures 5.2(k) and (l), respectively. RMSE also have intra-speaker variability (Figure 5.2(i) and (k)). One can observe that there is no direct relation between the distribution of energies of rectangular and mel subbands. This is because the former is the actual distribution where as the later is based on perceptual distribution. It has been shown that humans can recognize speakers by listening to the LP residual [12,53]. This indicates that the perceptual distribution of subband energies may be a good choice than the actual distribution. Therefore we expect that $RMSE$ features may provide better recognition accuracy than $RRSE$ features. Further, it is expected that $RMSE$ features may provide even better performance than $RTSE$ features based on actual distribution. To verify this we derive $RMSE$ features and used them for both identification and verification experiments. The *Divergence* of the $RMSE$ features is given in the seventh column of the Table 5.1. The discriminating ability of $RMSE$ features is better than other two features. The robustness of $RMSE$ features against noise is better than $RRSE$ features and approximately similar to $RTSE$ features. For example, the relative degradation in the discriminating ability due to more noisy speech for $RMSE$ features is around 15.92% against 21.21% and 15.62% for $RRSE$ and $RTSE$ features, respectively. The recognition performances achieved for both identification and verification tasks are given in the eighth column of the Table 5.1. In all cases the performance achieved by $RMSE$ features is better than the other two features. Also, the relative degradation in the recognition performance due to noisy speech for $RMSE$ features is less than $RRSE$ and $RTSE$ features. Hence we

conclude that, as compared to $RRSE$ and $RTSE$ features, speaker information associated with the excitation energy can be better represented by $RMSE$ features .

5.2.2 Speaker Information from Harmonic Structure of LP Residual Spectrum

Rate of vocal folds vibration and manner in which vocal folds open and close show variations across speakers [15]. Hard voice corresponds to rapid and complete closing of vocal folds. Then the flow is discontinuous and excitation is more impulse-like in nature. So the residual spectrum is more flat. For soft speakers, the folds never close completely and there is a smooth air flow. In this case residual spectrum is comparatively less flat. Thus the dynamic range of the residual spectrum varies from speaker to speaker. The variation in the dynamic range depends on the periodicity nature of the spectrum [10]. Larger the periodicity, more will be the difference between peaks and dips of the spectrum. This can be better observed from the power spectrum, $p(k) = [R(k)]^2$, of the LP residual. Power spectra of *FS-1* and *FS-2* are shown in Figure 5.3(a)-(c). The dynamic range of *FS-2* that shows strong periodicity than *FS-1* is relatively more. In [10], this periodicity information is represented by PDSS features. PDSS can be interpreted as the subband version of spectral flatness (SF) measure of the power spectrum. SF of a spectrum is defined as the ratio of geometric mean (GM) to arithmetic mean (AM) of the spectral samples. Since $0 \leq GM \leq AM$, SF values vary from 0 to 1. PDSS from residual subband power spectrum is given by [10]

$$V(i) = 1.0 - \frac{\left[\prod_{k=l_i}^{h_i} p(k)\right]^{\frac{1}{N_i}}}{\frac{1}{N_i}\sum_{k=L_i}^{H_i} p(k)} \tag{5.3}$$

where, $N_i = h_i - l_i + 1$ is the sample number of frequency points in the i^{th} filter.

Since, $0 < SF \leq 1$, $PDSS$ values vary from 0 to 1. If the power spectrum have less dynamic range, for example nearly flat, then $GM \simeq EM$ and PDSS is less than one. Alternatively, if PDSS is low, the spectrum is less periodic. If the spectrum has peaks and dips, for example the dynamic range is more, then GM is less than EM and PDSS value is close to one. In this

case the spectrum is more periodic. So PDSS measure gives information about the periodicity nature of a spectrum. In [10], this information is extracted from power subband spectra for speaker recognition. In this study it is expected that, the periodicity nature of the excitation in the subband spectra may have more effective speaker-specific information. Subband spectra are obtained by multiplying the power residual spectrum with a filterbank and $PDSS$ values are computed using Equation (5.3).

In this work, first we use rectangular filters to compute PDSS and use them as features for speaker recognition. We call these features as residual rectangular PDSS ($RPDSS$) features. As suggested in [10], we set M_b to be 20 filters with half window overlapping for computation of $RPDSS$ features. Since 20 filters are used, the dimension of each $PDSS$ feature is 20. For extraction of $RPDSS$ features, the training speech of each speaker is processed in blocks of 20 msec with a shift of 10 msec duration. $RPDSS$ features computed for *FS-1* and *FS-2* are shown in Figures 5.3(d)- (f). It can be observed that PDSS values of two speakers are different. Because of strong periodicity, PDSS values of *FS-2* are relatively more than *FS-1*. It may be observed that the average values of PDSS are more close for the same speaker (*FS-1*). This indicates that $RPDSS$ features are speaker dependant and hence may contain speaker-specific information. The usefulness of this information is verified from both identification and verification experiments. The *Divergence* values of $RPDSS$ features from both identification and verification tasks are given in the third column of the Table 5.2. They indicate that $RPDSS$ features have the ability for speaker discrimination. The *Divergence* value of $RPDSS$ features for identification task of *Set-2* database is relatively lower than *Set-1* database. This indicates that $RPDSS$ features have less discriminating ability for more noisy speech. This is also observed from recognition results. The results of the speaker identification and verification experiments are given in the fourth column of the Table 5.2. $RPDSS$ features provide good recognition performance and hence contain speaker information. The recognition performance of the $RPDSS$ features is degraded for noisy database. The relative degradation in divergence and performance are 28.48% and 28.36%, respectively.

Motivated from the results obtained by $RMSE$ features, we also employ non-uniformly

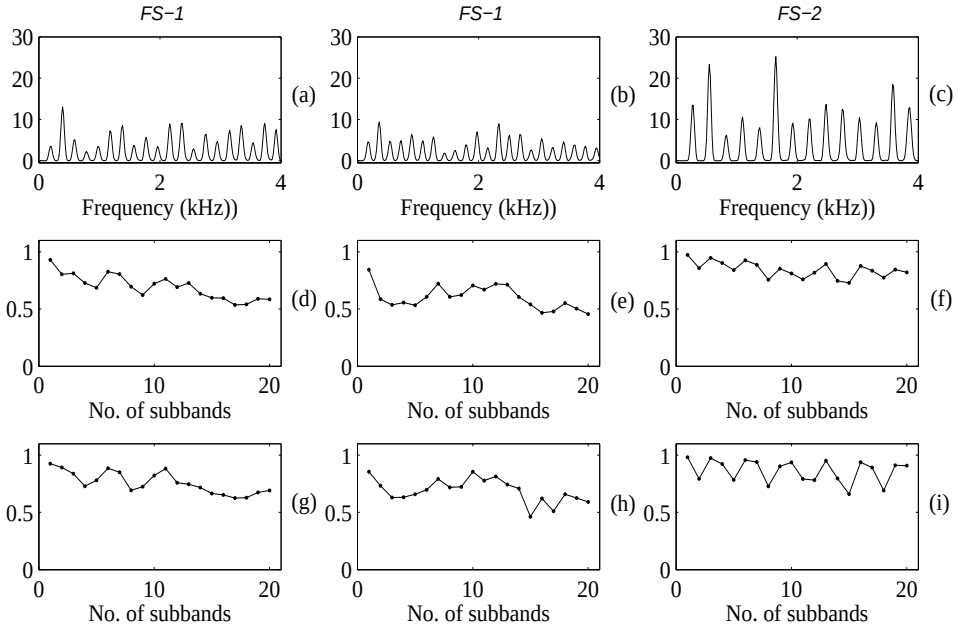

Figure 5.3: *PDSS* features from rectangular and mel filters for *FS-1* and *FS-2*. (a)-(b) Residual power spectra of *FS-1*. (c) Residual power spectrum of *FS-2*. (d)-(e) *RPDSS* features of *FS-1*. (f) *RPDSS* feature of *FS-2*. (g)-(h) *MPDSS* features *FS-1*. (i) *MPDSS* feature of *FS-2*. The first two columns represent two examples of the speaker *FS-1*.

placed triangular filters based on the mel scale for computation of PDSS features. The other motivation of using mel filters is from the property of the mel filterbank that provides less spectral samples to lower bands and more to higher bands (beyond 1 kHz). The dominant speaker information in the excitation is manifested in the higher range of frequency. Since PDSS is a statistical measure, with increase in number of samples, PDSS may be more accurately measured in the higher frequency range. So it is expected that PDSS computed from mel subbands may provide better recognition performance. To verify this, PDSS values are computed using mel filterbank. We call them as mel PDSS (*MPDSS*) features. The computational procedure of *MPDSS* features is similar to *RPDSS* features except the use of mel filters. Examples of *MPDSS* features for *FS-1* and *FS-2* are shown in Figures 5.3(g) - (i). In this case also, *MPDSS* values of *FS-1* is less than *FS-2*. This shows that the property of PDSS that is more for strong periodic spectrum still holds good, even if it is computed from mel warped spectrum. It can be observed that *MPDSS* values more accurately distinguish the periodicity nature of

Table 5.2: Speaker identification performance (in %) and verification performance (in EER) of $RPDSS$ and $MPDSS$ features. (DIV) represents the *Divergence*. ($Perf$) represents the performance.

Task	Feature Database	$RPDSS$		$MPDSS$	
		DIV	$Perf$	DIV	$Perf$
Identification	*Set-1*	34.97	67	42.39	73
	Set-2	25.01	48	31.33	48
	Relative Degradation	28.48	28.36	26.01	34.25
Verification	NIST-03	28.16	37.62	33.42	32.65

the speakers. For example, in case of *FS-1*, RPDSS values from subbands 2, 3 and 4 indicate that they are almost same. But this is not the case as observed from $MPDSS$ values. The $MPDSS$ value of subband 3 is comparatively less. This indicates that perceptually, the periodicity of the *FS-1* in the third subband is weak. Similar observation can also be made for *FS-2* in subbands 4, 5 and 6. Thus we may therefore expect that $MPDSS$ feature may provide relatively better recognition accuracy.

Both identification and verification experiments are conducted to verify the potential of $MPDSS$ feature. The *Divergence* values and results of $MPDSS$ feature are given in fifth and sixth columns of the Table 5.2, respectively. In all cases the discriminating ability of $MPDSS$ is better than $RPDSS$ feature. The relative degradation in the discriminating ability of $MPDSS$ feature due to noise is less as compared to $RPDSS$ feature. The recognition result of identification task indicates that for less noisy database, $MPDSS$ feature provides good recognition accuracy. For more noisy database the identification performance of $RPDSS$ and $MPDSS$ features are same. However, in verification task with same database, the performance achieved by $MPDSS$ features is significantly better than the $RPDSS$ features.

By comparing the recognition performance of energy and periodicity information of the excitation, we observed that $MPDSS$ provides better recognition performance than $RMSE$ feature. The reason may be that, $RMSE$ feature represent the amplitude of the excitation signal in subbands. Information associated with the instantaneous variation of the excitation

may be more useful. Because, instantaneous values represent both amplitude and sequence information. Hence it is expected that speaker-specific information associated with instantaneous variation of the excitation signal may give better recognition performance.

5.3 Speaker Information from Cepstral Analysis of LP Residual

In the previous section, the LP residual is processed in spectral domain to extract speaker-specific excitation energy information. It is shown that this information can be effectively represented by mel subband energies. These energies are computed by summing the squared mel warped spectral magnitude samples. Thus, subband energies mostly represent the envelope of the excitation signal in the spectrum. The envelope may be treated to be made of several slowly varying components and are being manifested as local variations. The envelope does not give much information about the local variation in the excitation signal. It is conjectured that information associated with the local variation of the excitation signal may be more speaker-specific by nature. This information can be extracted by processing the LP residual in cepstral domain. The cepstral domain representation of the LP residual, say $C(n)$, is given by [9],

$$C(n) = Re\left(\frac{1}{N}\sum_{k=0}^{N-1} ln|R(k)|e^{j\frac{2\pi nk}{N}}\right), \quad n \in [0, N/2] \tag{5.4}$$

Because of symmetric spectrum due to real nature of LP residual, cepstral duration is truncated to just half the segment. Cepstral samples essentially represent nearly the temporal variation of the excitation signal. In [9], cepstral representation computed for every 20 msec segment of LP residual with a shift of 10 msec are used for speaker recognition experiments. In our experiment we prefer to use only first few samples excluding the first one. The reason is as follows: the first cepstral value is relatively very large (Equation (5.4)) and hence may dominate the effect of other cepstral samples, and, use of all remaining cepstral samples may be of very large dimension. For example, for a speech signal sampled at 8 khz, 20 msec segment of residual spectrum have at least 80 cepstral samples. Such large dimension feature may not seem to be effective for speaker recognition. In general, for signals having 160 samples 256 points fft is

used which will further increase the dimension. For this, we prefer to use lower dimensional cepstral representation. In selecting the number of cepstral samples to represent as feature, *F-ratio* measure may be used [86].

Figure 5.4(a) and (b) show *F-ratio* values of 127 cepstral samples derived from 256 point fft residual spectrum for 90 speakers taken from *Set-1* and *Set-2* databases, respectively. It can be observed from these figures that in each case the *F-ratio* value of cepstral samples beyond 25 is significantly decreased. Similar observation is also made from *F-ratio* values shown in Figure 5.4(c) computed for larger data set consisting of 356 speakers taken from NIST-03 database. This observation indicates that cepstral samples beyond 25 may have large intra-speaker variability and hence may not be effective for speaker recognition purpose. Thus, in this work cepstral samples beyond 25 are ignored. Examples of cepstral representation of first 24 samples of LP residual excluding first one for *FS-1* and *FS-2* are shown in Figures 5.5(a) - (b). It can be observed that, cepstral representation of LP residuals are different from speaker to speaker. This indicates that cepstral samples of LP residual contain speaker-specific information and may be used as feature for speaker recognition. The cepstral shows intra-speaker variability (*FS-1*). To verify the significance of these features, different speaker recognition experiments are conducted using them. In our experiments we use only thirteen cepstral samples to represent a feature. The purpose of using thirteen cepstral samples is in accordance with the conventional use of thirteen dimensional MFCC features for comparison. Since cepstral sample features are derived directly from fft residual spectrum we call them as residual fast fourier transform cepstral coefficient ($RFFTCC$) features.

The *Divergence* and recognition performances of $RFFTCC$ features are given in the third and fourth columns of the Table 5.3, respectively. *Divergence* values indicate that $RFFTCC$ features have the speaker discriminating ability. The identification performances achieved for *Set-1* and *Set-2* databases are 78% and 41%, respectively. The verification performance achieved is 40.28%. These results show that $RFFTCC$ features indeed contain good amount of speaker information.

The recognition performance achieved by $RFFTCC$ features may be compared with the

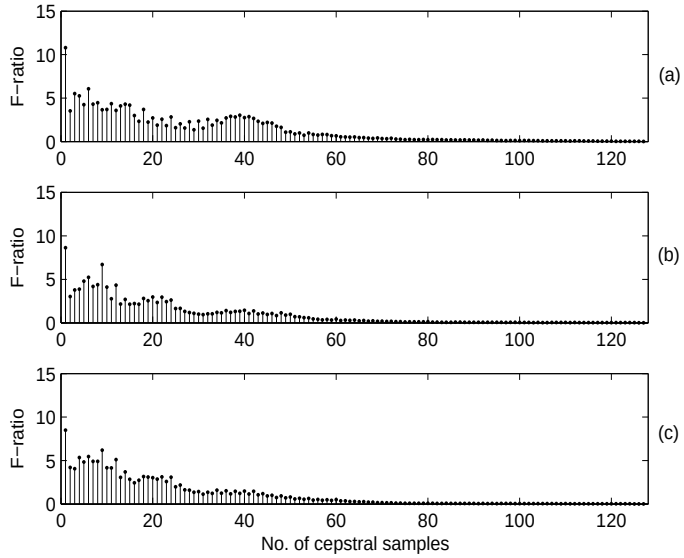

Figure 5.4: F-ratio values of 127 cepstral samples derived from (a) *Set-1*, (b) *Set-2* and (c) Larger set of 356 speakers from NIST-03 database.

subband energy features computed from spectral domain processing of the LP residual. For identification task, $RFFTCC$ features are providing relatively good recognition performance compared to subband features. This is because, cepstral values capture the variation in the energy of the excitation. Further, for verification task, despite more discriminating ability, the recognition performance of $RFFTCC$ features is relatively less. This indicates that compared to spectral energy features, $RFFTCC$ features are relatively less effective for more noisy and large database. The reason may be in the computational procedure involved in extracting these features. Spectral energy features are computed from subbands spectrum and $RFFTCC$ features are computed from the whole spectrum which is nearly flat. Features extracted directly from a flat spectrum may not seem to be effective for speaker recognition. Thus we propose to extract the cepstral features from subband spectrum. For this, first we obtain the subband spectrum from filtering operation and then cepstral analysis is performed to extract speaker-specific features. We call them as residual cepstral coefficient (RCC) features. The components of the RCC features are derived from discrete cosine transform of log magnitude spectrum of subbands. The subband spectra are obtained by using filterbank. As mentioned earlier, the

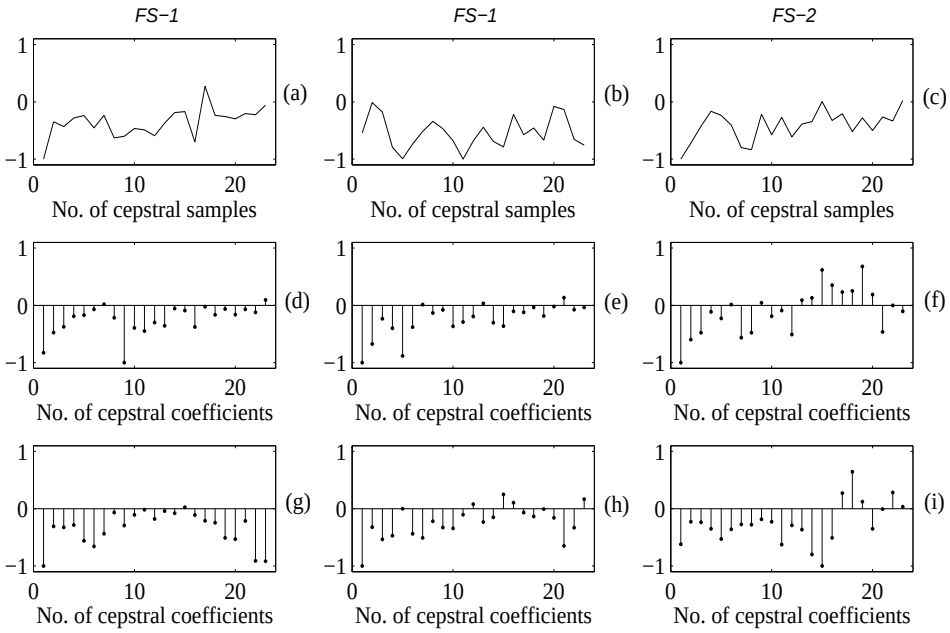

Figure 5.5: Cepstral features of *FS-1* and *FS-2*. (a)-(b) $RFFTCC$ features of *FS-1*. (c) $RFFTCC$ feature of *FS-2*. (d)-(e) $RRFCC$ features of *FS-1*. (f) $RRFCC$ feature of *FS-2*. (g)-(h) $RMFCC$ features of *FS-1*. (i) $RMFCC$ feature of *FS-2*. The first two columns represent two examples of the speaker *FS-1*.

filterbank used consists of 24 overlapping filters. First we employ rectangular filterbank and compute cepstral coefficients by using Equation (5.5).

$$C_{ir}(i) = \sum_{m=1}^{M} X(m) cos[i(m - \frac{1}{2})\frac{\pi}{M}] \tag{5.5}$$

where, i=1,2,. . . ,C is the number of cepstral coefficients, (usually C<M) and $\mathrm{X(m)} = \log_{10}(\sum_{\mathrm{k=0}}^{\mathrm{N-1}}[\mathrm{R_m(k)}]$ represents the log energy output of the m^{th} filter. In this work, these coefficients are termed as residual rectangular filter cepstral coefficients ($RRFCC$). Examples of $RRFCC$ of *FS-1* and *FS-2* are shown in Figures 5.5(d) and (f), respectively. The variations in the distribution of $RRFCC$ of speakers are different. This indicates that $RRFCC$ are speaker dependant and may provide useful information. RRFCC also exhibit more intra-speaker variability (*FS-1*).

To verify the effectiveness of the $RRFCC$, we use these coefficients as components of cepstral features and used them for speaker recognition. Earlier studies on speaker recognition using cepstral features like $LFCC$ or $MFCC$, mostly use first 13-19 coefficients excluding c_0 as

Table 5.3: Speaker identification performance (in %) and verification performance (in EER) of $RFFTCC$, $RRFCC$ and $RMFCC$ features. DIV)represents the *Divergence*. $Perf$ represents the performance.

Task	Feature Database	$RFFTCC$		$RRFCC$		$RMFCC$	
		DIV	$Perf$	DIV	$Perf$	DIV	$Perf$
Identification	*Set-1*	31.50	78	23.63	78	32.52	81
	Set-2	26.10	41	21.47	48	29.70	52
	Relative Degradation	17.14	47	9.14	38	8.67	36
Verification	NIST-03	28.83	40.28	23.13	37.85	30.84	35.14

feature components [3, 4]. In our case since similar feature extraction technique is employed except that the use of LP residual signal, we prefer to use thirteen cepstral coefficient to form features. for speaker recognition experiments. The discriminating ability and the recognition performance of these features are given in fifth and sixth column of the Table 5.3, respectively. *Divergence* values obtained for $RRFCC$ indicate that they have speaker discriminating ability. However, compared to $RFFTCC$ features, the discriminating ability of the $RRFCC$ features is relatively less but provides better recognition performance. This indicates that $RRFCC$ features contain less redundant information [8].

To verify the effect of the window shape in recognition performance, mel filters are employed to obtain the subbands. In this case the computational procedure remain same except that the use of the mel filterbank. The cepstral coefficients computed from mel subband spectra are termed as residual mel filter cepstral coefficients ($RMFCC$). Example of $RMFCC$ of *FS-1* and *FS-2* are shown in Figures 5.5(g) and (i). The variation in the distribution of $RMFCC$ of speakers are different and also different from the $RRFCC$ distribution. RMFCC seems to provide relatively less intra-speaker variability compared to RRFCC. Since mel scale based on human perception is applied, it is expected that $RMFCC$ may be a better choice than $RRFCC$. Features are formed by using thirteen $RMFCC$ as components and used for speaker recognition experiments. The discriminating ability and the recognition performance of these features are given in seventh and eighth column of the Table 5.3, respectively. *Divergence* values

of $RMFCC$ features are relatively more than $RRFCC$ features. This indicates that $RMFCC$ features have more discriminating ability. This is also observed from the recognition results. In all cases $RMFCC$ features are providing better performance. The relative degradation in performance due to noise in case of $RMFCC$ is around 8.67% which is less than $RMSE$ features of around 15.89%. This shows that $RMFCC$ features are more robust against noise. Thus we conclude that energy information associated with the excitation may be better represented by $RMFCC$ features.

The recognition performance achieved by cepstral features indicates that speaker-specific excitation information associated with the excitation can be extracted from the cepstral domain processing of the LP residual. Compared to $RMSE$, $RMFCC$ features are relatively more robust against noise and provide improved recognition performance and hence may be a better choice for representation of the excitation energy information.

5.4 Speaker Information from combined Spectral and Cepstral Domains

In the previous sections we demonstrated that periodicity and energy information associated with the excitation can be effectively represented by $MPDSS$ and $RMFCC$ features, respectively. However, they reflect different nature of speaker information. By way of deriving these features, the information present in $MPDSS$ and $RMFCC$ features are different. $MPDSS$ features are computed from each subband spectra. Thus they represent speaker information associated with local variation of the spectrum, where as, computation of $RMFCC$ features involves whole spectrum and thus represent the speaker information associated with gross variation of the spectrum. By comparing their *Divergence* values, it can be observed that $MPDSS$ features have more discriminating ability. By comparing their verification performance on large noisy database, it can be observed that $MPDSS$ features provide good performance. This indicates that speaker-specific information present in $MPDSS$ features is more robust against noise. This can also be observed by comparing the relative degradation in their identification performance due to noise. For example the relative degradation in case of $MPDSS$ features

is around 34.25% where as for $RMFCC$ feature it is 36%. These observations indicate that $MPDSS$ and $RMFCC$ features indeed contain different aspect of speaker-specific excitation information. In this section we use confusion patterns and scatter diagrams to further explain the different nature of the speaker information present in $MPDSS$ and $RMFCC$ features and their usefulness for combined use in speaker recognition.

Confusion patterns from identification results conducted using $MPDSS$ and $RMFCC$ features for *Set-1* and *Set-2* databases are shown in Figure 5.6. In each case, the confusion pattern is entirely different. The decisions for both true and false identification are different. This indicates that they reflect different aspect of excitation information. This may help in combining the evidences to further improve the recognition performance. For combination, we use score level combination scheme $Comb_1$ described in Section 3.4. In this work the combined representation of $MPDSS$ and $RMFCC$ features is abbreviated as Src_4. The identification performance of Src_4 for both datasets are given in the third row of the Table 5.4. In case of *Set-1* database, the performance is improved from 81% to 82%. In case of *Set-2* database, the performance is decreased from 52% to 51%. The reason may be due to the combination scheme employed.

Table 5.4: Speaker recognition performances of combined evidence from $RMFCC$ and $MPDSS$ features. Src_4 represents $RMFCC + MPDSS$ by score level ($Comb_1$) combination scheme.

Feature		$Performance$		
		$Identification(\%)$		$Verification$
		Set-1	*Set-2*	(EER)
$RMFCC$		81	52	35.14
$MPDSS$		73	48	32.65
Src_4	$Comb_1$	82	51	32.33

In case of verification task, the different aspect of the information present in $MPDSS$ and $RMFCC$ is observed from their respective distribution of scores for imposter and genuine trails [29]. Distribution of two dimensional (2-D) LLR scores for genuine and imposter trials for $MPDSS$ and $RMFCC$ features are shown in Figure 5.8. In region I, $MPDSS$ feature

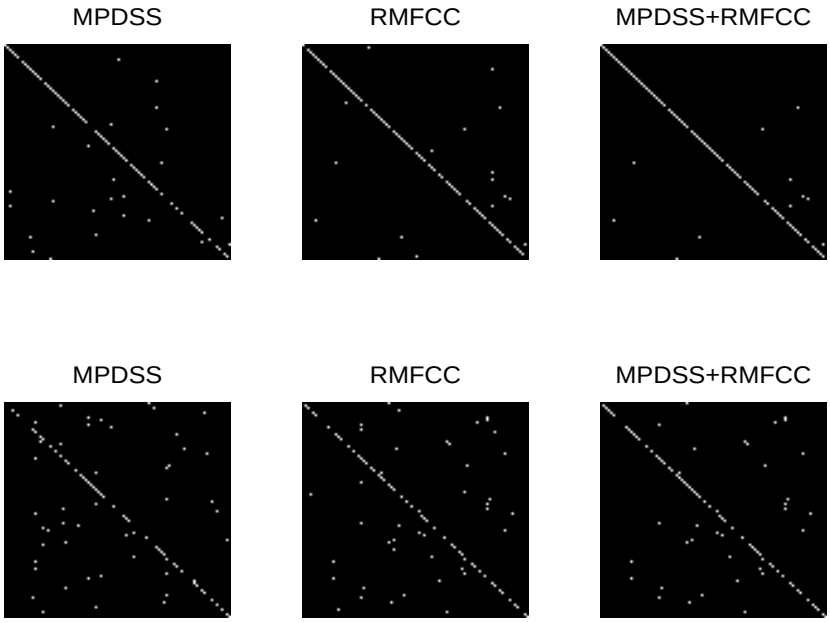

Figure 5.6: Confusion patterns from identification from identification results of $RMFCC$ and $MPDSS$ features their combinations. (Top) *Set-1* database. (Bottom) *Set-2* database.

rejects but $RMFCC$ accepts. Similarly in region II, $MPDSS$ feature accepts but $RMFCC$ rejects. In the regions I and II, some genuine rejected and imposters accepted by one feature are corrected by other. These observations indicate the different nature of speaker information present in $MPDSS$ and $RMFCC$ features. In combining the evidences from $MPDSS$ and $RMFCC$, the verification accuracy is further improved. The verification performance of Src_4 is also given in the third row of the Table 5.4. The best individual performance of $MPDSS$ feature of 32.65% is improved to 32.33%. Improvements in both identification and verification performances due to combination of $MPDSS$ and $RMFCC$ features indicate that they reflect different aspect of excitation information and combined well to further improve the recognition performance. Thus, we propose that periodicity and energy of excitation information can be effectively represented by combination of $MPDSS$ and $RMFCC$ features.

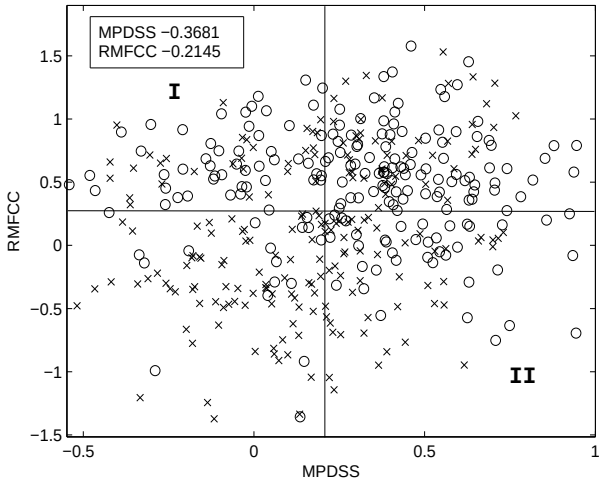

Figure 5.7: Distribution of 2-D LLR scores for genuine and imposter trails of $RMFCC$ and $MPDSS$ features.

5.5 Comparison of Processing LP Residual in Temporal, Spectral and Cepstral Domains

In this section a comparative study is made on processing the LP residual in temporal, spectral and cepstral domains. The comparison is made based on computational complexity and recognition accuracy. Although, there is no decimation involved, the dominant information in the combined representation of the $RMFCC$ and $MPDSS$ features is the segmental excitation information. Thus, the segmental level excitation information in time domain may be compared with the information present in the combined representation of the $RMFCC$ and $MPDSS$ feature. The excitation information is represented by combining the subsegmental, segmental and suprasegmental levels information. However, in frequency domain approach, due to non-stationarity nature of the speech signal, it is difficult to extract the suprasegmental level excitation information from the source spectrum. The first comparison is made for extracting the segmental level excitation information. Then, second comparison is made for extracting the improved excitation information by adding suprasegmental level information to both temporal and spectral-cepstral features separately. Finally, their significance is further verified by combing them with the vocal tract information.

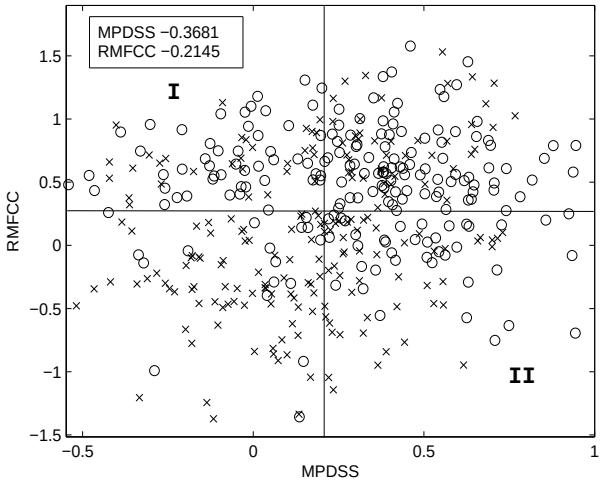

Figure 5.8: Distribution of 2-D LLR scores for genuine and imposter trails of $RMFCC$ and $MPDSS$ features.

For comparison study, the subsegmental, segmental and suprasegmental level speaker-specific excitation information is extracted by processing the LP residual in the time domain as suggested in Chapter 3. Time domain features are derived directly from the sequence of samples of LP residual and themselves are used to build the speaker models. The duration of blocks and block shift considered in the extraction of time domain features is relatively more than the frequency domain approach. The dimension of the time domain features is relatively more than frequency domain features. For example, for a speech signal sampled at 8 kHz, the dimension of segmental feature is 40 and that of $MPDSS$ and $RMFCC$ features is 20 and 13, respectively. Thus, time domain features are computationally more intensive. Since, there is no parameterization involved, time domain features give lossless representation of the information. On the other hand, the frequency domain approach involves parametric representation and hence computationally less intensive and provides lossy representation of the information.

To verify the time and frequency domain features potential in recognition performance, we compare their respective recognition performances and different nature in providing the improved representation of the excitation information. The speaker identification and verification results are given in the Table 5.5. The results show that, for clean speech, features derived from frequency domain approach provide relatively good performance. For example, the Src_4

provides 82% identification accuracy for *Set-1*, as against 60% by segmental (Seg). For noisy speech case the time domain feature provides relatively good performance. For example, the Src_4 provides 51% identification accuracy and EER of 32.33%, as against 58% and 26.96% for *Set-2* dataset and NIST-03 database, respectively. The improvement in the performance of time domain features in case of noisy speech may be due to the lossless representation of the information. For less noisy speech, the speaker-specific information may be better modeled by the frequency domain features. Also, in case of less noisy speech, Src_4 representation combines well with the vocal tract features. This indicates that frequency domain features may be a preferred choice for less noisy data for individual modelling of the segmental excitation information. By adding the subsegmental and suprasegmental levels information, we observe that the frequency domain features are providing relatively more evidence for improved representation of the excitation information. For example, $Sub + Src_4 + Supra$ provides identification accuracy of 83% and EER of 32.33% for *Set-1* dataset and NIST-03 database, respectively. As mentioned earlier, the poor performance of $Sub + Src_4 + Supra$ for *Set-2* dataset may be due to the combination scheme employed. The excitation information either extracted from

Table 5.5: Speaker recognition performances from processing the LP residual in temporal, spectral and cepstral domains. Src_4 represents $RMFCC + MPDSS$ by score level ($Comb_1$) combination scheme. Seg represents segmental and Src_2 represents combined ($Comb_1$) representation of subsegmental, segmental and suprasegmental excitation information from time domain processing of the LP residual.

Feature		$Performance$		
		$Identification(\%)$		$Verification$
		Set-1	*Set-2*	(EER)
$MFCC$		87	66	22.94
Seg		60	58	26.96
Src_4	$Comb_1$	82	51	32.33
$Src_4 + MFCC$	$Comb_1$	89	66	22.94
$Sub + Supra$	$Comb_1$	60	51	40.96
Src_2	$Comb_1$	74	58	33.28
$Sub + Src_4 + Supra$	$Comb_1$	83	53	32.33
$Src_2 + MFCC$	$Comb_1$	87	70	22.99
$Sub + Src_4 + Supra + MFCC$	$Comb_1$	91	70	22.42

the time or frequency domain approach helps in improving the recognition performance of the vocal tract features. The contribution of the excitation information extracted from the time domain processing is slightly better than frequency domain approach. In case of identification task, by combining the frequency domain features the $MFCC$ performance for *Set-1* and *Set-2* datasets is improved from 87% to 91% and 66% to 70%, as compared to 87% and 70% with time domain features, respectively. Similarly for verification task, the $MFCC$ performance is improved by from 22.94% to 22.42%, as compared to 22.99% with time domain features. To summarize, in general time domain processing of LP residual is computationally more intensive than frequency domain approach. The recognition performance of time domain features is slightly better than corresponding frequency domain features. Both time and frequency domain features are well combined with subsegmental and suprasegmental and vocal tract information. Vocal tract information is more useful with frequency domain features. Since frequency domain approach is computationally less intensive and almost equally effective with time domain approach, we suggest that combined use of $RMFCC$ and $MPDSS$ features may be the better way of representing the segmental excitation information.

5.6 Summary

In this chapter, the LP residual is processed in spectral and cepstral domains to extract the speaker-specific information associated with periodicity nature and energy of the excitation. Information associated with the periodicity nature of the excitation is extracted from the power differences measure of the subband source spectra and is represented by $MPDSS$ features. Information associated with the excitation energy is extracted by the ceptral analysis performed on subband spectra and is represented by $RMFCC$ features. Experimental results show that $MPDSS$ and $RMFCC$ features provide useful speaker-specific information. By combining the individual evidence from $MPDSS$ and $RMFCC$ features, the improvement in the recognition performance indicates the different nature of speaker information present in them. These features mostly represent the dominant segmental level excitation information and therefore do not represent the complete excitation information. The combination of spectral and cepstral

domains features together with the subsegmental and suprasegmental level information further improves the recognition performance. A comparative study is made between the time and frequency-cepstral domains processing of the LP residual for the extraction of speaker-specific excitation information. The time domain processing of the LP residual provides lossless representation of the information but computationally more intensive. The spectral and cepstral domains processing of the LP residual is computationally less due to significantly reduced computational complexity, it is suggested that for the extraction of speaker-specific excitation information, the spectral and cepstral domains approaches of processing the LP residual may be used with a slight compromise in the recognition performance. The next chapter explores approaches for modeling the suprasegmental level excitation information explicitly by processing the LP residual in time and cepstral domains.

6

Explicit Suprasegmental Processing of LP Residual for Speaker Information

Contents

This chapter develops an explicit representation of the speaker-specific information at the suprasegmental level of the excitation signal by processing the LP residual in the time and quefrency domains. This will be an alternative for the method developed in the temporal processing of LP residual described in Chapter 3. In the proposed approach, the pitch related suprasegmental information is modeled by pitch and epoch strength vectors. Pitch and epoch strength are computed from the HE of LP residual using the recently proposed *zero-frequency filtering* approach [62, 63]. Successive ten pitch and epoch strength values with a shift of one value are used as feature vectors to model the suprasegmental information. The individual $RMFCC$ temporal trajectories are also explored as additional suprasegmental information. The cepstral trajectories are selected based on the statistical F-ratio measure. A block of ten with shift of one value are used as cepstral trajectory vectors for each cepstral coefficient. Experimental results show that pitch, epoch strength and cepstral trajectory vectors individually well model the respective suprasegmental information. The evidences from all the cepstral trajectory vectors together provide the best performance. The speaker-specific evidences from pitch, epoch strength and cepstral trajectories are observed to be different. By combining these evidences, the performance is improved further. The experimental results show that the proposed approach models the suprasegmental information at par with the subsegmental and segmental levels. Finally, a comparative study is made with the temporal domain processing of the LP residual at the suprasegmental level. This study indicates that the present approach provides better performance. Thus, we suggest the combined representation of pitch and epoch strength vectors together with the $RMFCC$ trajectories as the suprasegmental level speaker-specific excitation information for speaker recognition.

6.1 Introduction

The suprasegmental speaker-specific excitation information represents the temporal variations of the excitation characteristics across several segments, in particular, over 100 msec. Humans continuously change the tension of the vocal folds and the subglottal air pressure for speech production [2,8]. As a result, the average rate of vocal folds vibration (pitch) varies with

time. Similarly, the strength of excitation at the instants of vocal folds closing (epoch strength) also varies with time. Thus pitch and epoch strength vary with time. Due to physiological differences in the vocal folds and associated muscle structure, the variations in the pitch and epoch strength show speaker dependant characteristics and are found to be effective for speaker recognition [2,8,19,91]. Due to changes in the tension and mass lesions of the vocal folds across speakers, variations in the cepstral trajectories of LP residual may also be speaker dependant as in the case of cepstral trajectories of speech [19,92]. In Chapter 3, the suprasegmental level information is modeled by the time domain processing of the LP residual. Although, the LP residual processed at 100 msec blocks demonstrated capturing the suprasegmental information, the performance was poor and the approach is computationally intensive. It was also observed that the 100 msec blocks mostly look like mere noise sequences (see, Figure 3.2 (g)). We may benefit by modeling the information at the suprasegmental level in an explicit manner. Hence the motivation for the work described in this chapter.

The effectiveness of the suprasegmental level pitch contour information depends upon the accurate estimation of the pitch. There are several methods in the literature for pitch estimation [62,75,93]. These methods are broadly classified into two categories: block based and event based. In the block based method, pitch period is estimated mostly by either the autocorrelation analysis of speech or LP residual or finding peak in the cepstrum [8,94]. These methods provide good estimate of the pitch period but suffer from the limitation of block processing, such as only providing the average value not the exact pitch value. Alternatively, the event based pitch estimation measures the pitch period by locating the instants of glottal closure in the speech signal. The recently proposed event based *zero-frequency filtering* approach is demonstrated to be providing the most accurate estimation. Another advantage of the zero-frequency filtering approach is that the epoch strength can also be easily measured from the zero-frequency filtered signal (ZFFS). Therefore, in this work the *zero-frequency filtering* approach is used to measure the pitch and epoch strength contours. As mentioned in chapter 4, the *zero-frequency filtering* approach may not work well in the case of telephonic speech [62]. To avoid this difficulty, we propose to use the HE of the LP residual as the input to the zero-frequency filter

for pitch and epoch strength estimation. The sequences of pitch and epoch strength values are used as feature vectors to model the suprasegmental information. The significance of the speaker-specific information present in the pitch and epoch strength vectors are demonstrated by speaker identification and verification experiments. Since, the pitch and epoch strength vectors are computed from the HE of the LP residual by employing *zero-frequency filtering*, the approach is also a temporal processing one, but gives explicit modeling of pitch and epoch strength contours, as opposed to the implicit modeling in Chapter 3.

The cepstral trajectories of the LP residual can be modeled by processing the LP residual in the quefrency domain. In this work, an approach is developed to model the $RMFCC$ trajectories. The cepstral coefficients essentially represent the slowly varying components of the source spectrum at the segmental level. The variation of individual cepstral coefficients across several segments may therefore be useful for modeling the suprasegmental information. In this work, the blocks of overlapping individual cepstral coefficients are used as feature vectors to model the suprasegmental information. First, the significance and different nature of the speaker-specific excitation information present in the $RMFCC$ trajectory vectors are experimentally demonstrated from different speaker identification and verification studies. Then, the individual evidences from cepstral trajectory vectors are combined. Since, the cepstral trajectory vectors are computed from $RMFCC$, the procedure may be viewed from the quefrency domain processing of the LP residual.

The pitch, epoch strength and cepstral trajectories represent different aspect of excitation characteristics. Thus, they seem to be providing different speaker-specific evidence. The different nature of the suprasegmental information is studied from the respective confusion patterns and score distribution diagrams. The individual evidences are then combined to represent the complete suprasegmental information.

The rest of the chapter is organized as follows: Section 6.2 describes the method for the extraction of pitch and epoch strength information. This section also describes the speaker recognition study made using pitch and epoch strength vectors. Section 6.3 describes the method for the extraction of $RMFCC$ cepstral trajectory vectors and demonstrates their significance

by conducting different speaker recognition studies. In Section 6.4, the individual evidence are combined. In this section the significance of the proposed representation is compared with subsegmental and segmental levels excitation and also with the vocal tract information. The last section summarizes the work presented in this chapter.

6.2 Speaker-specific Information from Pitch and Epoch Strength Contours

This section begins with the brief description of the *zero-frequency filtering* method for epoch extraction. Methods for computing the pitch and epoch strengths from the extracted epochs are then described. With suitable dimension, the vector representation of the pitch and epoch strength contours are proposed as the representative features to model the suprasegmental information. The significance of speaker-specific information present in the pitch and epoch strength vectors are demonstrated by speaker identification and verification studies.

6.2.1 Zero-frequency filtering Method for Pitch and Epoch Strength Estimation

The zero-frequency filtering method estimates the pitch and epoch strength by locating epochs or glottal closure instants (GCIs) [62, 63, 81]. As described in Section 4.4.1.1, the GCIs in telephone speech can be computed by passing the HE of the LP residual through a zero-frequency filter. The resulting signal is called as the zero-frequency filtered signal (ZFFS). The positive zero crossings in the ZFFS correspond to the locations of the epochs or GCIs [63]. The interval between successive positive zero-crossings gives the instantaneous pitch period T_0. The reciprocal, $P_0 = \frac{1}{T_0}$ is the instantaneous pitch frequency [62]. The slope of the ZFFS around the zero crossings corresponding to the location of the epochs gives a measure of epoch strength A_0 [81]. The slope around the epochs can be computed as the absolute difference between the preceding and succeeding sample amplitudes of ZFFS around the GCIs [81].

6.2.2 Speaker Recognition Studies using Pitch and Epoch Strength Contours Information

In this work, modified zero-frequency filtering approach as described in Section 4.4.1.1 is used for the computation of pitch and epoch strength contours to model the suprasegmental excitation information. Figure 6.1 shows the examples of pitch and and epoch strength contours of two female speakers, *FS-1* and *FS-2* collected from TIMIT database. It can be observed that the contours are different across the speakers. This shows that pitch and epoch strength contours contain speaker-specific information. The contours also show variation for the same speaker (*FS-1*) indicating large intra-speaker variation. Since the contours are computed across a longer segment of the voiced speech, the speaker-specific information present in them is attributed to the suprasegmental level. The suprasegmental information is usually extracted from 100-300 msec segments and hence we need on an average around 25-50 pitch values to represent a feature vector. Since the nature of pitch and epoch strength contours have high intra-speaker variability, the dimension of the feature vectors consisting of 25-50 values may not seem to be effective for the recognition task. Pitch and epoch strength values are computed from the voiced speech only. The number of feature vectors obtained with 25-50 dimension may be comparatively less.

With large dimension and less number of feature vectors, speaker information may not be modeled well. Thus due to intra-speaker variability and poor modeling, matching may be difficult. For this reason we prefer to use lower dimension feature vectors. To select suitable dimension, we conduct a speaker identification study for different dimensions of pitch values for 30 speakers set collected (both train and test data) from NIST-99 database. In this experiment the feature vectors are made by the sequence of pitch values with a shift of one. The reason for considering every sample shift of the pitch values is to get maximum number of feature vectors. The result of this experiment is shown in Figure 6.2. With increase in the dimension from 1 to 10, the performance is increased. Any further increase in dimension results in decreasing performance. The reason may be that with increase in dimension, the intra-speaker variability may also be increased. We therefore use 10 pitch values with shift of one to represent pitch

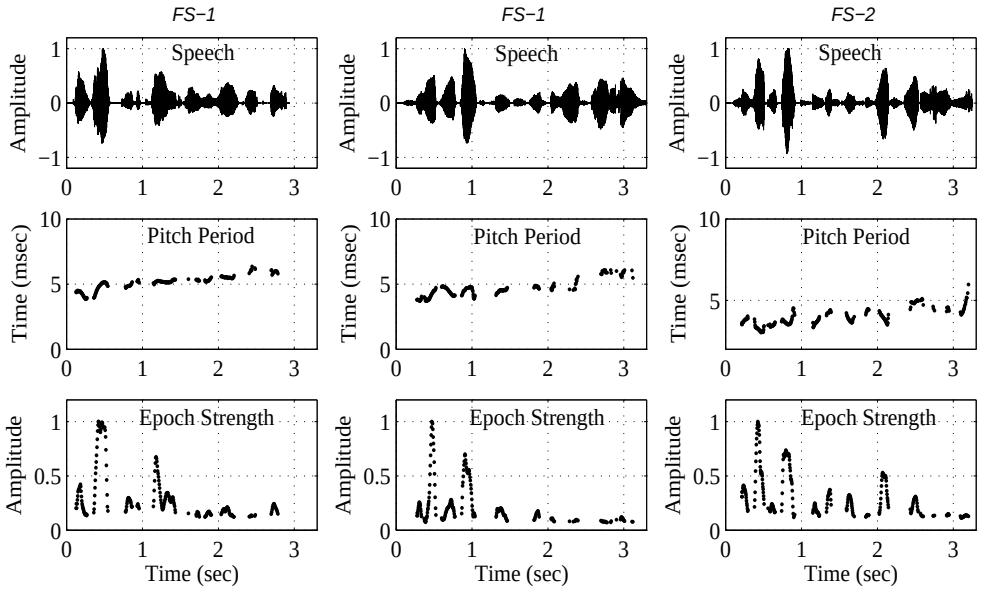

Figure 6.1: Examples of speech waveforms, pitch and epoch strength contours computed using *zero-frequency filtering* approach of two female speakers *FS-1* and *FS-2*. The text of the speech of both speakers is same. The first two columns represent two examples of speech waveforms and corresponding pitch and epoch strength contours of *FS-1* from two speech signals.

feature vectors. We call them as T_0 vectors. Similarly we use 10 epoch strength values with shift of one to represent epoch strength features. We call them as A_0 vectors.

The speaker recognition results of the proposed T_0 and A_0 vectors are given in the Table 6.1. The results show that for both speaker identification and verification tasks, the T_0 vectors provide relatively better performance. It indicates that the pitch information contains more speaker-specific evidence than the epoch strength information at the suprasegmental level. The identification performance of the T_0 vectors is significantly better and the verification performance of A_0 vectors is very poor. This shows that the speaker-specific evidence provided by the proposed T_0 vectors may have large inter-speaker variability. On the other hand, the speaker-specific evidence provided by the proposed A_0 vectors may have large intra-speaker variability.

Although, for both identification and verification tasks the proposed epoch strength vectors provide poor recognition performance, they may help in providing the additional speaker-specific evidence to pitch vectors. This is because T_0 and A_0 vectors represent different aspect

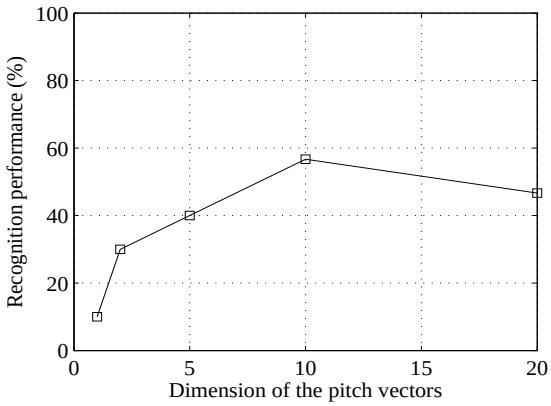

Figure 6.2: Speaker identification performance of pitch vectors for different dimension.

Table 6.1: Speaker recognition performance of T_0, A_0 vectors and their linear score level ($Comb_1$) combination scheme.

Feature		*Performance*		
		Identification(%)		*Verification*
		Set-1	*Set-2*	(*EER*)
T_0 vectors		29	18	45.39
A_0 vectors		9	7	49.27
$T_0 + A_0$	$Comb_1$	32	13	45.32

of suprasegmental excitation information. For example, the T_0 and A_0 vectors represent the temporal fluctuation in the rate and level of the vocal folds vibration, respectively. Further, the different nature of the speaker-specific information present in T_0 and A_0 vectors can also be observed from their respective confusion patterns and the 2-D score distribution diagrams. Figures 6.3(a) and (b) show the confusion patterns from the identification results of *Set-1* database using T_0 and A_0 vectors, respectively. The confusion patterns are entirely different. The decisions for both true and false identification are different. Similar observation can also be made from the verification results of T_0 and A_0 vectors. The 2-D log-likelihood score distribution diagram of T_0 and A_0 vectors is shown in Figure 6.4. In the regions *I* and *II* the T_0 and A_0 vectors give different decisions. These observations indicate that the speaker-specific information present in T_0 and A_0 vectors is different. This may help us in combining their individual

evidences to further improve the recognition performance.

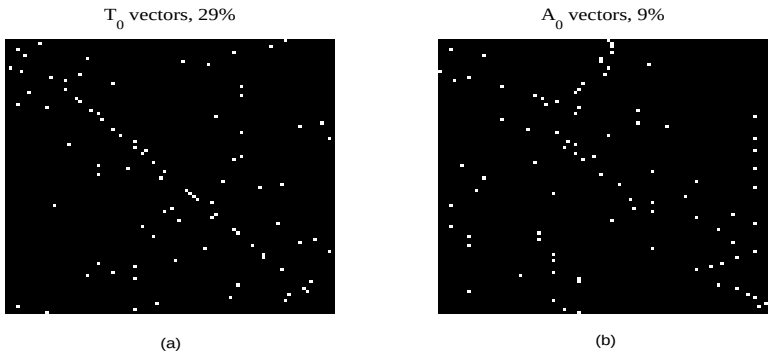

Figure 6.3: Confusion patterns from identification results for *Set-1* using pitch and epoch strength information represented by, (a) T_0 and, (b) A_0 vectors, respectively.

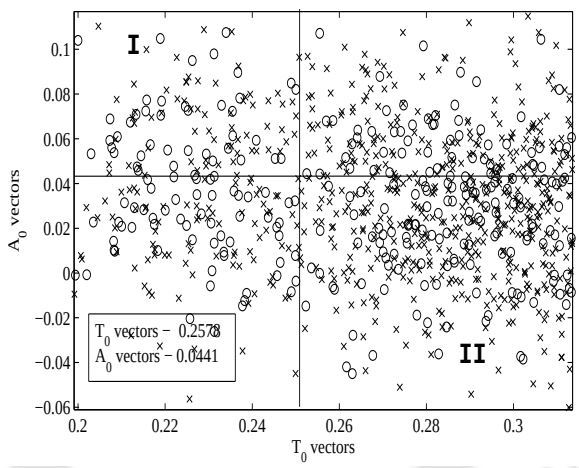

Figure 6.4: 2-D log-likelihood score distribution of genuine and imposter trails using pitch and epoch strength information represented by T_0 and A_0 vectors, respectively.

In combining the evidences from T_0 and A_0 vectors, we employ the $Comb_1$ schemes described earlier. In this work the combined evidence from T_0 and A_0 vectors is represented by $T_0 + A_0$. The speaker identification and verification results of $T_0 + A_0$ are given in the third row of the Table 6.1. It can be observed that the in most of the cases the best individual performance given by T_0 vectors is further improved. For example, the best identification accuracy that we can achieve for *Set-1* is 32%. For speaker verification task, the best EER we can achieve is 45.32%. Thus, we suggest that the suprasegmental pitch and epoch strength information can be represented by the combined representation of T_0 and A_0 vectors.

The combined use of the proposed pitch and epoch strength vector represents one aspect of the suprasegmental excitation information. In the next section we describe a method to model the suprasegmental cepstral trajectory information for speaker recognition.

6.3 Speaker-specific Information from LP Residual Cepstral Trajectories

The variation in the amplitude of the cepstral coefficient across several segments is called as cepstral trajectory [92]. In Chapter 5, we observe that the LP residual cepstral coefficients extracted using segments of 10-30 msec contain speaker-specific excitation information. Both the static and dynamic nature of speaker information are demonstrated. Like pitch and epoch strength contours, the LP residual cepstral trajectories may also contain speaker-specific information. The LP residual cepstral trajectories are spanned across several segments and hence the speaker-specific evidence in them may be viewed as the suprasegmental information. In this section, a method is developed to model the suprasegmental excitation information from few selected $RMFCC$ trajectories. First, based on the statistical *F-ratio* measure, few $RMFCC$ trajectories are selected and their speaker-specific nature is observed. The selected $RMFCC$ trajectories are processed in overlapping blocks to capture the speaker-specific information present in them. The significance of the suprasegmental excitation information present in the selected $RMFCC$ trajectories is demonstrated by the speaker identification and verification studies.

6.3.1 Speaker Information in RMFCC Trajectories

In the speaker recognition study, mostly 10-20 cepstral coefficients (excluding c_0) derived from the speech signal are typically chosen to represent the speaker-specific features [67]. In Chapter 4, we used first 13 cepstral coefficients excluding c_0 derived from the frequency warped LP residual spectrum to represent the segmental excitation information. We observe that the features represented by 13 $RMFCC$s are relatively more effective in capturing the speaker-specific information. The cepstral trajectories of all these $RMFCC$s can be used to model

the suprasegmental information. However, all these trajectories may not be equally useful for speaker recognition. In particular, the higher order cepstral trajectories may not contain much speaker-specific information [19, 67, 92]. Also the use of all these cepstral trajectories increases the computational complexity. Thus, we prefer to use the trajectory of the lesser (selective) number of *RMFCC*s to represent the suprasegmental information. To select the coefficients, statistical *F-ratio* measure that evaluates the discriminating ability may be used [86]. The *F-ratio* of a cepstral coefficient is defined as the ratio of its variance of means to the average intra-variance. Variance of means represents how the mean of a cepstral coefficient varies from speaker to speaker. Average intra-variance represents the variation of a cepstral coefficient within a speaker. An ideal cepstral coefficient should have large variance of means and small average intra-variance for discriminating speakers. *F-ratio* have been extensively used for measuring the discriminating ability of a feature and selecting optimized feature for speaker recognition [2, 23]. However, it should be noted here that cepstral coefficients with smaller *F-ratio* value may not imply that they are less effective in capturing the speaker information but may be redundant. Thus, when we purposefully want to select some few coefficients from a given set, *F-ratio* measurement may be a good measure for selection.

In selecting the *RMFCC*s by *F-ratio* measure, *Set-1* and *set-2* data sets are considered. The *RMFCC*s are computed as described in Chapter 4, Section 5.3. The *F-ratio* value of 13 individual *RMFCC*s for *Set-1* and *Set-2* data sets are given in the Table 6.2. It can be observed from third and sixth rows of this table that, the first five higher *F-ratio* value coefficients for both the data sets are from their first seven coefficients. For example, cepstral trajectories c_{t1}, c_{t2}, c_{t3}, c_{t4} and c_{t6} in case of *Set-1* and c_{t1} c_{t2}, c_{t4}, c_{t6} and c_{t7} for *Set-2*. The common higher *F-ratio* value cepstral coefficients in both the cases are c_{t1}, c_{t2}, c_{t4} and c_{t6}. Therefore, we consider the trajectory of only these four cepstral coefficients to represent the suprasegmental excitation information.

Figure 6.5 shows the example of c_{t1}, c_{t2}, c_{t4}, c_{t6} trajectories for *Speaker-1* and *Speaker-2*. In both cases, the text of the speech signal remains same. Any variations in the cepstral trajectories may be due to their speaker-dependant characteristics. It can be observed that in

Table 6.2: *F-ratio* value of $RMFCC$s (c_{t1}-c_{t13}) for *Set-1* and *Set-2*. The down arrow (↓) represents the arrangement of the cepstral coefficients in descending order.

Cepstral Coefficients	*Set-1*												
	c_{t1}	c_{t2}	c_{t3}	c_{t4}	c_{t5}	c_{t6}	c_{t7}	c_{t8}	c_{t9}	c_{t10}	c_{t11}	c_{t12}	c_{t13}
F-ratio	10.23	9.11	9.18	12.57	8.68	8.71	5.73	4.64	4.55	3.07	2.70	2.30	2.96
Order (↓)	c_{t4}	c_{t1}	c_{t3}	c_{t2}	c_{t6}	c_{t5}	c_{t7}	c_{t8}	c_{t9}	c_{t10}	c_{t13}	c_{t11}	c_{t12}
Cepstral Coefficients	*Set-2*												
	c_{t1}	c_{t2}	c_{t3}	c_{t4}	c_{t5}	c_{t6}	c_{t7}	c_{t8}	c_{t9}	c_{t10}	c_{t11}	c_{t12}	c_{t13}
F-ratio	7.58	6.65	5.97	6.79	5.54	7.17	6.98	5.37	4.81	5.93	5.75	4.91	3.13
Order (↓)	c_{t1}	c_{t6}	c_{t7}	c_{t4}	c_{t2}	c_{t3}	c_{t10}	c_{t11}	c_{t5}	c_{t8}	c_{t12}	c_{t9}	c_{t13}

each case, apart from their duration differences, the variation in the amplitudes of the sequence of cepstral trajectory samples are also significantly different across speakers. This shows that $RMFCC$ trajectories are speaker dependant and may be useful in modeling the suprasegmental information. This is indeed we observe from the speaker identification and verification studies made in the next section.

6.3.2 Speaker Recognition Studies using RMFCC Trajectories

So far we observed that $RMFCC$ trajectories contain speaker-specific evidence that corresponds to the suprasegmental excitation information and can be modeled by using features from c_{t1}, c_{t2}, c_{t4} and c_{t6} trajectories. To verify the significance of the suprasegmental $RMFCC$ trajectory information, we conduct both speaker identification and verification experiments. Like pitch and epoch strength vectors, the speaker-specific $RMFCC$ trajectory features are represented by overlapping blocks of 10 cepstral values. The sequence of 10 cepstral coefficients that span across 10 segments is considered to capture suprasegmental excitation information. In this case also, every sample shift is considered to get the maximum number of feature vectors. The feature vectors are derived from each chosen cepstral trajectory and modeled independently for speaker recognition studies.

The speaker identification and verification results of c_{t1}, c_{t2}, c_{t4} and c_{t6} trajectory feature vectors are given in the Table 6.3. The results show that each trajectory contains good amount

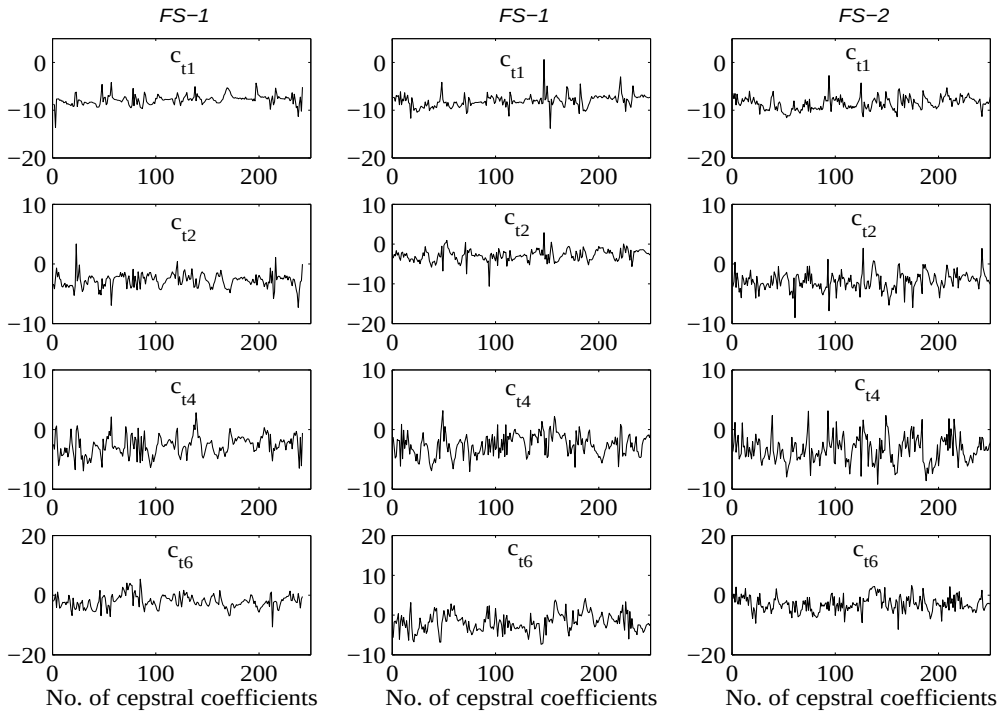

Figure 6.5: Examples of four $RMFCC$s $(c_{t1}, c_{t2}, c_{t4}, c_{t6})$ trajectories from two female speakers. The text of the speech of both speakers is same. The first two columns represent two examples of $RMFCC$s $(c_{t1}, c_{t2}, c_{t4}, c_{t6})$ trajectories of *FS-1* from two speech signals.

of speaker-specific excitation information. In case of more noisy speech the recognition performance of cepstral trajectory feature vectors is degraded. For example, the degradations in the identification performance from *Set-1* to *Set-2* for c_{t1}, c_{t2}, c_{t4} and c_{t6} trajectory feature vectors are 32%, 35%, 43% and 35%, respectively. This may be due to the fact that cepstral processing is mostly affected by noise. For both speaker identification and verification tasks, lower order cepstral trajectory vectors, for example c_{t1} and c_{t2} feature vectors are providing good recognition accuracy. Although, the computation of individual $RMFCC$s is contributed by all the subband energies, the low quefrency variations seem to be more speaker-specific.

Although, the higher order $RMFCC$ trajectory vectors are providing relatively poor performance, but they may reflect different speaker-specific evidence. The different nature of the speaker-specific information present in c_{t1}, c_{t2}, c_{t4} and c_{t6} trajectory feature vectors can be observed from their respective confusion patterns. The confusion patterns of the identification

Table 6.3: Speaker recognition performance of c_{t1}, c_{t2}, c_{t4}, c_{t6} trajectory vectors and their different combination. $C_t = c_{t1} + c_{t2} + c_{t4} + c_{t6}$.

Feature		*Performance(%)*		
		Identification		*Verification*
		Set-1	*Set-2*	
c_{t1}		40	27	42.63
c_{t2}		34	22	42.41
c_{t4}		21	12	43.27
c_{t6}		26	17	43.85
C_t	$Comb_1$	56	37	41.59

results of c_{t1}, c_{t2}, c_{t4} and c_{t6} trajectory feature vectors for *Set-1* dataset is shown in Figure 6.6. The confusion patterns of c_{t1}, c_{t2}, c_{t4} and c_{t6} trajectory feature vectors are different. They give different decisions for both true and false cases. This indicates that the c_{t1}, c_{t2}, c_{t4} and c_{t6} trajectory feature vectors reflect different aspect of speaker-specific information and can be combined to further improve the recognition accuracy.

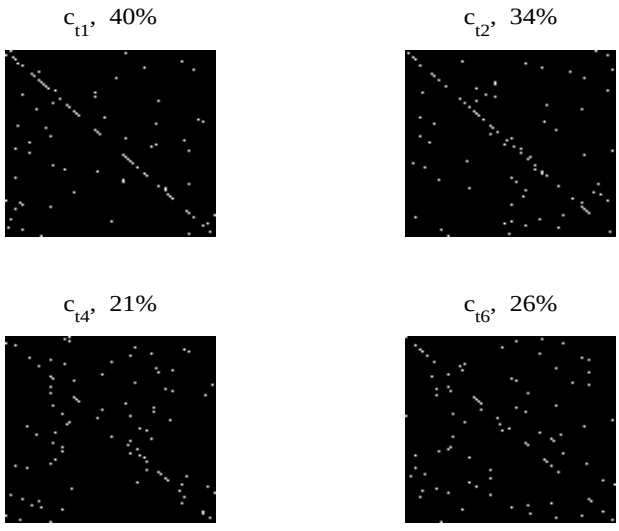

Figure 6.6: Confusion patterns from speaker identification results of *Set-1* dataset using c_{t1}, c_{t2}, c_{t4}, c_{t6} trajectory vectors.

In this work the representative feature for combined use of c_{t1}, c_{t2}, c_{t4} and c_{t6} feature

vectors is abbreviated as C_t. The speaker identification and verification performances of C_t for $Comb_1$ scheme are given in the fifth row of the Table 6.3. In case of *Set-1*, the best individual performance 40% by c_{t1} is improved to 56%. In case of *Set-2*, the best individual performance 27% by c_{t1} is improved to 37%. As expected, the relative improvement in the identification accuracy by C_t is more in case of clean speech. For example, the improvement in the identification accuracy by C_t for *Set-1* is around 40%, as against 37% in case of *Set-2* data set. In case of verification task, the best individual performance 42.41% by c_{t2} is improved to 41.59%. The improvement in the recognition accuracy by C_t for both verification and identification tasks indicates that the c_{t1}, c_{t2}, c_{t4} and c_{t6} feature vectors contain different aspect of suprasegmental information. From these observations we conclude that the suprasegmental $RMFCC$ trajectory information can be effectively represented by the combined representation of evidences from c_{t1}, c_{t2}, c_{t4} and c_{t6} vectors.

6.4 Speaker Information from Combined Pitch, Epoch Strength and RMFCC Trajectory Vectors

In this section, the suprasegmental excitation information is modeled explicitly by combining the evidences from pitch, epoch strength and $RMFCC$ trajectory vectors for speaker recognition. First, we demonstrate the different nature of speaker information present in pitch, epoch strength and cepstral trajectory vectors and then combine the respective evidences for explicit representation of the speaker-specific suprasegmental excitation information. The significance of the proposed representation is demonstrated by speaker identification and verification studies.

From the previous sections of this chapter we observe that the suprasegmental pitch and epoch strength and cepstral trajectory information can be effectively represented by $T_0 + A_0$ and C_t, respectively. By comparing their respective results from Tables 6.1 and 6.3, it can be observed that for both identification and verification, individually the performance of $T_0 + A_0$ is relatively poor than C_t. For more noisy speech the performance of both $T_0 + A_0$ and C_t is affected. However, the degradation in case of $T_0 + A_0$ is relatively more than C_t. For example,

in case of identification task, the performance of $T_0 + A_0$ and C_t feature vectors degrades by 59% and 34%, respectively. This may due to the large intra-speaker variability of $T_0 + A_0$ and also due to text-independent mode of operation. Further, $T_0 + A_0$ and C_t reflect different aspect of the suprasegmental excitation information. This can also be observed from their confusion patterns. Figure 6.7 shows the confusion patterns of $T_0 + A_0$ and C_t from the identification results of *Set-1* dataset. The confusion patterns are different. They give different decisions for both true and false identification. Thus, the evidence provided by $T_0 + A_0$ and C_t is different and hence may be combined to further improve the performance from suprasegmental excitation information perspective.

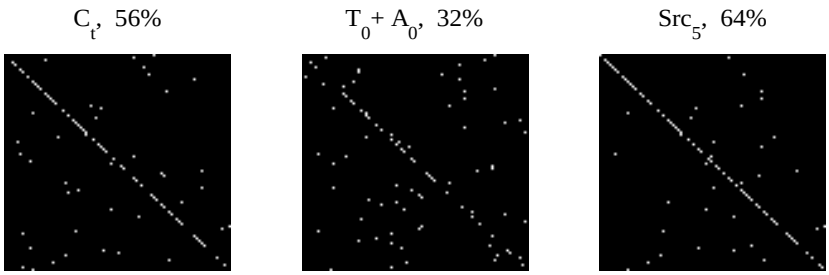

Figure 6.7: Confusion patterns from speaker identification results for *Set-1* dataset using C_t, $T_0 + A_0$ and their combination ($Comb_1$).

In combining the evidences from $T_0 + A_0$ and C_t we use linear $Comb_1$ scheme. In this work, the combined evidences from $T_0 + A_0$ and C_t is represented by Src_5. The speaker identification and verification results of Src_5 are given in the Table 6.4. As compared to $T_0 + A_0$ and C_t, the Src_5 provides improved performance for both identification and verification tasks. For more noisy speech the performance of Src_5 is less affected. For example, the relative degradation in the identification performance due to noise from *Set-1* to *Set-2* data sets is around 32%. This indicates that Src_5 provides relatively more robust speaker-specific evidence. Thus, we conclude that combined representation of cepstral trajectory, pitch and epoch strength vectors may be the best possible way of modeling of the suprasegmental excitation information.

Table 6.4: Speaker recognition performance of combined (linear combination $Comb_1$) pitch, epoch strength and cepstral trajectory vectors. $Src_5 = T_0 + A_0 + C_t$.

Feature		*Performance*		
		Identification(%)		*Verification*
		Set-1	*Set-2*	(*EER*)
Src_5	$Comb_1$	64	43	39.47

6.5 Comparison of Explicit and Implicit Modeling of Suprasegmental Speaker Information

In Chapter 3, the suprasegmental excitation information is modeled by processing the LP residual directly in the time domain without extracting any feature and hence corresponds to the implicit modeling. Alternatively, in this chapter we propose a method to model the suprasegmental excitation information explicitly by deriving the speaker-specific parameters from the LP residual and hence corresponds to explicit modeling. In this section we make a comparative study between the implicit and explicit approaches to model the suprasegmental excitation information. The comparison is made based on computational complexity, recognition accuracy and in providing the different evidence to other levels of speaker information, namely, subsegmental and segmental of excitation, and vocal tract.

In implicit modeling of the suprasegmental excitation information, the LP residual is decimated by a factor of 50 and processed in blocks of 250 msec with a shift of 6.25 msec. For a given 2 min speech signal samples at 8 kHz, the number of suprasegmental LP residual blocks may be around 19,000 and the dimension of each block is 40. On the other hand, for the same speech signal the number of pitch, epoch strength and cepstral trajectory vectors will be very less. Because, the pitch and epoch strength vectors are computed only from the voiced part of the speech signal and the *RMFCC* trajectory vectors from 20 msec segment of speech and their dimension is 10. Further, there is no additional computation required for *RMFCC*, as it is already used in modeling the segmental excitation information. Therefore, it seems that relatively more computational complexity is involved in implicit modeling of the suprasegmental

excitation information.

By comparing the performance of the *Supra* feature with the Src_5 from Table 6.5, it can be observed that the later provides significantly better performance. The performance achieved by Src_5 is at par with the subsegmental and segmental levels speaker-specific excitation information. Thus, it is possible to achieve the better recognition accuracy from the suprasegmental excitation information alone by using the proposed representation. Further, the relative degradation in the performance due to noise in case of Src_5 is less. For example, the relative degradation in the identification performance of Src_5 from *Set-1* to *Set-2* is 32%, as compared to 58% in case of *Supra* feature. This shows that the proposed explicit representation of the suprasegmental excitation information provides good and robust recognition accuracy than the *Supra* feature.

To verify the significance of the explicit approach in providing the additional information to other levels, we combine the evidences from subsegmental and segmental level with Src_5. The speaker identification and verification results of combined subsegmental, segmental and Src_5 are given in seventh row of the Table 6.5. By comparing the results of $Sub + Seg + Src_5$ with Src_2 it can be observed that in all cases the proposed feature provides relatively good performance. The proposed Src_5 representation also combines well with the vocal tract features. In case of identification task the $Sub + Seg + Src_5 + MFCC$ provides 87% and 66%, as against 87% and 70% by $Src_2 + MFCC$ for *Set-1* and *Set-2* datasets, respectively. As, mentioned earlier, the poor performance in case of *Set-2* may be due to the combination schemes employed. Because, $Sub + Seg + Src_5 + MFCC$ provides improved verification performance of 22.46% as compared to $Src_2 + MFCC$ of 22.99%. These results show that Src_5 provides relatively more different evidence to other levels to achieve improved performance.

The above observations indicate that the proposed explicit approach of modeling the suprasegmental excitation information is relatively more compact, robust against noise and provide more additional information to other levels of excitation and vocal tract for speaker recognition. Thus we conclude that the combined representation of pitch, epoch strength and cepstral trajectory vectors may be the best possible way of modeling the suprasegmental excitation information

Table 6.5: Speaker recognition performances of suprasegmental excitation information using explicit and implicit modelling approaches. $Comb_1$ represents linear combination scheme. $Src_5 = T_0 + A_0 + C_t$. Src_2 represents combined ($Comb_1$) representation of subsegmental (Sub), segmental (Seg) and suprasegmental ($Supra$) excitation information from time domain processing of the LP residual.

Feature		$Performance(\%)$		
		$Identification(\%)$		$Verification$
		Set-1	*Set-2*	(EER)
$MFCC$		87	66	22.94
$Supra$		31	13	44.49
Src_5	$Comb_1$	64	43	39.47
$Src_5 + MFCC$	$Comb_1$	89	66	22.94
$Sub + Seg$	$Comb_1$	64	60	32.02
Src_2	$Comb_1$	74	58	33.28
$Sub + Seg + Src_5$	$Comb_1$	74	61	31.39
$Src_2 + MFCC$	$Comb_1$	87	70	22.99
$Sub + Seg + Src_5 + MFCC$	$Comb_1$	91	66	22.46

for speaker recognition.

6.6 Summary

In this chapter a combined method is proposed for explicit modeling of the suprasegmental excitation information for speaker recognition. In the proposed approach, the suprasegmental pitch and epoch strength information is modeled by processing the respective contours in overlapping blocks. The pitch and epoch strength contours are computed by using the modified zero-frequency filtering approach that is more effective for telephone speech. In the modified approach the HE of the LP residual is used as the input to the zero-frequency filter. The suprasegmental cepstral trajectory information is modeled by processing the few selected $RMFCC$ trajectories in blocks. Results from speaker recognition experiments show that the proposed pitch, epoch strength and cepstral trajectory feature vectors well capture the suprasegmental excitation information. The suprasegmental cepstral trajectory information is relatively more robust and provides better recognition accuracy for both speaker identification and verification tasks. Pitch, epoch strength contours and $RMFCC$ trajectories reflect different aspect

of the speaker information and combine well to further improve the recognition performance. A comparative study made between the proposed approach and the implicit modeling of the suprasegmental excitation information revealed that the speaker-specific information present in the combined representation of pitch, epoch strength and cepstral trajectory vectors is relatively more compact, robust and well combine with other levels of excitation and vocal tract information. Thus, it is suggested that the combined use of pitch, epoch strength and cepstral trajectory vectors may be the best possible way of representing the suprasegmental excitation information for speaker recognition.

So far we observe that, the speaker recognition performance achieved by the improved representation of the excitation information is still poorer than the conventional vocal tract information. The studies so far made are on the baseline system. State-of-the-art approaches like, post feature processing, GMM-UBM modeling technique have not been used. It is expected that by using all these techniques the speaker recognition performance from excitation perspective may be improved further. Then a comparative study with the conventional vocal tract features may enable us to verify the real potential of the excitation information for speaker recognition. This aspect is explored in the next chapter.

7

Speaker Verification using Excitation Information

Contents

In this chapter, we develop a speaker verification system based on the excitation information and demonstrate its significance by comparing with the corresponding vocal tract information based system. The speaker-specific excitation information is extracted from the subsegmental, segmental and suprasegmental levels processing of the LP residual, using possible approach developed in the earlier chapters. Post feature processing like cepstral mean subtraction (CMS) is employed to achieve robustness against channel variations. The speaker-specific information from the subsegmental, segmental and suprasegmental levels of the excitation signal is modeled independently using GMM-UBM modeling technique. The speaker-specific evidence from each of these levels are combined at the score level. The significance of the proposed speaker recognition system is demonstrated by conducting speaker verification experiments on the whole NIST-03 database. For each trail two different tests, termed as *Clean test* and *Noisy test* are conducted. For both tests, common speaker models are used, built using the speech signal available for training. At the time of testing, in case of *Clean test*, the speech signal available for testing is used as it is for verification. In case of *Noisy test*, the test speech is corrupted by factory noise (9 dB) and then used for verification. This experiment is performed to demonstrate the robustness of the proposed excitation information based system. The speaker verification results show that for *Clean test* case, the proposed source based speaker recognition system provides relatively poor performance than the conventional vocal tract information. On the other hand, for *Noisy test* case, the proposed system provides relatively better performance than the conventional vocal tract system. For both clean and noisy cases, by providing different and robust speaker-specific evidences, the proposed system helps the conventional vocal tract system to further improve the overall recognition performance. We also observed that the real potential of the excitation information depends upon how well the speaker-specific evidence from different levels of the excitation signal are combined. If we have a suitable combination scheme, then it is indeed possible to achieve better performance from the excitation itself.

7.1 Introduction

The function of the SR system at different stages described in Chapter 1 shows that the success rate mostly depends upon extracting and then modeling the speaker-dependant characteristics of the speech signal. In this work, we assume that the well established modeling techniques like GMM and GMM-UBM are general enough for building the speaker models and focus on the feature extraction stage. As mentioned in the introduction chapter, state-of-the-art speaker recognition system that uses the vocal tract information represented by $MFCC$ feature has suffered from certain limitations. The most important among them is that the standard $MFCC$ feature depends upon the quality and quantity of the data [29]. On the other hand, the excitation information extracted from the LP residual is relatively more robust and require less amount of data for speaker recognition [12]. Hence, the work presented in this thesis concentrated on exploring the different methods for the extraction of the excitation information for speaker recognition. The results of these explorations revealed that the excitation information can indeed be effectively extracted by processing the LP residual at subsegmental, segmental and suprasegmental levels. The speaker-specific information present at each of these levels of the LP residual can be modeled either by implicit or by explicit approaches. As demonstrated in earlier chapters, the later approach is relatively more compact, robust and provides improved speaker recognition recognition accuracy by combining evidence from subsegmental, segmental and suprasegmental levels. Individually, the $GFD + \Delta + \Delta\Delta$, $RMFCC + MPDSS$ and $T_0 + A_0 + C_t$ features are the possible ways of representing the subsegmental, segmental and suprasegmental excitation information, respectively.

The objective of the work presented in this chapter is to develop a speaker recognition system based on the excitation information. The literature review made in the Chapter 2 revealed that there are several attempts made in developing the speaker recognition system using the excitation information and demonstrated their significance [9, 10, 16, 29]. The major limitations of these attempts are that they use the then available methods for feature extraction and building the speaker models. For example, in the feature extraction stage, use of the

recently proposed zero-frequency filtering approach is found to be more effective for modeling the excitation pitch and epoch strength information (Chapter 6). Post feature extraction processing have not been applied. It is well known fact the cepstral features are severely affected by the channel characteristics [7]. Since, the $RMFCC$ features correspond to cepstral features, they are also expected to be affected by channel effect. By performing the cepstral mean substraction (CMS) and also by including their dynamic information, the performance may further be improved. On the similar lines, for building the speaker models, the modeling technique like GMM-UBM have not be used, except the time frequency feature like $WOCOR$ (Table 2.3). The $WOCOR$ features are extracted from 30 msec blocks of the LP residual that correspond to segmental level information and hence may not represent the complete excitation information. This may be the reason for which the performance achieved by the $WOCOR$ feature is relatively poor than the vocal tract information. It is expected that by incorporating the best possible feature extraction methods and incorporating the state-of-the-art techniques for each level of the excitation information, we may achieve good recognition accuracy from the excitation information perspective. Hence, the motivation for the present work.

In this work, to develop a speaker recognition system using the excitation information, the possible approaches as suggested so far in this thesis work are used for the extraction of speaker-specific features from the subsegmental, segmental and suprasegmental levels excitation information. The CMS is performed to eliminate the channel effect in case of the $RMFCC$ feature. It is already shown that the dynamics of the cepstral coefficients contains speaker-specific information [3, 33]. The dynamic information of the $RMFCC$ feature is included by concatenating their Δ and $\Delta\Delta$ values. The GMM-UBM modeling technique is used for building the speaker models. The significance of the proposed system is demonstrated by the speaker verification experiments. Towards the end of this chapter, the speaker verification performance of the state-of-the-art vocal tract features on the similar experimental conditions is evaluated and a comparative study is made with the developed system.

The rest of this chapter is organized as follows: Section 7.2 describes the basic block diagram of the developed excitation information based SR system. In this section significance of the

developed system is demonstrated by conducting different speaker verification experiments. In Section 7.3, the speaker verification performance of the vocal tract features is evaluated on similar experimental condition and comparative study is made with the proposed system. The last section summarizes the present work described in this chapter.

7.2 SR System using excitation Information

In this section, first we describe the block diagram of the proposed excitation information based speaker recognition system. The developed system uses the post feature extraction processing like, temporal dynamics and channel compensation of the cepstral features. GMM-UBM modeling technique is employed for building the speaker models. The potential of the proposed system is demonstrated by speaker verification studies. We perform the verification experiments on clean and noisy cases to verify the robustness of the proposed system.

7.2.1 Block Diagram of the Proposed Speaker Recognition System

The block diagram of the proposed speaker recognition system is shown in Figure 7.1. The function of each block is similar to the one described in the introduction section, except the use of the recent approaches at each level. The feature extraction block for both training and testing phases is shown as a common block. In this block diagram the actual input to the feature extraction stage during training and testing phase is the LP residual computed from the train and test speech signals, respectively. Hence, the system is called as the excitation based speaker recognition system. Due to simple and computationally less intensive, the LP residual is computed by the autocorrelation method [1].

In the feature extraction stage, the $GFD + \Delta + \Delta\Delta$ is extracted as described in Section 4.5, and used as the feature to represent the subsegmental level excitation information. For segmental level excitation information, the $MPDSS$ and $RMFCC$ feature are extracted as described in Section 5.2.2 and 5.3, respectively. In case of $RMFCC$ feature, the CMS is applied to remove the channel effect. It should be noted here that although the CMS reduces some channel effect, but also eliminates some speaker information [7]. In applications where the speech data

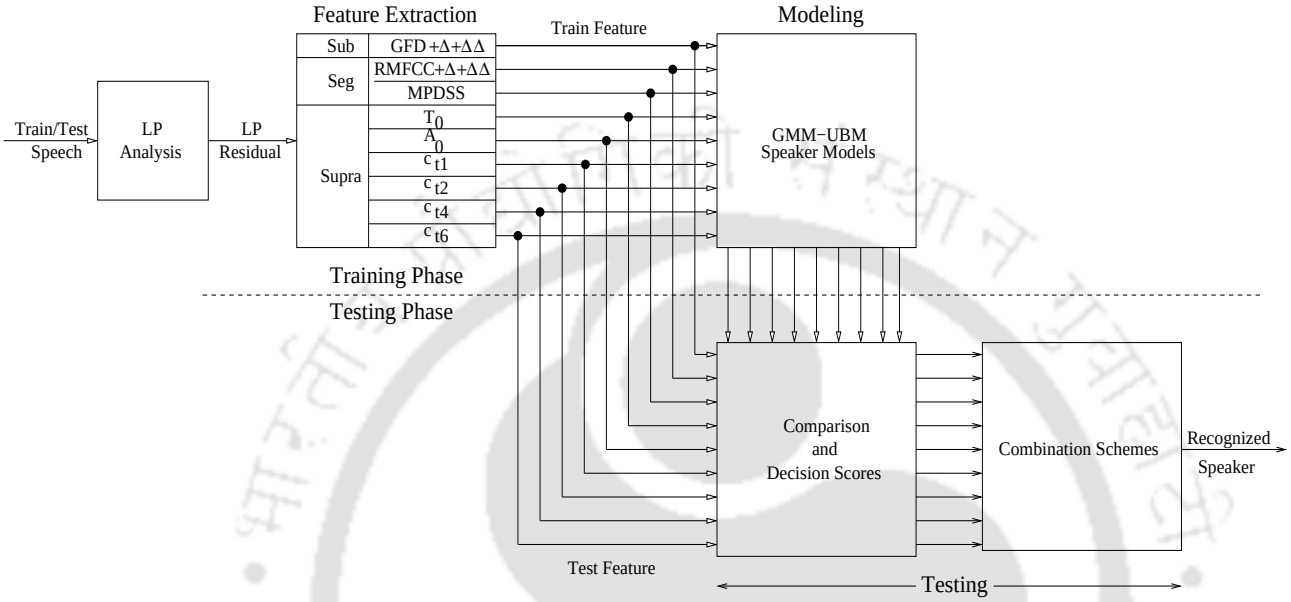

Figure 7.1: Block diagram of the proposed speaker recognition system using excitation information.

available for training and testing are collected in clean condition, the CMS may not be useful. The $RMFCC$ is concatenated with its Δ and $\Delta\Delta$ values to incorporate the temporal dynamic information of segmental excitation energy. In this work they are called as $RMFCC + \Delta + \Delta\Delta$ feature. The $MPDSS$ and $RMFCC + \Delta + \Delta\Delta$ features are used independently for building the speaker models. Their individual evidences are added to represent improved segmental excitation information. In this work the combination of $RMFCC + \Delta + \Delta\Delta$ and $MPDSS$ feature to represent the segmental level information is abbreviated as Src_6. For building the speaker models from the suprasegmental pitch and epoch strength contours excitation information, the T_0 and A_0 vectors are computed as described in Section 6.2.1. For building the speaker models from the suprasegmental level cepstral trajectory excitation information, c_{t1}, c_{t2}, c_{t4} and c_{t6} trajectory vectors are extracted from $RMFCC$ as described in Section 6.3. The pitch, epoch strength and the cepstral trajectory vectors are used independently for building the speaker models. The complete suprasegmental excitation information is represented by Src_5. The combination of speaker-specific evidence from $GFD + \Delta + \Delta\Delta$ feature, Src_6 and

Src_5 is abbreviated as Src. The Src represents the excitation information for the proposed speaker verification system.

For building the speaker models, GMM-UBM modeling technique is employed. The UBM is a GMM built on large amount of data collected from several speakers, mostly not included in the target set [35]. The UBM represents speakers independent distribution of the feature vectors and serves as the average speaker model and also can be used as an imposter model for speaker verification task [29]. In [35], it was shown that it is advantageous to train two separate background models, one for male and the other for female speakers, and then build the single UBM by pooling the two models. The target speaker GMM model is then adapted by the single UBM. In practice, adapting the means only is found to work well than adapting all the three parameters [35,36,65]. Therefore, in most of the existing speaker recognition systems, only the means are adapted, and the weights and variances of the speaker model remain unaltered. The adaption technique is described in *Appendix C*. For each target speaker, separate models are built using different features. For example, in the proposed system, each target speaker has nine separate target and background models.

At the time of testing, the features are extracted from the input test speech signal similar to the training phase and compared with respective feature target model. The decision is taken based on the LLR scores assigned to the corresponding target models. The LLR that depends on both the target model and background model is given by

$$LLR = logP(\lambda_c) - logP(\lambda_u) \tag{7.1}$$

where, $P(\lambda_c)$ and $P(\lambda_u)$ are the likelihoods given by the claimed speaker model and the UBM, respectively.

In the combination stage, the speaker-specific evidence from different excitation features are combined to achieve the improved recognition performance. Combining evidences from the multiple levels may be broadly grouped into three categories, namely, abstract level, rank level and the measurement level combination [74]. Majority of the speaker recognition systems that use information from multiple levels mostly use the measurement level combination [74,95].

This is because, among the three levels of combination schemes, the measurement level contains the maximum amount of information and the abstract level contains the least [95]. Further, whenever the recognition performance from different levels not at par with each other, the abstract and rank levels combination schemes give relatively poor performance [95]. This is indeed the case for different excitation features, as we observe from the results presented in previous chapters. Thus, in this work we prefer to use the measurement level combination schemes for combining the speaker-specific evidence from different excitation features. Of course, the limitation of measurement level is also due to the poor performing system and is handled as described next.

As mentioned in the introduction section, the excitation information is extracted from subsegmental, segmental and suprasegmental levels. In some levels like segmental and suprasegmental levels, different features are combined to represent their respective complete information. So, to model the complete excitation information, there are two approaches that may be used for combining the evidences from all the three levels. In one approach, the speaker-specific evidence from all the features from all levels can be combined. Alternatively, feature from individual levels can be combined first and then then these combined evidences can be further combined to obtain the overall excitation information evidence. The later approach seem to give better performance when some of the systems are performing poor. Because, from the previous experimental results presented in this thesis, we can observe that the performance of the individual features are significantly different. For example, the verification performance of the $MPDSS$ and A_0 feature vectors for whole NIST-03 database on a GMM based speaker recognition system is 32.25% and 49.25%, respectively. Due to the large difference in the performance, the weighting of the LLR of the poor feature may further reduce its significance. As a result, the effectiveness of the poor performing feature may be diminished by good performing features. On the other hand, in the similar condition, the best possible speaker verification performances achieved by subsegmental, segmental and suprasegmental levels excitation information are 39.79%, 32.33% and 39.47%, respectively. They are almost at par with each other, so their respective weighting factors may be proportional. As a result, the effectiveness of the

speaker-specific information from each level of the excitation can be properly manifested in the combined representation. So we suggest that, first the combination of the evidences may be at segmental and suprasegmental levels. The combined evidence from each of these levels may further be combined with the subsegmental level to represent improved excitation information. By this combination scheme, it is expected that we may achieve better recognition performance from the proposed speaker recognition system.

7.2.2 Speaker Verification Studies using Excitation Source Features

To demonstrate the significance of the proposed excitation information based system, we conduct speaker verification experiments on the whole NIST-03 database [58]. The verification performance is evaluated for individual excitation features from subsegmental, segmental, suprasegmental levels, the complete evidence from each level and then finally the complete excitation information under clean and noisy conditions. A comparative study is also made on the effectiveness of the different excitation features for speaker verification task on clean and noisy cases.

The speech signal available in the NIST-03 database is sampled at 8 kHz [58]. Thus, for LP residual computation we choose the prediction order as 10 to best represent the speaker-specific excitation information [12]. For training and testing, the features are extracted from the LP residual as mentioned earlier. Equations (4.21) and (4.22) are used to compute the Δ and $\Delta\Delta$ values from $RMFCC$ [3, 7]. By setting the energy threshold, the residual cepstral features corresponding to the voiced frames are considered. The selected cepstral features are subjected to CMS for eliminating the channel effect [7].

For building the speaker independent background model, the UBM is built from approximately forty hours speech data of 200 speakers that include 100 males and 100 females, collected from the switchboard database [96]. These speakers are not included in the NIST-03 evaluation set. As mentioned earlier, two separate male and female background models using 512 Gaussian mixtures are made. The two independent models are pooled together to form a 1024 single UBM model. The number of iterations for the expectation maximization is chosen as 10 [4].

In this work, only the means are adapted and the weights and variances of the speaker models and the UBM remain same [35, 36, 65].

In case of combining the speaker-specific evidence from different features, the scalar value assigned to a model (for example the LLR in the present case) from different features is considered as their respective speaker-specific evidence. For combination, we use the linear score level combination ($Comb_1$) scheme.

Each test is performed on two cases. The first case refers to *Clean* test, where the test signal is directly used for the verification. The second case refers to *Noisy* test, where we add the factory noise with SNR around 9 dB to the test signal and then used for the verification. The objective of doing the *Noisy* test is to verify the robustness of the features against noise with respect to the vocal tract features for speaker verification task. In both test cases, the experimental conditions remain same for fair comparison. The speaker verification performance is given by the DET curve based on genuine and imposter LLRs and the EER [40]. The DET curve is a two dimensional graph plotting the false alarm probability verses miss probability [40]. The horizontal axis represents the false alarm probability and the vertical axis represents the miss probability. For an ideal feature, both the false alarm probability and the miss probability should be zero or as minimum as possible. Further, a feature with less false alarm probability may have more inter-speaker variability. On the other hand a feature with less miss probability may have less intra-speaker variability. A feature with more inter-speaker variability and less intra-speaker variability is expected to give minimum false alarm and miss probabilities and in turn good EER.

The DET curves of the speaker verification results of all the excitation features for both *Clean* and *Noisy* test cases are shown in the Figures 7.2 and 7.3, respectively. Their corresponding EER are given in second and third columns of the Table 7.1, respectively. The results show that individually the features from the parametric representation of the segmental excitation information are providing good recognition performance. For example, for *Clean* test case the $RMFCC + \Delta + \Delta\Delta$ provides the best performance of 18.33% followed by $MPDSS$. In case of *Noisy* test cases, the $GFD + \Delta + \Delta\Delta$ feature provides the best performance of

31.31%. In case of *Noisy* test case, relatively poor performance of $RMFCC + \Delta + \Delta\Delta$ features may be due to the effect of the noise on the LP residual magnitude spectrum. The features from the parametric representation of the suprasegmental excitation information provide the least verification performance. For example, for *Clean* and *Noisy* test cases, the A_0 and c_{t6} vectors provide the least performance of 44.26% and 48.33%, respectively. As mentioned earlier, the speaker-specific evidence present in the suprasegmental excitation features may have large intra-speaker variability. This can also be observed from their respective DET curves for *Clean* and *Noisy* test cases shown in Figures 7.2 and 7.3, respectively. The A_0 and c_{t6} vectors provide relatively more miss probability than the $RMFCC + \Delta + \Delta\Delta$ and $MPDSS$ features for *Clean* test case. In *Noisy* test case, although they all provide higher miss and false alarm probabilities, but the minimum miss probability in case of A_0 and c_{t6} vectors is relatively more than the $RMFCC + \Delta + \Delta\Delta$ and $MPDSS$ features. This shows that the suprasegmental excitation information may have large intra-speaker variability.

Table 7.1: Speaker verification performances (EER) of different excitation features for *Clean* and *Noisy* test cases using GMM-UBM modeling technique. $Src_6 = RMFCC + \Delta + \Delta\Delta + MPDSS$. $C_t = c_{t1} + c_{t2} + c_{t4} + c_{t6}$. $Src_5 = T_0 + A_0 + C_t$. Src represents combination of subsegmental, segmental and suprasegmental excitation information extracted using respective proposed explicit approaches. $Comb_1$ represents linear score level combination schemes.

$Feature$		$Performance(EER)$		$Relative$
		$Clean$	$Noisy$	$Degradation(\%)$
$GFD + \Delta + \Delta\Delta$		20.28	31.31	54
$MPDSS$		18.65	35.99	93
$RMFCC + \Delta + \Delta\Delta$		18.33	39.34	115
T_0		33.73	40.74	21
A_0		44.26	46.12	4
c_{t1}		31.39	47.11	50
c_{t2}		31.21	47.24	51
c_{t4}		32.02	47.47	48
c_{t6}		32.83	48.33	47
Src_6	$Comb_1$	16.39	34.41	110
C_t	$Comb_1$	26.73	46.97	76
Src_5	$Comb_1$	25.15	40.69	62
Src	$Comb_1$	14.13	30.85	118

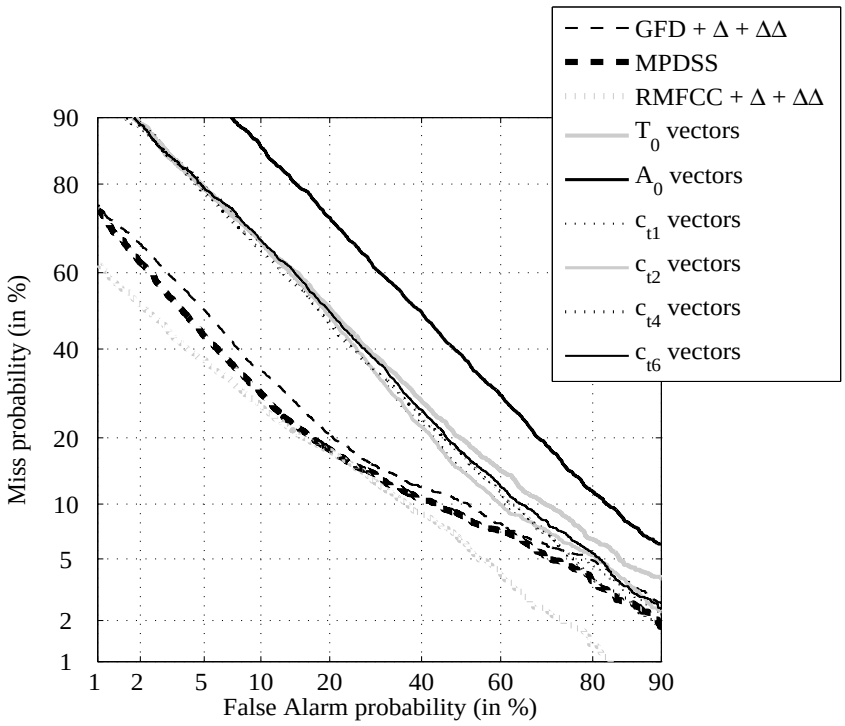

Figure 7.2: DET curves from the speaker verification experiments using different excitation features for *Clean* test case.

The DET curves of the speaker verification results from the subsegmental, segmental and suprasegmental levels of excitation information and their combined evidence by $Comb_1$ scheme for *Clean* and *Noisy* test cases are shown in Figures 7.4 and 7.5, respectively. Their corresponding EER are given in first, tenth and twelveth rows of the Table 7.1, respectively. It should be noted here that the subsegmental level excitation information is effectively represented by $GFD + \Delta + \Delta\Delta$ feature (Chapter 4). The DET curve and the EER of the subsegmental excitation information is same as the $GFD + \Delta + \Delta\Delta$ feature. The EER achieved by subsegmental, segmental and suprasegmental level information with $Comb_1$ scheme for *Clean* test case are 20.28%, 16.39% and 25.15%, and for *Noisy* test case 31.31%, 34.41% and 40.69%, respectively. These results show that for the *Clean* test case, the segmental excitation information provides the best performance followed by the subsegmental level. It can be observed from the Figure 7.4 that, the good performance in case of the segmental excitation information may be due to less miss probability. This indicates that the speaker-specific evidence provided by the segmental

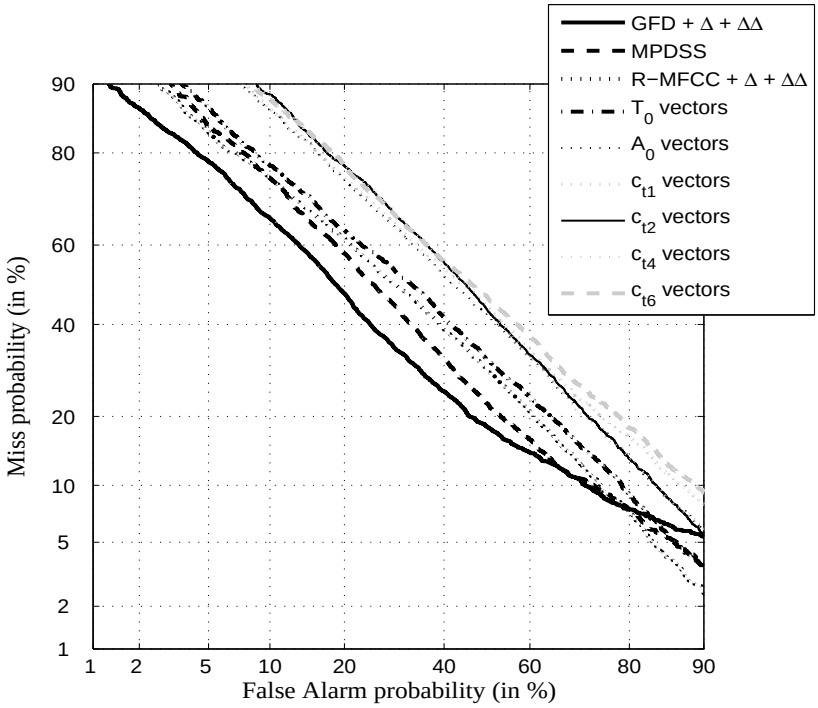

Figure 7.3: DET curves from the speaker verification experiments using different excitation features for *Noisy* test case.

excitation features have less intra-speaker variability.

For the *Noisy* test case, the subsegmental excitation information provides relatively good performance followed by the segmental level. The segmental excitation information is extracted from the LP residual magnitude spectrum. The spectrum is severely affected when the signal is corrupted by noise. For both test cases the suprasegmental excitation information provides the least performance. As mentioned earlier, the poor performance in case of combined evidences from the suprasegmental level may be due large intra-speaker variability. In combing the evidence from subsegmental, segmental and suprasegmental levels, for both *Clean* and *Noisy* test cases the verification performance is further improved. The EER achieved by *Src* for *Clean* and *Noisy* test cases from $Comb_1$ scheme are 14.13% and 30.85%, respectively.

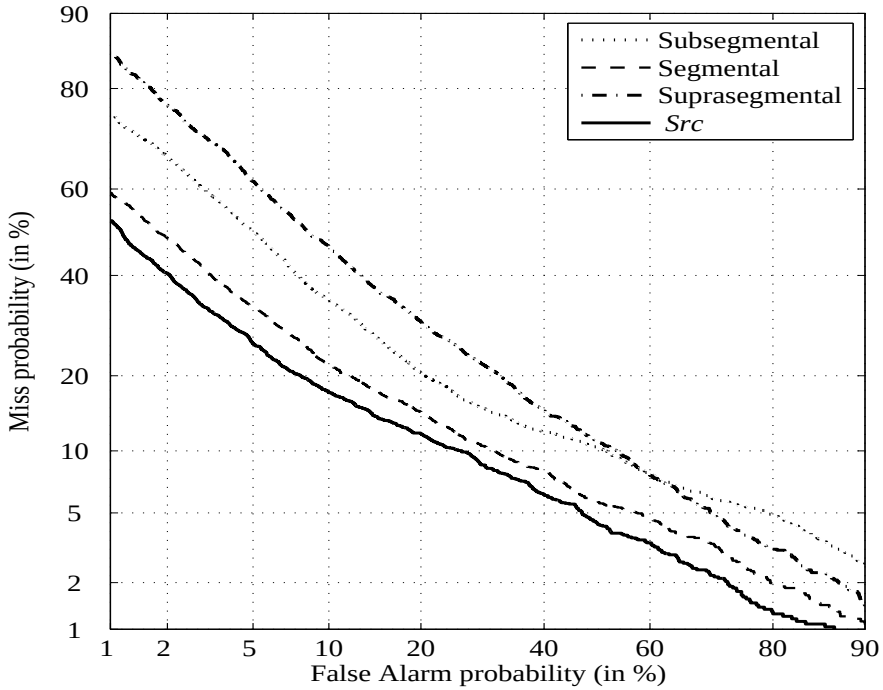

Figure 7.4: DET curves from the speaker verification experiments using the evidence from subsegmental, segmental, suprasegmental excitation information and their combination by $Comb_1$ scheme for *Clean* test case.

7.2.3 Effect of Noise on Excitation Source Features

To verify the effect of noise on the recognition performance of the excitation features, we compare their respective EER and relative degradation measure. The relative degradation in the EER for all cases are given in the last column of the Table 7.1. It can be observed that, in all cases the EER is increased when the test signal is corrupted by noise. This indicates that the excitation features are also affected by noise. Individually, the most severely affected feature is $RMFCC + \Delta + \Delta\Delta$ followed by $MPDSS$. In these cases the relative degradation in the performance are around 115% and 93%, respectively. The $RMFCC + \Delta + \Delta\Delta$ and $MPDSS$ feature are extracted from the LP residual spectrum. The spectrum of a signal is severely affected when the high factory noise is added to it [7, 97]. This may be the reason for severe degradation in the recognition performance of the $RMFCC + \Delta + \Delta\Delta$ and $MPDSS$ features for the *Noisy* test case. Since, both $RMFCC + \Delta + \Delta\Delta$ and $MPDSS$ features are

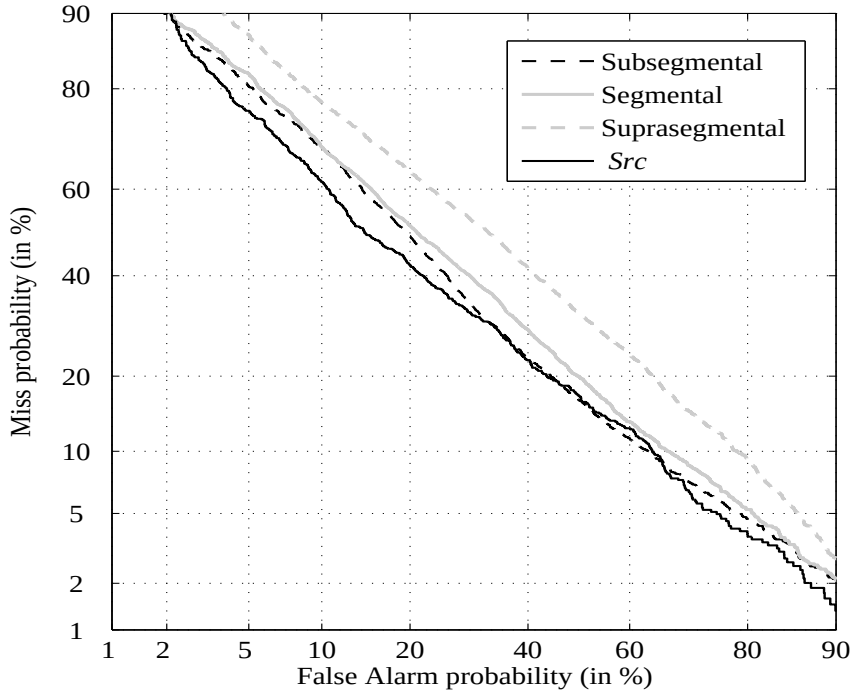

Figure 7.5: DET curves from the speaker verification experiments using the evidence from subsegmental, segmental, suprasegmental excitation information and their combination by $Comb_1$ scheme for *Noisy* test case.

affected by noise, the performance of the evidence from their combination representing the segmental level information is also affected maximum. The maximum relative degradation in the EER in case of subsegmental, segmental and suprasegmental excitation information are 54%, 110% and 62%, respectively. When the individual speaker-specific evidence from all the three levels are combined, the relative degradation in the performance is 118%. Moreover, the benefit we achieve by combining the evidence from all the three levels is relatively poor in case of *Noisy* test case. For example, for *Clean* test case the maximum individual of 18.33% by $RMFCC + \Delta + \Delta\Delta$ feature is reduced to 14.13% by *Src*. A relative improvement of 23% is achieved. On the other hand for the noisy case, the maximum individual performance of 31.33% by $GFD + \Delta + \Delta\Delta$ feature is reduced to 30.85% by *Src*. A relative improvement of 2% is achieved. These results show that the speaker verification performance is affected by noise when the evidences from the multiple levels of the excitation signal are combined. As

mentioned earlier, the reason may be due to the combination scheme employed. With suitable combination scheme, the EER may further be improved and the relative degradation may be decreased. Although, the speaker verification performance is affected by noise when evidences are combined from multiple levels, but we achieve relatively improved recognition accuracy. This suggests that the proposed SR system may be effective even in noisy case also.

7.3 Comparison of Excitation Source and Vocal Tract SR Systems

In this section, to demonstrate the significance of the proposed speaker recognition system based on excitation information, a comparative study is made with the corresponding vocal tract information based system for the verification task. First, we evaluate the verification performance of the vocal tract features for whole NIST-03 database on *Clean* and *Noisy* test cases. The comparison is made based on the verification performance and the robustness against noise.

7.3.1 Speaker Verification Studies using Vocal tract Features

State-of-the-art SR system based on the vocal tract information mostly use $MFCC$ together with their temporal dynamic information as the speaker-specific features. The specification and the computational procedure of the $MFCC+\Delta+\Delta\Delta$ feature is similar to $RMFCC+\Delta+\Delta\Delta$ feature described in Section 5.3, except the use of the speech signal. In this work the $MFCC$ concatenated with its Δ and $\Delta\Delta$ is called as the $MFCC+\Delta+\Delta\Delta$ feature. In building the speaker models and later for testing, all the experimental conditions as used for excitation based system remain same for fair comparison.

The speaker verification results of $MFCC+\Delta+\Delta\Delta$ feature for *Clean* and *Noisy* test cases are given in the first row of the Table 7.2. The EER achieved by $MFCC+\Delta+\Delta\Delta$ feature for *Clean* and *Noisy* test cases are 6.92% and 22.58%, respectively. The corresponding DET curves are shown in Figures 7.6 and 7.7, respectively. In these figures the respective DET curve of Src feature are also shown for comparison. It can be observed from the DET curves that the both

miss probability and false alarm probability of $MFCC + \Delta + \Delta\Delta$ feature are relatively better than *Src*. This indicates that the speaker-specific evidence present in the $MFCC + \Delta + \Delta\Delta$ feature have more inter-speaker variability and less intra-speaker variability, hence provides the good EER. Since $MFCC + \Delta + \Delta\Delta$ feature are derived from the cepstral processing of the speech signal, their performance also degrades under *Noisy* test condition. The relative degradation in the performance from *Clean* to *Noisy* test case is 226%. It can be observed from the DET curves that, the miss probability of the $MFCC + \Delta + \Delta\Delta$ feature is relatively more affected by noise. For example, due to noise the miss probability degrades from 25% to 85% and the false alarm probability degrades from 40% to 85%. The relative degradation in the miss and false alarm probabilities are 240% and 112%, respectively. This shows that, when the speech signal is corrupted by noise, the intra-speaker variability nature of the vocal tract information represented by $MFCC + \Delta + \Delta\Delta$ feature is relatively more affected.

Table 7.2: Speaker verification performances (%) of vocal tract and its combination with excitation features for *Clean* and *Noisy* test cases using GMM-UBM modeling technique. *Src* represents combination of subsegmental, segmental and suprasegmental excitation information extracted using respective proposed explicit approaches. $Comb_1$ represents linear score level combination schemes.

Feature		*Performance(EER)*		*Relative Degradation(%)*
		Clean	*Noisy*	
$MFCC + \Delta + \Delta\Delta$		6.92	22.58	226
Src	$Comb_1$	14.13	30.85	118
$Src + MFCC + \Delta + \Delta\Delta$	$Comb_1$	5.89	17.44	196

7.3.2 Comparison of Speaker Verification Performance of *Src* and $MFCC + \Delta + \Delta\Delta$ Features

By comparing the performance of the proposed *Src* and $MFCC + \Delta + \Delta\Delta$ features from the Table 7.2, it can be observed that the former provides relatively poor performance for both *Clean* and *Noisy* test cases. By comparing their respective DET curves shown in Figures 7.6 and 7.7, it can be observed that in most of the cases the false alarm probability and miss probability of the *Src* feature is relatively poor than the $MFCC + \Delta + \Delta\Delta$. In particular, the

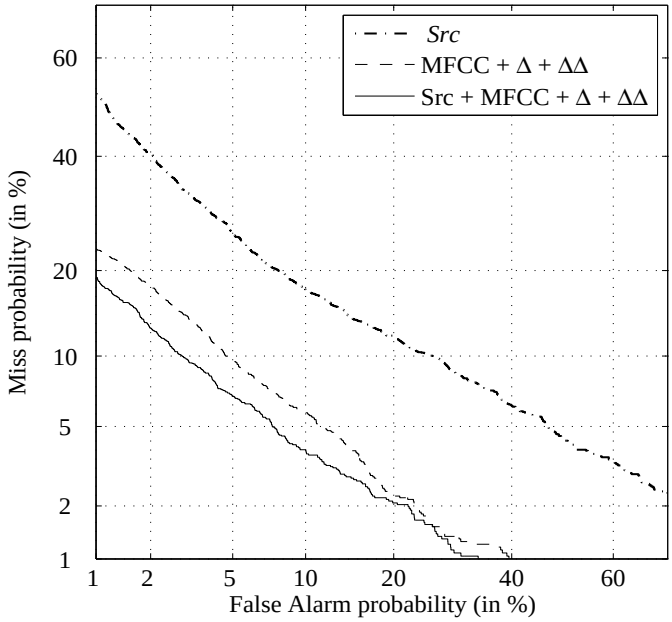

Figure 7.6: DET curves from the speaker verification experiments using evidence from the vocal tract and its combination with excitation using $Comb_1$ scheme for *Clean* test case.

false alarm probability is significantly large. The poor performance in case of Src feature may be due to the large false alarm probability. This indicates that the speaker-specific evidence present in Src may have less inter-speaker variability as compared to the $MFCC + \Delta + \Delta\Delta$ feature.

For both *Clean* and *Noisy* test cases, the verification performance is improved by the combined use of Src and $MFCC + \Delta + \Delta\Delta$ features. In this work the combined evidence from Src and $MFCC + \Delta + \Delta\Delta$ is represented by $Src + MFCC + \Delta + \Delta\Delta$. The DET curves of $Src + MFCC + \Delta + \Delta\Delta$ from $comb_1$ scheme for *Clean* and *Noisy* test cases are shown in Figures 7.6 and 7.7, respectively. The corresponding EER values are given in the last row of the Table 7.2. The verification performance achieved by $Src + MFCC + \Delta + \Delta\Delta$ for *Clean* and *Noisy* test cases are 5.89% and 17.44%, respectively. It can be observed from the respective DET curves that, in most of the cases the Src helps $MFCC + \Delta + \Delta\Delta$ in reducing both the miss and false alarm probabilities. This shows that the speaker-specific evidence present in Src is different from $MFCC + \Delta + \Delta\Delta$ and they together further improve the recognition

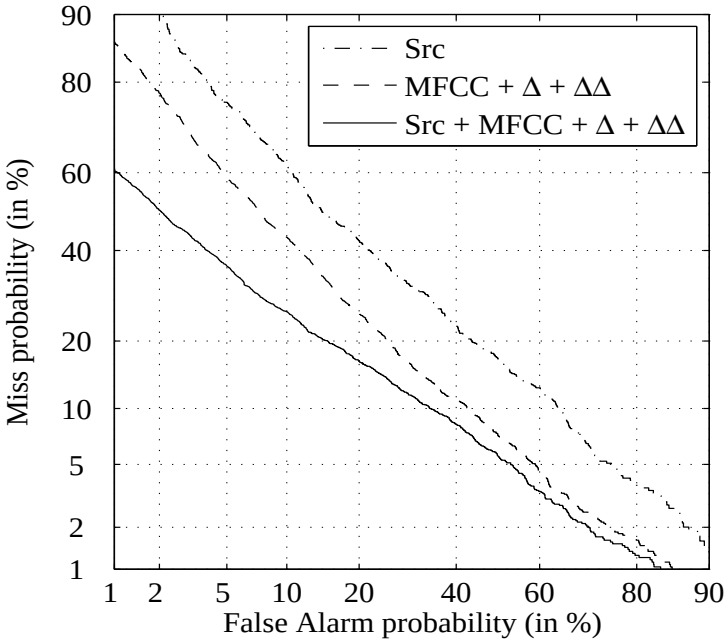

Figure 7.7: DET curves from the speaker verification experiments using evidence from the vocal tract and its combination with excitation using $Comb_1$ scheme for *Noisy* test case.

performance.

It is interesting to observe that, to improve the verification performance, the contribution of the Src to $MFCC + \Delta + \Delta\Delta$ is relatively better in case of *Noisy* test case. For example, the relative improvement in case of *Clean* test case is 15%, as compared to 23% in case of *Noisy* test case. In case of *Noisy* test case, the Src helps $MFCC + \Delta + \Delta\Delta$ feature in reducing the miss probability significantly. This can be observed from the DET curves of $Src + MFCC + \Delta + \Delta\Delta$ for *Noisy* test case shown in Figure 7.7. The maximum miss probability of $MFCC + \Delta + \Delta\Delta$ feature is around 85% and of $Src + MFCC + \Delta + \Delta\Delta$ is around 60%. A relative improvement of 29% is achieved. These observations indicate that, the recognition performance of the state-of-the-art vocal tract feature degrades when the speech signal is corrupted by noise. This can be compensated by combining the evidences from the excitation information.

7.3.3 Robustness of Vocal Excitation Source Features

Experimental results from *Clean* and *Noisy* test cases demonstrate that, the performance of both *Src* and $MFCC + \Delta + \Delta\Delta$ features degrade when the speech signal is corrupted by noise. By comparing the relative degradation in the performance of excitation and vocal tract features, it can be observed that the maximum degradation is occurred in case of the $MFCC + \Delta + \Delta\Delta$ feature, as compared to all the excitation features. This shows that the speaker-specific evidence provided by excitation features are relatively more robust against noise than the conventional vocal tract features.

Although, the relative degradation measurement shows the robustness of a feature but for comparison purpose, it may not seem to be a better assessment approach. Because, features that provide very poor performance in the clean speech case may have little speaker information. The effect of noise in this little speaker information may not be significant. Thus, the performance degradation of the poor performing features due to noise may be less. For example, in case of *Clean* test case, the least performance achieved by A_0 vectors is degraded by 4% due to noise may not be considered as a robust feature. Therefore, to further verify the robustness of the excitation features, we consider error rate reduction (ERR) measure [98]. In this measure, the relative degradation in the performance due to noise is measured with a common good performing system, called as the baseline system. In the present case, the vocal tract system using the $MFCC + \Delta + \Delta\Delta$ feature is considered as the baseline system. The ERR measures the relative degradation of a feature with the baseline system. Let, X_1 and Y_1 are the EER achieved by the baseline and the proposed system, respectively. Then, the ERR measure is given by the Equation 7.2, [98].

$$ERR(\%) = \frac{Y_1 - X_1}{X_1} \times 100 \tag{7.2}$$

By comparing the ERR for *Clean* and *Noisy* test cases, the robustness of the excitation features can be demonstrated. For example, if the excitation features have relatively lower ERR value for *Noisy* test case, it indicates they are more robust against noise. A good feature should have minimum ERR. The relative decrement measure between the ERR values for *Clean*

and *Noisy* test cases of a feature will indicate its robustness against noise. A good feature should have maximum relative decrement in the ERR from *Clean* to *Noisy* test cases. Moreover, the ERR is a relative measure with a common good performing feature ($MFCC + \Delta + \Delta\Delta$). So a fair comparison among the different excitation features can also be made to demonstrate their individual robustness against noise. In this work, for computation of ERR X_1 is considered as 6.92 and 22.58 for *Clean* and *Noisy* test cases, respectively.

The ERR values of all the excitation features for *Clean* and *Noisy* test cases are given in second and third columns of the Table 7.3, respectively. In all cases the ERR in case of *Noisy* test case is relatively less than the corresponding *Clean* test. This shows that the speaker-specific evidence provided by the proposed excitation features are relatively less affected by noise. By comparing the ERR from the *Noisy* test case, it can be observed that, individually the $GFD + \Delta + \Delta\Delta$ feature provides the least ERR of 39% followed by $MPDSS$ feature of around 59%. The possible reason may be that these features mostly associated with the harmonics of the excitation may be less affected by the noise. The pitch, epoch strength and cepstral trajectory vectors show large ERR for the *Noisy* test case. This indicates that, the suprasegmental contour information are largely affected by noise. Also, the higher ERR in case of T_0 and A_0 vectors may be due to the method employed for their accurate estimation. For example, the HE of the LP residual from the signal corrupted by factory noise may need further processing to obtain the accurate zero-frequency filtered signal. By comparing the relative robustness measure given in the last column of the Table 7.3, it can be observed that the maximum relative decrement in the ERR is in case of $GFD + \Delta + \Delta\Delta$ and T_0 vectors, around 80%, indicating more robustness against noise.

By comparing the ERR of the excitation information from three different levels, we observed that the segmental excitation information have the least ERR of 39% followed by segmental of 52%. The corresponding maximum relative ERR decrements are 81% and 62%, respectively. These observations indicate that, the subsegmental and segmental excitation information represented by their proposed respective features may provide robust speaker-specific evidence for speaker recognition. By combining the evidence from all the three levels the ERR is further

reduced. The ERR measures of the Src is given in the last row of the Table 7.3. The ERR for *Clean* and *Noisy* test cases are 104% and 50%, respectively. Decrement in the ERR is 52%. These results show that the combined representation of subsegmental, segmental and suprasegmental excitation information provides relatively robust representation of the excitation information.

Table 7.3: Robustness of excitation features against noise. Factory noise (SNR 9dB) is added only to test speech signals. $MFCC + \Delta + \Delta\Delta$ is considered as the baseline system for ERR computation. $Src_6 = RMFCC + \Delta + \Delta\Delta + MPDSS$. $C_t = c_{t1} + c_{t2} + c_{t4} + c_{t6}$. $Src_5 = T_0 + A_0 + C_t$. Src represents combination of subsegmental, segmental and suprasegmental excitation information extracted using respective proposed explicit approaches. $Comb_1$ represents linear score level combination schemes.

$Features$		ERR (%)		$Robustness$(%)
		Clean	*Noisy*	
$GFD + \Delta + \Delta\Delta$		193	39	81
$MPDSS$		169	59	65
$RMFCC + \Delta + \Delta\Delta$		165	74	55
T_0		387	80	80
A_0		539	100	80
c_{t1}		354	109	69
c_{t2}		351	109	69
c_{t4}		363	110	70
c_{t6}		374	114	69
Src_6	$Comb_1$	137	52	62
C_t	$Comb_1$	286	108	62
Src_5	$Comb_1$	263	80	70
Src	$Comb_1$	104	50	52

The experimental results described in this chapter suggest that the proposed Src feature contains good amount of speaker-specific excitation information that can be used for the development of the state-of-the-art speaker recognition system. The performance of the Src is relatively poor but helps the state-of-the-art $MFCC + \Delta + \Delta\Delta$ vocal tract feature to further improve the performance for both *Clean* and *Noisy* test cases. The performance of the $MFCC + \Delta + \Delta\Delta$ feature is severely suffered for *Noisy* test case, indicating less robustness against noise. On the other hand, the speaker-specific evidence provided by Src is relatively more robust against noise.

7.4 Summary

In this chapter, we developed a speaker verification system using excitation information and made a comparative study with the corresponding vocal tract system. In the proposed system, the excitation features are extracted as suggested in our studies made in the previous chapters. Post feature processing like CMS and incorporation of the dynamic information were included. GMM-UBM modeling technique is used for building the speaker models. The significance of the proposed SR system is demonstrated by speaker verification experiments on clean and noise conditions. The speaker verification performance of the vocal tract features ($MFCC+\Delta+\Delta\Delta$) is also evaluated on similar experimental condition for comparative study. We observe that the $MFCC+\Delta+\Delta\Delta$ have less intra-speaker and more inter-speaker variability and hence provide relatively better performance. On the other hand, the excitation feature (Src) has relatively more inter-speaker variability that increases the false alarm probability. As a result the Src provides relatively poor performance. The relative degradation in the performance due to noise and ERR measure demonstrates that, the Src provides relatively more robust speaker-specific evidence. Although, the $MFCC+\Delta+\Delta\Delta$ feature perform well in favorable environment, but suffers from noise effect. In this sense, the proposed excitation features based SR system is useful for the speaker recognition task.

8

Summary and Conclusions

Contents

8.1 Summary of the Work

The present work first studied the amount of speaker-specific excitation information present in the subsegmental, segmental and suprasegmental levels of the LP residual by processing the LP residual directly in the time domain. Since there are no explicit speaker-specific parameters extracted, the approach is termed as implicit modelling. After this methods are developed for explicit modelling of the subsegmental, segmental and suprasegmental levels excitation information. The significance of the proposed methods is demonstrated by performing the speaker identification and verification experiments. Finally, a speaker verification system based on excitation information is developed and a comparative study is made with the corresponding vocal tract based system.

The summary of the observations from the implicit modelling of the LP residual are as follows: At the subsegmental level, 5 msec with a shift of 2.5 msec blocks of LP residual are used as the feature vectors to represent the speaker-specific information. The 20 msec blocks with a shift of 2.5 msec blocks of LP residual decimated by a factor four are used to represent the segmental level excitation information. The 250 msec blocks with a shift of 6.25 msec blocks of LP residual decimated by a factor of fifty are used to represent the suprasegmental level excitation information. The evidences from all these levels are modeled independently to evaluate the amount of speaker information present in them. The experimental results show that good amount of speaker-specific excitation information is present at each of these levels of the LP residual and is useful for speaker recognition task. The segmental level blocks provide the best performance followed by the subsegmental blocks. The suprasegmental level blocks provides the least performance. The poor performance may be due to large intra-speaker variability and also due to text-independent mode of operation. The evidence from these levels are observed to be different from each other and combined well to provide useful speaker-specific excitation information to further improve the recognition performance. A comparative study with the conventional vocal tract feature demonstrates that, if suitable combination technique is available, it is indeed possible to achieve good recognition performance from the excitation

information perspective alone.

The speaker-specific excitation information in the LP residual may be due to the variation in both the amplitude and sequence. To avoid the dominance of one on the other, the amplitude and sequence information are modeled independently by Hilbert envelope (HE) and cosine phase of LP residual, respectively. The HE and cosine phase of the LP residual are processed in the subsegmental, segmental and suprasegmental levels to capture the respective amplitude and sequence information. The detailed results show that, the evidences in HE and residual phase (RP) are different and combined well at each level to represent the complete excitation information. It is also observed that, it may be better to first combine the HE and RP at the subsegmental, segmental and suprasegmental levels separately and then combine them for effective representation of the complete excitation information. The combined evidence from amplitude and sequence information at each level provides relatively better performance than the corresponding LP residual vectors.

The objective of the implicit modelling was to demonstrate the amount of speaker-specific excitation information present in the LP residual. It is observed that the implicit modelling of the LP residual is useful for speaker recognition tasks but computationally intensive. For effective and compact representation of the speaker-specific excitation information, we process the LP residual in an explicit modelling approach.

The LF model of the GFD cycle is used to model the subsegmental excitation information. The LF parameters are computed from the LP residual. A simple and computationally efficient approach is proposed for the approximate estimation of the LF parameters from the LP residual blocks. These blocks are identified by locating GCIs and GOIs. The zero-frequency filtering approach is used to locate the GCIs. To avoid the difficulty in using the zero-frequency filtering approach for telephone speech, we proposed to use the HE of the LP residual as the input to the zero-frequency filter. The significance of the proposed approach of estimating the GCIs is demonstrated by computing and comparing the GCIs for a speech signal collected through microphone and telephone. GOIs are obtained as the fixed fraction of the close-phase intervals. Initially, the LF parameters are computed with an assumption that instants of the first zero

crossing and the slope of the return phase are fixed fraction of the GFD cycle interval. Then, the estimated LF parameters are optimized by using the constraint that the flow returns to zero at the end of each glottal cycle. The proposed approach significantly reduces the computation needed to implement the LF model. The LF parameters together with their Δ and $\Delta\Delta$ are used as features to model the subsegmental level excitation information. The Δ and $\Delta\Delta$ values of LF parameters are included to capture the aspiration and ripple information. Experimental results show that the $GFD+\Delta+\Delta\Delta$ feature well model the subsegmental level information for speaker recognition. A comparative study between $GFD+\Delta+\Delta\Delta$ feature and subsegmental LP residual vectors revealed that the later provides significantly better performance in speaker identification task and the former for the speaker verification task. $GFD+\Delta+\Delta\Delta$ features may have relatively less intra and inter-speaker variability and hence providing good performance for speaker verification task. For both speaker identification and verification tasks the $GFD+\Delta+\Delta\Delta$ feature well combined with other level excitation information and relatively more robust against noise. Thus, we suggest $GFD+\Delta+\Delta\Delta$ as a possible way of parameterizing the speaker-specific subsegmental excitation information from the LP residual.

The excitation periodicity information is captured by the spectral flatness measure of the LP residual. The excitation energy information is captured by the cepstral coefficients of the LP residual. We found that, MPDSS and RMFCC computed from mel warped subband spectra well capture the excitation periodicity and energy information, respectively. The different aspect of speaker-specific information present in MPDSS and RMFCC is different and well combined to further improve the recognition performance. The MPDSS and RMFCC features are computed from the mel warped LP residual spectra computed from 20 msec with a shift of 10 msec blocks of the LP residual and correspond to segmental level excitation information in the time domain. A comparative study made on combined frequency and cepstral domains processing and corresponding temporal domain processing of the LP residual revealed that with a little compromise in the recognition performance, the combined use of MPDSS and RMFCC features is a possible way of representing the segmental level excitation information. The evidence from the combined MPDSS and RMFCC representation provides more different information to other

levels excitation information.

In the suprasegmental level, the speaker-specific excitation information is manifested in pitch and epoch strength contours and LP residual cepstral trajectories spanning across several pitch periods. To model the suprasegmental pitch and epoch strength contours information, the pitch and epoch strength vectors are used. To estimate the pitch and epoch strength contours, the modified zero-frequency filtering approach is used. Pitch is computed as the interval between successive positive zero-crossings in the zero-frequency filtered signal (ZFFS). The slope of the ZFFS around the zero crossings is measured as the epoch strength. Every ten pitch and epoch strength values with a shift of one value are used as the pitch and epoch strength vectors (T_0 and A_0 vectors). The dimension of the pitch and epoch strength vectors are decided based on the identification results conducted on a small database. Every sample shift is considered to collect maximum number of feature vectors for better modelling. The excitation information present in the pitch and epoch strength vectors are modeled independently. The experimental results show that the combined evidence from pitch and epoch strength vectors well capture the suprasegmental pitch and epoch strength information for speaker recognition.

To model the suprasegmental LP residual cepstral trajectory information, we proposed to use first few $RMFCC$ cepstral trajectories except the first coefficient. For this, four cepstral trajectories such as, c_{t1}, c_{t2}, c_{t4} and c_{t6} having the highest F-ratio value are selected. Blocks of ten with a shift of one value from each cepstral trajectory are used as the feature vectors to represent the speaker-specific information present in them. The dimension of the feature vectors are considered based on the study for pitch and epoch strength vectors. The feature vectors from the individual cepstral trajectory are modeled independently. Experimental results show that the individual cepstral trajectory feature vectors contain speaker-specific excitation information. The speaker-specific evidence present in individual cepstral trajectory vectors are observed to be different and combined well to further improve the recognition performance. The detailed recognition results demonstrate that the speaker-specific evidence present in the combined representation of the cepstral trajectory vectors and, pitch and epoch strength vectors are different. They together well represent the suprasegmental level excitation information. By

a comparative study, we found that the proposed suprasegmental level representative feature provides better performance than the corresponding LP residual vectors and well combine with other level excitation information. Thus, we suggest that $T_0 + A_0 + C_t$ representation as a possible way of parameterizing the speaker-specific suprasegmental excitation information from the LP residual.

The objective of the above studies was to verify the amount of the speaker-specific excitation information present at the subsegmental, segmental and suprasegmental levels of the LP residual and develop different methods for their effective modelling. We found that significant speaker-specific excitation information is present at each level of the LP residual. Due to computationally less intensive and more robustness, the proposed approaches of extracting the speaker-specific excitation information for each level by parameterizing the LP residual are found to be the possible way for speaker recognition. Since, the objective was to develop effective methods for modelling speaker-specific excitation information at each level, the initial study is made on baseline system. We use the proposed feature from each level of the LP residual and develop a speaker verification system based on the excitation information. The *RMFCC* features are subjected to CMS and their temporal dynamic information is incorporated by concatenating the Δ and $\Delta\Delta$ values. GMM-UBM modelling technique is used to build the speaker models. The speaker-specific evidence from each level is combined at the measurement level to give the decisions.

The significance of the proposed system is demonstrated by conducting the speaker verification experiments on both clean and noise test cases and a comparative study with the corresponding vocal tract based system. Results of speaker verification experiments for both clean and noisy cases show that the proposed excitation information based system is effective for speaker recognition task. Independently, it gives relatively less recognition accuracy but provides complementary and robust speaker-specific evidences to the conventional vocal tract information based system. The conventional vocal tract based system provides good performance in the favorable environment, but suffers severely from the noise effect. In this sense, the proposed excitation features based SR system is relatively more robust and hence may be

useful for the speaker recognition task under noisy conditions.

8.2 Contributions of the Work

The contributions of the work reported in this thesis for processing the LP residual for speaker-specific excitation information include,

- Implicit processing of the LP residual in the time domain with different frame size and shift for the extraction of subsegmental, segmental and suprasegmental speaker-specific excitation information.
- Implicit processing of the analytic representation of the LP residual in the time domain for independent modelling of the amplitude and sequence information.
- Modification suggested to the zero-frequency filtering method for telephone speech for the accurate estimation of glottal closure instants (GCIs) and then pitch and epoch strength.
- Proposed efficient approach for computation of the LF parameters from the LP residual blocks.
- Explicit modelling of the subsegmental level excitation information by using LF parameters.
- Investigation on filter shapes for processing the LP residual in frequency and cepstral domains for explicit modelling of the segmental level excitation information that results in $RMFCC$ and $MPDSS$ features.
- Explicit modelling of the suprasegmental level excitation information by the combined use of pitch, epoch strength and cepstral trajectory vectors.
- Development of the speaker verification system using excitation information.

8.3 Scope for the Future Work

- We made an attempt to process the LP residual in subsegmental, segmental and suprasegmental levels by considering the fixed frame size and frame shift. The corresponding LP residual blocks are not pitch synchronized. Significant speaker-specific excitation information is known to be present around the epochs [13]. Thus, the LP residual samples around the epochs at each levels may be useful for modelling the respective excitation information. By selecting the pitch synchronous blocks we may able to reduce the computational complexity involved in the time domain processing of the LP residual directly. With pitch synchronized LP residual blocks, the number of feature vectors at the respective levels may be reduced. The difficulty in building the speaker models with the less number of feature vectors also needs to be explored.

- In processing the LP residual in cepstral domain we use standard mel warped spectrum that provides high resolution to the lower bands (below 1 kHz) and low to the higher bands. In general, the distribution of the excitation energy in the higher band is relatively more than the lower bands [21, 99]. Since the cepstral coefficients essentially capture the excitation energy information, a filter with higher resolution in the higher bands may be useful. For this, inverse mel bank filters may be used [100].

- In this work we use GMM and GMM-UBM modelling techniques. The usefulness of the AANN [27, 66], support vector machines (SVM) [101, 102] and joint factor analysis [103] modelling techniques need to be verified with the parameterization of the subsegmental, segmental and suprasegmental levels excitation information. Although, they are computationally intensive but may provide better recognition accuracy from the excitation information perspective.

- The combination scheme we used to get the maximum benefit is based on the the assumption that the ground truth information is available ($Comb_2$) and hence may not be useful for real time application. Suitable combination techniques need to be developed

for combing the evidences from different features for improving the speaker recognition performance. For this, the recently proposed feature switching technique that measures the quantity and quality of the information present in the feature may be explored [104].

- In this work, the robustness of the proposed methods is verified on the telephone speech data and adding factory noise (9 dB) to test data only. The robustness of the proposed methods needs to be verified by adding different noises like, car noise, pink noise and bubble noise to both train and test data at different SNR levels.

- In our work, we used LP residual as the representation of the excitation signal. The LP residual is an approximation of the excitation signal representation. Future work should focus on using the glottal waves directly recorded from speakers and/or derived from the speech signal as the excitation signal.

- MFCC features are computed directly from the speech signal. So, they contain both vocal tract as well as the excitation information. But, the dominant speaker information present in MFCC features is the vocal tract information. The effectiveness of the excitation information in MFCC features is suppressed. It is expected that, if we independently estimate the vocal tract and excitation information from the speech signal and then combine their respective evidence, one may achieve improved recognition performance. For this, the LP cepstral coefficients may be used as the representation of only vocal tract information and can be combined with the proposed excitation information based system.

A

Linear Prediction Coefficients Computation

Contents

A.1 Linear Prediction Coefficients (LPC)

In linear prediction (LP) analysis of speech, each sample is predicted as a linear weighted combination of the past p samples, where p represents the order of prediction [1,23,24]. If $s(n)$ is the present sample, then it is predicted by the past p samples as

$$\hat{s}(n) = -\sum_{k=1}^{p} a_k s(n-k) \tag{A.1}$$

where, $a_k s$ are the LP coefficients (LPCs) computed by minimizing the mean square prediction error, such that

$$s(n) = a_1 s(n-1) + a_2 s(n-2) + ... + a_p s(n-p) \tag{A.2}$$

where the coefficients a_1, a_2,...,a_p are assumed constant over speech analysis frame.

A.2 Estimation of Linear Prediction Coefficients

There are two methods for estimating LPC:

- Autocorrelation
- Covariance

Both methods choose the short-term filter coefficients a_k in such a way that the energy in the error signal (residual) is minimized. For speech processing tasks, the autocorrelation method is exclusively used because of its computational efficiency and inherent stability, whereas the covariance method does not guarantee the stability of the all-pole LP synthesis filter [1, 43]. The autocorrelation method of computing LPC is described below:

First, speech signal $s(n)$ is multiplied by a window $w(n)$ to get the windowed speech segment $s_w(n)$. Normally, a Hamming or Hanning window is used. The windowed speech signal is expressed as

$$s_w(n) = s(n)w(n). \tag{A.3}$$

The next step is to minimize the energy in the residual signal. The residual energy E_p is defined as [24]

$$E_p = \sum_{n=-\infty}^{\infty} e^2(n) = \sum_{n=-\infty}^{\infty} \left(s_w(n) + \sum_{k=1}^{p} a_k s_w(n-k) \right)^2. \tag{A.4}$$

The values of a_k that minimize E_p are found by by setting the partial derivatives of the energy E_p with respect to the LPC parameters equal to zero.

$$\frac{\partial E_p}{\partial a_k} = 0, \quad 1 \leq k \leq p. \tag{A.5}$$

This results in the following p linear equations for the p unknown parameters $a_1, \ldots, a_p$

$$\sum_{k=1}^{p} a_k \sum_{n=-\infty}^{\infty} s_w(n-i) s_w(n-k) = -\sum_{n=-\infty}^{\infty} s_w(n-i) s_w(n), \quad 1 \leq i \leq p. \tag{A.6}$$

This linear equations can be expressed in terms of the autocorrelation function. This is because the autocorrelation function of the windowed segment $s_w(n)$ is defined as

$$R_s(i) = \sum_{n=-\infty}^{\infty} s_w(n) s_w(n+i), \quad 1 \leq i \leq p. \tag{A.7}$$

Exploiting the fact that the autocorrelation function is an even function i.e., $R_s(i) = R_s(-i)$. By substituting the values from Equation (A.7) in Equation (A.6), we get

$$\sum_{k=1}^{p} R_s\left(|i-k|\right) a_k = -R_s(i), \quad 1 \leq i \leq p. \tag{A.8}$$

These set of p linear equations can be represented in the following matrix form as [1, 43]

$$\begin{bmatrix} R_s(0) & R_s(1) & \cdots & R_s(p-1) \\ R_s(1) & R_s(0) & \cdots & R_s(p-2) \\ \vdots & \vdots & \ddots & \vdots \\ R_s(p-1) & R_s(p-2) & \cdots & R_s(0) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix} = - \begin{bmatrix} R_s(1) \\ R_s(2) \\ \vdots \\ R_s(p) \end{bmatrix} \tag{A.9}$$

This can be summarized using vector-matrix notation as

$$\mathbf{R_s a} = -\mathbf{r_s} \tag{A.10}$$

where the $p \times p$ matrix $\mathbf{R_s}$ is known as the autocorrelation matrix. The resulting matrix is a Toeplitz matrix where all elements along a given diagonal are equal. This allows the linear equations to be solved by the Levinson-Durbin algorithm. Because of the Toeplitz structure of $\mathbf{R_s}$, $A(z)$ is minimum phase [1, 89, 105]. At the synthesis filter $H(z) = 1/A(z)$, the zeros of $A(z)$ become the poles of $H(z)$. Thus, the minimum phase of $A(z)$ guarantees the stability of $H(z)$.

B

MFCC Feature Extraction

Contents

B.1 MFCC Feature Extraction

The various steps involved in the MFCC feature extraction is shown in Figure B.1. The brief description of the steps are as follows [21, 45, 99]:

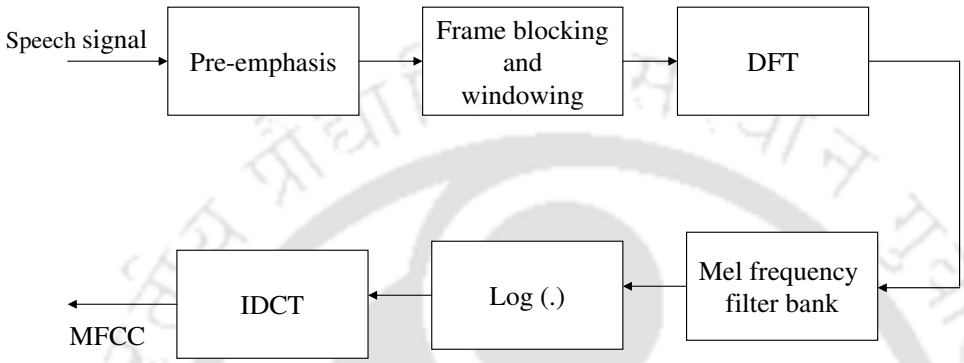

Figure B.1: MFCC feature extraction process

(i) Pre-emphasis: This refers to filtering that emphasizes the higher frequencies. Its purpose is to balance the spectrum of voiced sounds that have a steep roll-off in the high frequency region. The most commonly used pre-emphasis filter is given by the following transfer function

$$H(z) = 1 - az^{-1} \tag{B.1}$$

where the value of γ controls the slope of the filter and is usually between 0.9 to 1.0.

(ii) Frame blocking and windowing: The speech is slow varying quasi-stationary signal. Therefore, speech analysis must always be carried out on short segments across which the speech signal is assumed to be stationary. Short-term spectral measurements are typically carried out over the range of 10-30 ms frame size and shift [1, 21]. The blocked frames are Hamming windowed. This helps to reduce the edge effect while taking the DFT on the signal.

(iii) DFT spectrum: Each windowed frame is converted into magnitude spectrum by applying

DFT.

$$X(k) = \sum_{n=0}^{N-1} x(n) e^{\frac{-j2\pi nk}{N}}; \qquad 0 \leq k \leq N-1 \tag{B.2}$$

where N is the number of points used to compute the DFT.

(iv) Mel-spectrum: This can be computed by passing the Fourier transformed signal through a set of band-pass filters known as mel-filter bank. A mel is a unit of perceived speech frequency or a unit of tone. The mel scale is therefore a mapping between the physical frequency scale (Hz) and the perceived frequency scale (Mels). The approximation of mel from physical frequency can be expressed as [21, 99]:

$$f_{mel} = 2595 \log_{10}\left(1 + \frac{f}{700}\right) \tag{B.3}$$

where f denotes the physical frequency and f_{mel} denotes the perceived frequency.

The mel spectrum values or mel frequency coefficients of the magnitude spectrum $X(k)$ is computed by multiplying the magnitude spectrum by each of the triangular mel weighting filters.

$$S(m) = \sum_{k=0}^{N-1} |X(k)|^2 H_m(k); \qquad 0 \leq m \leq M-1 \tag{B.4}$$

where M is total number of triangular mel weighting filters.

(v) Inverse Discrete Cosine Transform (IDCT): The log operation is performed on the mel frequency coefficients. The IDCT is then applied to obtain cepstral coefficients. This results in a signal in the cepstral domain. MFCC is computed as :

$$c(n) = \sum_{m=0}^{M-1} \log_{10}\left(S(m)\right) \cos\left(\frac{\pi n(m-0.5)}{M}\right) \qquad n = 0, 1, 2, ..., C-1 \tag{B.5}$$

where $c(n)$ are the cepstral coefficients and C is the number of MFCCs. The zeroth coefficient is often excluded since it represents the average log-energy of the input signal, which only carries little speaker-specific information. Silence and low-energy speech parts are removed using an energy-based voice activity detection (VAD) technique [106].

C

Gaussian Mixture Models

Contents

C.1 Gaussian Mixture Model (GMM) Description

In the speech and speaker recognition the acoustic events are usually modeled by Gaussian probability density functions (PDFs), described by the mean vector and the covariance matrix. However unimodel PDF with only one mean and covariance are unsuitable to model all variations of a single event in speech signals. Therefore, a mixture of single densities is used to model the complex structure of the density probability. For a D-dimensional feature vector denoted as x_t, the mixture density for speaker s is defined as weighted sum of M_g component Gaussian densities as given by the following equation [5, 67]

$$P(x_t|s) = \sum_{i=1}^{M_g} w_i P_i(x_t) \tag{C.1}$$

where w_i are the weights and $P_i(x_t)$ are the component densities. Each component density is a D-variate Gaussian function of the form

$$P_i(x_t) = \frac{1}{(2\pi)^{D/2}\,|\Sigma_i|^{\frac{1}{2}}} e^{-\frac{1}{2}\left[(x_t-\mu_i)'\Sigma_i^{-1}(x_t-\mu_i)\right]} \tag{C.2}$$

where μ_i is a mean vector and Σ_i covariance matrix for i^{th} component. The mixture weights have to satisfy the constraint [5, 67]

$$\sum_{i=1}^{M_g} w_i = 1. \tag{C.3}$$

The complete Gaussian mixture density is parameterized by the mean vector, the covariance matrix and the mixture weight from all component densities. These parameters are collectively represented by

$$s = \{w_i, \mu_i, \Sigma_i\}\,; \qquad i = 1, 2, ...M_g. \tag{C.4}$$

C.2 Training the GMMs

To determine the model parameters of GMM of the speaker, the GMM has to be trained. In the training process, the maximum likelihood (ML) procedure is adopted to estimate model parameters. For a sequence of training vectors $X = \{x_1, x_2, .., x_T\}$, the GMM likelihood can be

written as (assuming observations independence) [5, 67]

$$P(X|s) = \prod_{t=1}^{T} P(x_t|s). \tag{C.5}$$

Usually this is done by taking the logarithm and is commonly named as log-likelihood function. From Equations (C.1) and (C.5), the log-likelihood function can be written as

$$\log\left[P(X|s)\right] = \sum_{t=1}^{T} \log\left[\sum_{i=1}^{M} w_i P_i(x_t)\right]. \tag{C.6}$$

Often, the log-likelihood value is divided by T to normalize out duration effects. Also, since the incorrect assumption of independence is underestimating the actual likelihood value with dependencies, scaling by T can be considered a rough compensation factor [5, 67]. The parameters of a GMM model can be estimated using maximum likelihood (ML) estimation. The main objective of the ML estimation is to derive the optimum model parameters that can maximize the likelihood of GMM. The likelihood value is, however, a highly nonlinear function in the model parameters and direct maximization is not possible. Instead, maximization is done through iterative procedures. Of the many techniques developed to maximize the likelihood value, the most popular is the iterative expectation maximization (EM) algorithm [68, 107].

C.2.1 Expectation Maximization (EM) Algorithm

The EM algorithm begins with an initial model s and tends to estimate a new model such that the likelihood of the model increasing with each iteration. This new model is considered to be an initial model in the next iteration and the entire process is repeated until a certain convergence threshold is obtained or a certain predetermined number of iterations have been made. A summary of the various steps followed in the EM algorithm are described below.

(i) **Initialization:** In this step an initial estimate of the parameters is obtained. The performance of the EM algorithm depends on this initialization. Generally, LBG [69] or K-means algorithm [69] is used to initialize the GMM parameters.

(ii) **Likelihood Computation:** In each iteration the posterior probabilities for the i^{th} mix-

ture is computed as [5, 67]:

$$\Pr(i|x_t) = \frac{w_i P_i(x_t)}{\sum_{j=1}^{M} w_j P_j(x_t)}. \tag{C.7}$$

(iii) **Parameter Update:** Having the posterior probabilities, the model parameters are updated according to the following expressions [5, 67].

Mixture weight update:

$$\overline{w_i} = \frac{\sum_{i=1}^{T} \Pr(i|x_t)}{T}. \tag{C.8}$$

Mean vector update:

$$\overline{\mu_i} = \frac{\sum_{i=1}^{T} \Pr(i|x_t) x_t}{\sum_{i=1}^{T} \Pr(i|x_t)}. \tag{C.9}$$

Covariance matrix update:

$$\overline{\sigma}_i^2 = \frac{\sum_{i=1}^{T} \Pr(i|x_t) \left| x_t - \overline{\mu_i} \right|^2}{\sum_{i=1}^{T} \Pr(i|x_t)}. \tag{C.10}$$

In the estimation of the model parameters, it is possible to choose, either full covariance matrices or diagonal covariance matrices. It is more common to use diagonal covariance matrices for GMM, since linear combination of diagonal covariance Gaussians has the same model capability with full matrices [108]. Another reason is that speech utterances are usually parameterized with cepstral features. Cepstral features are more compactable, discriminative, and most important, they are nearly uncorrelated, which allows diagonal covariance to be used by the GMMs [5, 67]. The iterative process is normally carried out 10 times, at which point the model is assumed to converge to a local maximum [5, 67].

C.2.2 Maximum *a posteriori* (MAP) Adaptation

Gaussian mixture models for a speaker can be trained using the modeling described earlier. For this, it is necessary that sufficient training data is available in order to create a model of

the speaker. Another way of estimating a statistical model, which is especially useful when the training data available is of short duration, is by using maximum *a posteriori* adaptation (MAP) of a background model trained on the speech data of several other speakers [109]. This background model is a large GMM that is trained with a large amount of data which encompasses the different kinds of speech that may be encountered by the system during training. These different kinds may include different channel conditions, composition of speakers, acoustic conditions, etc. A summary of MAP adaptation steps are given below.

For each mixture i from the background model, $Pr(i|x_t)$ is calculated as [5, 67]

$$\Pr(i|x_t) = \frac{w_i P_i(x_t)}{\sum_{j=1}^{M} w_j P_j(x_t)}. \tag{C.11}$$

Using $Pr(i|x_t)$, the statistics of the weight, mean and variance are calculated as follows [5, 67]

$$n_i = \sum_{i=1}^{T} \Pr(i|x_t) \tag{C.12}$$

$$E_i(x_t) = \frac{\sum_{i=1}^{T} \Pr(i|x_t) x_t}{n_i} \tag{C.13}$$

$$E_i(x_t^2) = \frac{\sum_{i=1}^{T} \Pr(i|x_t) x_t^2}{n_i}. \tag{C.14}$$

These new statistics calculated from the training data are then used adapt the background model, and the new weights ($\hat{w}_i$), means ($\hat{\mu}_i$) and variances ($\hat{\sigma}_i^2$) are given by [35]

$$\hat{w}_i = \left[\frac{\alpha_i n_i}{T} + (1 - \alpha_i) w_i\right] \gamma \tag{C.15}$$

$$\hat{\mu}_i = \alpha_i E_i(x_t) + (1 - \alpha_i)\mu_i \tag{C.16}$$

$$\hat{\sigma}_i^2 = \alpha_i E_i(x_t^2) + (1 - \alpha_i)(\sigma_i^2 + \mu_i^2) - \hat{\mu}_i^2. \tag{C.17}$$

A scale factor γ is used, which ensures that all the new mixture weights sum to 1. α_i is the adaptation coefficient which controls the balance between the old and new model parameter

estimates. α_i is defined as [35]

$$\alpha_i = \frac{n_i}{n_i + r} \tag{C.18}$$

where r is a fixed relevance factor, which determines the extent of mixing of the old and new estimates of the parameters. Low values for α_i ($\alpha_i \to 0$), will result in new parameter estimates from the data to be de-emphasized, while higher values ($\alpha_i \to 1$) will emphasize the use of the new training data-dependent parameters. Generally only mean values are adapted [35,36,65]. It is experimentally shown that mean adaptation gives slightly higher performance than adapting all three parameters [35].

C.3 Testing

In identification phase, mixture densities are calculated for every feature vector for all speakers and speaker with maximum likelihood is selected as identified speaker. For example, if S speaker models $\{s_{1,}s_{2,...,}s_S\}$ are available after the training, speaker identification can be done based on a new speech data set. First, the sequence of feature vectors $X = \{x_1, x_2, .., x_T\}$ is calculated. Then the speaker model $\hat{s}$ is determined which maximizes the a posteriori probability $P(s_S|X)$. That is, according to the Bayes rule [5, 67]

$$\hat{s} = \max_{1 \le s \le S} P(s_S|X) = \max_{1 \le s \le S} \frac{P(X|s_S)}{P(X)} P(s_S). \tag{C.19}$$

Assuming equal probability of all speakers and the statistical independence of the observations, the decision rule for the most probable speaker can be redefined as

$$\hat{s} = \max_{1 \le s \le S} \sum_{t=1}^{T} \log P(x_t|s_s) \tag{C.20}$$

Decision in verification is obtained by comparing the score computed using the model for the claimed speaker s_S given by $P(s_S|X)$ to a predefined threshold θ. The claim is accepted if $P(s_S|X) > \theta$, and rejected otherwise [36, 65].

Bibliography

[1] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*. Englewood Cliffs, NJ: Prenntice-Hall, 1978.

[2] J. J. Wolf, "Efficient acoustic parameters for speaker recognition," *J. Acoust. Soc. Amer.*, vol. 51, no. 2, pp. 2044–2055, 1972.

[3] S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Trans. Acoust. Speech, and Signal Process.*, vol. 29, no. 2, pp. 254–272, Apr. 1981.

[4] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using gaussian mixture speaker models," *IEEE Trans. Speech and Audio Process.*, vol. 3, no. 1, pp. 4–17, Jan. 1995.

[5] D. A. Reynolds, "Speaker identification and verification using gaussian mixture speaker models," *Speech Commun.*, vol. 17, pp. 91–108, Aug. 1995.

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

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

[8] B. S. Atal, "Automatic speaker recogntion based on pitch contours," *J. Acoust. Soc. Amer.*, vol. 52, no. 6, pp. 1687–1697, 1972.

[9] P. Thevenaz and H. Hugli, "Usefulness of the LPC-residue in text-independent speaker verification," *Speech Commun.*, vol. 17, pp. 145–157, Aug. 1995.

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

[11] B. Yegnanarayana, K. S. Reddy, and S. P. Kishore, "Source and systsem feature for speaker recognition using AANN Models." Proc. IEEE Int. Con. Acoust. Speech and Signal Process., Salt Lake City, UT, USA, May 2001, pp. 409–412.

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

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

[14] L. Mary and B. Yegnanarayana, "Prosodic features for speaker verification," in *INTERSPEECH, Pittsburg, Pennsylvania*, 2006, pp. 917–920.

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

[16] N. Wang, P. C. Ching, and T. Lee, "Exploration of vocal excitation modulation features for speaker recognition," in *Proc. INTERSPEECH-09, Brighton UK*, 2009, pp. 892–895.

[17] M. Chetouani, M. Faundez-Zanuy, B. Gas, and J. L. Zarader, "Investigation on LP-residual representation for speaker identification," *Pattern Recognition*, vol. 42, pp. 487–494, Nov. 2009.

[18] L. Mary and B. Yegnanarayana, "Extraction and representation of prosodic features for language and speaker recognition," *Speech Commun.*, vol. 50, pp. 782–796, 2008.

[19] A. G. Adami, R. Mihaescu, D. A. Reynolds, and J. J. Godfrey, "Modeling prosodic dynamics for speaker recognition," in *Int. Conf. on Acoust. Speech and Signal Process., Hong Kong*, April 2003, pp. IV 788–791.

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

[21] J. R. Deller, J. H. Hansen, and J. G. Proakis, *Discrete Time Processing of Speech Signals*, 1st ed. Upper Saddle River, NJ, USA: Prentice Hall PTR, 1993.

[22] J. Laver, *Principles of Phonetics*. Cambridge Textbooks in Linguistics, 1994.

[23] B. S. Atal, "Automatic recognition of speakers from their voices," *Proc. IEEE*, vol. 64, no. 4, pp. 460–475, Apr. 1976.

[24] J. Makhoul, "Linear prediction: A tutorial review," *Proc. IEEE*, vol. 63, no. 4, pp. 561–580, Apr. 1975.

[25] T. C. Feustel, G. A. Velius, and R. J. Logan, "Human and machine performance on speaker identity verification," *Speech Technology*, pp. 169–170, 1989.

[26] G. R. Doddington, "Speaker recognition- Identifying people by their voices," *Proc. IEEE*, vol. 73, no. 11, pp. 1651–1664, Nov. 1985.

[27] B. Yegnanarayana and S. P. Kishore, "AANN: An alternative to GMM for pattern recognition," *Neural Networks*, vol. 15, no. 3, pp. 459 – 469, Apr. 2002.

[28] L. Mary and B. Yegnenarayana, "Neural network models for capturing duration and intonation knowledge for language and speaker identification." Proc. Int. Conf. Cognitive Neural System, Boston, Massachusetts, May 2004.

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

[30] J. P. Campbell, Jr., "Speaker recognition: A tutorial," *Proc. IEEE*, vol. 85, no. 9, pp. 1437–1462, Sept. 1997.

[31] D. O' Shaughnessy, "Speaker recognition," *IEEE Acoust., Speech, Signal Process. Mag.*, vol. 3, pp. 4–17, Oct. 1986.

[32] H. Gish and M. Schmidt, "Text- independent speaker identification," *IEEE Signal Process. Mag.*, vol. 11, pp. 18–32, Oct. 1994.

[33] S. Furui, "Speaker-dependent feature extraction, recognition and processing techniques," *Speech Commun.*, vol. 10, pp. 505–520, Dec. 1991.

[34] F. Socing, A. Rosenberg, L. Rabiner, and B. Juang, "A vector quantization approach to speaker recognition," *Proc. Int. Conf. Acoust., Speech and Signal Processing, Tampa, FL*, vol. 10, pp. 387–390, Apr. 1985.

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

[36] F. Bimbot, J. F. Bonastre, C. Fredouille, G. Gravier, M. I. Chagnolleau, S. Meignier, T. Merlin, O. J. Garcia, P. Delacretaz, and D. A. Reynolds, "A tutorial on text-independent speaker verification," *EURASIP J. Applied Signal Process.*, vol. 2004, no. 4, pp. 430-451, 2004.

[37] Y. Linde, A. Buzo and R. M. Gray, "An algorithm for vector quantizer design," *IEEE Trans. on Communication*, vol. 28, pp. 84–95, Jan. 1980.

[38] A. E. Rosenberg, "Automatic speaker verification: A review," *Proc. IEEE*, vol. 64, no. 4, pp. 475–486, Apr. 1976.

[39] F. Itakura, "Minimum prediction residual principle applied to speech recogntion," *IEEE Trans. Acoust., Speech and Signal Proc.*, vol. 23, no. 1, pp. 67–72, Feb. 1975.

[40] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The DET curve in assessment of detection task performance," in *Proc. Eur. Conf. on Speech Communication Technology, Rhodes, Greece*, vol. 4, 1997, pp. 1895–1898.

[41] H. S. Jayanna and S. R. M. Prasanna, "Analysis, feature extraction, modeling and testing techniques for speaker recognition," *IETE Technical Review*, vol. 26, no. 3, pp. 181–190, May-June 2009.

[42] Kenneth N. Stevens, *Acoustic Phonetics*, 4th ed. Cambridge, Massachusetts: The MIT Press, 2000.

[43] T. F. Quatieri, *Discrete-Time Speech Signal Processing: Principles and Practice*, 1st ed. Prentice Hall, 2001.

[44] J. Benesty, M. M. Sondhi, and Y. A. Huang, *Springer Handbook of Speech Processing*. Springer-Verlag New York, Inc., 2008.

[45] L. Rabiner and B.-H. Juang, *Fundamentals of speech recognition*. Upper Saddle River, NJ, USA: Prentice Hall PTR, 1993.

[46] R.E. Slyh, W. T. Nelson and E. G. Hansen, "Analysis of mrate, shimmer, jitter, and F0 contour features across stress and speaking style in the SUSAS database," vol. 4. in Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing, March 1999, pp. 2091–2094.

[47] M. Farrus, J. Hernando and P. Ejarque, "Jitter and shimmer measurement for speaker recognition," *TALP Research Center, Department of Signal Theory and Communication, Universitat Politcnica de Catalunya, Barcelona, Spain*, 2001.

[48] *http://www.nist.gov/speech/tests/spk/2001/doc/2001-spkrec-evalplan-v05.9.pdf*, The NIST year 2001 speaker recognition evaluation plan, 2001.

[49] R. Veldhuish, "A computationally efficient alternative for the Liljencrants-Fant model and its perceptual evaluation," *J. Acoust. Soc. Amer.*, vol. 103, no. 1, pp. 566–571, Jan. 1998.

[50] D. W. Farnsworth, "High speed motion pictures of the human vocal cords," *Bell Labs. Rec.*, vol. 18, pp. 203–208, 1940.

[51] T. V. Ananthapadmanabha and G. Fant, "Calculation of true glottal flow and its components," *Speech Commun.*, vol. 1, pp. 167–184, Dec. 1982.

[52] B. Yegnanarayana and R. N. J. Veldhuis, "Extraction of vocal-tract system characterstics from speech signals," *IEEE Trans. Speech Audio Proc.*, vol. 6, no. 4, pp. 313–327, Jul. 1998.

[53] T. C. Feustel, G. A. Velius and R. J. Logan, "Human and machine performance on speaker identity verification," *Speech Technology*, pp. 169–170, 1989.

[54] M. Przybocky and A. Martin, "The NIST-1999 speaker recognition evaluation- An overview," *Digital signal processing*, vol. 10, pp. 1–18, 2000.

[55] L. Cohen, "Time frequency distribution: A review," *IEEE Proc.*, vol. 77, pp. 941–979, July 1989.

[56] ——, *Time-Frequency Analysis: Theory and Application*, ser. Signal Processing Series. Englewood Cliffs: Prentice Hall, 1995.

[57] K. S. R. Murthy and S. R. M. Prasanna, "Speaker specific information from residual phase," *in Proc. Int. Conf. on Signal Processing Communication, Bangalore, India*, Dec. 2004.

[58] "Nist speaker recognition evaluation plan," in Proc. NIST Speaker Recognition Workshop, College Park, MD, 2003.

[59] J. Gudnanson and M. Brookes, "Voice source cepstrum coefficients for speaker identification," in *Int. Conf. on Acoust. Speech and Signal Process. Las Vegas, Nevada, USA*, 2008, pp. 4821–4824.

[60] G. Strang and T. Nguyen, *Wavelets and Filter Banks*. Wellesley-Cambridge Press, 1996.

[61] D. Talkin, *A robust algorithm for pitch tracking (RAPT), in Speech Coding and Synthesis, W. B. Kleijn and K. K. Paliwal, Eds. New York: Elsevier, 1995.*

[62] B. Yegnenarayana and K. S. R. Murthy, "Event based instantaneous fundamental frequency estimation from speech signals," *IEEE Trans. Audio, Speech and Language Process.*, vol. 17, no. 4, pp. 614–624, May 2009.

[63] K. S. R. Murthy and B. Yegnanarayana, "Epoch extraction from speech signal," *IEEE Trans. Audio Speech and Language Process.*, vol. 16, no. 8, pp. 1602–1613, November 2008.

[64] K. S. R. Murty, S. R. M. Prasanna, and B. Yegnanarayana, "Speaker specific information from residual phase," in *Int. Conf. on signal proces. and comm. (SPCOM)*, 2004.

[65] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech Commun.*, vol. 52, pp. 12–40, 2009.

[66] L. P. Heck, Y. Konig, M. K. Sonmez, and M. Weintraub, "Robustness to telephone handset distortion in speaker recognition by discriminative feature design," *Speech Commun.*, vol. 31, no. 2-3, pp. 181–192, June 2000.

[67] D. A. Reynolds and R. C. Rose, "Robust text -independent speaker identification using gaussian mixture speaker models," *IEEE Trans.Speech and Audio Proc.*, vol. 3, no. 1, pp. 72–83, Jan. 1995.

[68] J. A. Bilmes, "A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models," *International Computer Science Institute, Berkeley , CA*, April 1998.

[69] Y. Linde, A. Buzo, and R. Gray, "An algorithm for vector quantizer design," *IEEE Trans. Communications*, vol. 28, no. 1, pp. 84–95, Jan. 1980.

[70] B. S. Atal, "Effetiveness of linear prediction characterstics of the speech wave for automatic speaker identification and verification," *J. Acoust. Soc. Amer.*, vol. 55, no. 6, pp. 1304–1312, 1974.

[71] T. V. Ananthapadmanabha and B. Yegnanarayana, "Epoch extraction from linear prediction residual for identification of closed glottis interval," *IEEE Trans. on Acoust., Speech and Signal Process.*, vol. ASSP-27, pp. 309–319, Aug. 1979.

[72] M. Farrus and J. Hernando, "Using jitter and shimmer in speaker verification," *IET signal proc.*, vol. 3, no. 4, pp. 247–257, Nov 2009.

[73] S. R. M. Prasanna, B. Yegnanarayana, J. Praveen, and H. Hermansky, "Analysis of confusion matrix to combine evidence for phoneme recognition," in *IDIAP Research Report*, July 2007.

[74] D. J. Mashao and M. Skosan, "Combining classifier decisions for robust speaker identification," *Pattern Recognition*, vol. 39, pp. 147–155, Jan. 2006.

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

[76] M. R. Iseli and A. Alwan, "Inter- and intra-speaker variability of glottal flow derivative," in *Int. conf. on Spoken Language Processing (ICSLP, 2000), Beijing, Chaina*, 2000.

[77] J. kominek and A. Black, "CMU-Arctic speech database," in *5th ISCA Speech Synthesis Workshop, Pittsburg, PA*, 2004, pp. 223–224.

[78] H. Strik, "Automatic parameterization of differentiated glottal flow: Comparing methods by means of synthetic flow pulses," *J. Acoust. Soc. Amer.*, vol. 103, no. 5, pp. 2659–2669, May 1998.

[79] R. Carlson, G. Fant, C. Gobl, B. Granstrom, I. Karlsson, and Q.-G. Lin, "Voice source rules for text-to-speech synthesis," in *Int. conf. on Acoust. Speech and Signal Process., Glasgow, Scotland,*, vol. 1, 1989, pp. 223–226.

[80] Y. Qi and N. Bi, "A simplified approximation of the four-parameter LF model of voice source," *J. Acoustic. Soc. Amer.*, vol. 96, no. 2, pp. 1182–1185, August 1994.

[81] K. S. R. Murthy and B. Yegnanarayana, "Characterization of glottal activity from speech signal," *IEEE Signal Process. Lett.*, vol. 16, no. 6, pp. 469–472, June 2009.

[82] B. Yegnanarayana and S. R. M. Prasanna, "Analysis of instantaneous F0 contours from two speakers mixed signal using zero frequency filtering," in *Int. conf. on Acoust. Speech and Signal Process.*, Dallas, Texas, USA, 2010, pp. 5074–5077.

[83] V. Zue, S. Seneff, and J. Glassa, "Speech database development at MIT: Timit and beyond," *Speech Communication*, vol. 9, no. 4, pp. 351–356, August 1990.

[84] C. Jankowski, A. Kalyanswamy, S. Basson, and J. Spitz, "NTIMIT: A phonetically balanced, continuous speech, telephone bandwidth speech database," in *Int. conf. on Acoust. Speech and Signal Process.*, Albuquerque, NM, 1990, pp. 109–112.

[85] S. Furui, "Comparison of speaker recognition methods using static features and dynamic features," *IEEE Trans. Acoust., Speech and Signal Process.*, vol. ASSP-29, no. 3, pp. 342–350, June 1981.

[86] S. Pruzansky and M. V. Mathews, "Talker-Recognition procedure based on Analysis of Variance," *J. Acoust. Soc. Amer.*, vol. 36, no. 11, pp. 2041–2047, 1964.

[87] R. O. Duda and P. E. Hart, *Pattern Classification*, 2nd ed. Willy, 2001.

[88] R. Haeb-Umbach, "Investigation on inter-speaker variability in the feature space," in *Int, conf. on Acoust. Speech and Signal Process., Phoenix, AZ*, 1999, pp. 397–400.

[89] A. V. Oppenhiem and R. W. Schafer, *Digital signal processing.* Englewood Cliffs, NJ: Prentice Hall, 1975.

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

[91] M. J. Carey, E. S. Parris, H. L., and S. B., "Robust prosodic features for speaker identification," in *Int. Conf. on Speech and Lang. Process.*, vol. 3, Oct. 1996, pp. 1800–1803.

[92] T. Arai, M. Pavel, H. Hermansky, and C. Avendano, "Syllable intelligibility for temporally filtered LPC cepstral trajectories," *J. Acouctic. Soc. Amer.*, vol. 105, no. 5, pp. 2783–2791, May 1999.

[93] L. Rabiner, M. J. Cheng, A. Rosenberg, and C. A. McGonegal, "A comparative performance study of several pitch detection algprithms," *IEEE Trans on Acoust., Speech, Signal Processing*, vol. 24, pp. 394–418, 1976.

[94] A. M. Noll, "Cepstrum pitch determination," *J. Acoust. Soc. Amer.*, vol. 41, no. 2, pp. 293–309, June 1967.

[95] L. Xu, A.Krzyzak, and C. Y. Suen, "Methods of combining multiple classifiers and their applications to handwriting," *IEEE Trans. Syst., Man Cybern.*, vol. 22, no. 3, pp. 412–435, May-June 1992.

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

[97] G. K. Parikh, "The effect of noise on the spectrum of speech," Master's thesis, The University of Texas, Dallas, Aug. 2002, http://www.utdallas.edu/~loizou/thesis/gaurang_ms_thesis.pdf.

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

[99] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust. Speech and Signal Process.*, vol. 28, no. 28, pp. 357–366, Aug. 1980.

[100] S. Chakroborty, A. Roy, and G. Saha, "Improved closed set text-independent speaker identification by combining MFCC with evidence from flipped filter banks," *Int. J. of Signal Process.*, vol. 4, no. 2, pp. 114–122, June 2007.

[101] H. Fenglei and W. Bingxi, "Text-independent speaker recognition using support vector machine," in *Int. Conf. on Info-tech and Info-net, Beijing*, vol. 3, 2001, pp. 402–407.

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

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

[104] R.Padmanabhan and H. A. Murthy, "Acoustic feature diversity and speaker verification," in *INTERSPEECH 2010, Sept., Makuhari, Chiba, Japan*, 2010, pp. 2010–2013.

[105] J. G. Proakis and D. G. Manolakis, *Digital signal processing-principles, algorithms, and applications*, 3rd ed. Prentice Hall, 1996.

[106] Hari Krishnan P., R. Padmanabhan and Hema A Murthy, "Robust voice activity detection using group delay functions," in *Proc. IEEE International conference on Industrial technology*, 2006, pp. 2603–2607.

[107] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Royal Statis. Soc.*, vol. 39, no. 1, pp. 1–38, Nov. 1977.

[108] Q. Y. Hong and S. Kwong, "A discriminative training approach for text-independent speaker recognition," *Signal Process.*, vol. 85, no. 7, pp. 1449–1463, 2005.

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

LIST OF PUBLICATIONS

Refereed Journals:

1. Debadatta Pati and S R Mahadeva Prasanna, "Speaker recognition from excitation source perspective" *J. IETE Technical Review*, vol. 27, Issue 2, pp. 138-157, Mar. 2010.
2. Debadatta Pati and S R Mahadeva Prasanna, "Subsegmental, segmental and suprasegmental processing of linear prediction residual for speaker information" *Int. J. of Speech technology (Springer)*, vol 14, no. 1, pp. 49-63, Feb. 2011.
3. Debadatta Pati and S R Mahadeva Prasanna, "Speaker recognition using suprasegmental excitation information" *Int. J. of Computer and Communication Technology*, vol. 2, Issue 6, pp. 62-69, 2011.

Manuscripts Submitted

1. Debadatta Pati and S R Mahadeva Prasanna, "A comparative study of explicit and implicit modelling of subsegmental speaker-specific excitation information" Communicated to *Int. J. of Circuits, Systems and Signal Processing (Springer)*, 2011.
2. Debadatta Pati and S R Mahadeva Prasanna, "Processing linear prediction residual in spectral and cepstral domains for speaker information" Communicated to *SADHANA (Springer)*, 2011.
3. Debadatta Pati and S R Mahadeva Prasanna, "Speaker verification using excitation source information" Communicated to *Int. J. of Speech Technology (Springer)*, 2011.

Refereed International/National Conferences:

1. Debadatta Pati and S. R. M. Prasanna, "Non-parametric vector quantization of excitation source information for speaker recognition," in *IEEE Proc. TENCON 2008* , 2008, Hyderabad, India.
2. Debadatta Pati and S. R. M. Prasanna, "Speaker information using subsegmental and segmental analysis of linear prediction residual," *IEEE Proc. NCC*, 2009, Guwahati, India.
3. Debadatta Pati and S. R. M. Prasanna, "Speaker information from subband energies of linear prediction residual," in *IEEE Proc. NCC*, 2010, Chennai, India.
4. Debadatta Pati and S. R. M. Prasanna, "Modeling speaker information from source spectrum," *in IEEE Proc. INDICON*, 2010, kolkata, India.