SIGNIFICANCE OF SOURCE FEATURES FOR SPEAKER RECOGNITION

A THESIS

submitted by

CHEEDELLA S GUPTA

for the award of the degree

of

MASTER OF SCIENCE

(by Research)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY, MADRAS.

APRIL 2003

To my

Mother- **Sarojini Lakshmi**

Wife- **Suneetha**

THESIS CERTIFICATE

This is to certify that the thesis entitled **Significance of Source Features for Speaker Recognition** submitted by **Cheedella S. Gupta** to the Indian Institute of Technology, Madras for the award of the degree of Master of Science (by Research) is a bonafide record of research work carried out by him under our supervision. The contents of this thesis, in full or in parts, have not been submitted to any other Institute or University for the award of any degree or diploma.

Chennai-36

Date:

Prof. B. Yegnanarayana
Dept. of Computer Science and Engg.

ACKNOWLEDGEMENTS

It is with a deep sense of gratitude that I thank Prof. B. Yegnanarayana for his constant guidance and help throughout the course of this work. He introduced me to the fascinating field of "Speaker Recognition". I am very fortunate for having numerous technical discussions with him from which I benefited enormously. I am highly indebted to him for his support.

I especially acknowledge the many useful discussions I had with Dr. C. Chandra Sekhar and Dr. Hema A. Murthy. These discussions helped me to understand some of the subtle research problems better.

Kishore and Sharat have contributed to the development of the speaker recognition system at IIT Madras, which has evolved over a number of years. I have not only borrowed their ideas but also have used their programs liberally. I thank them for their help and cooperation.

My special thanks to Surya, Prasanna, KSR and Nayeem who have been like elder brothers to me, and who helped me immensely on numerous occasions. Thank you, Guru and Dhanu, for patiently going through the thesis and giving your feedback.

I cannot forget to mention the interest shown and help rendered by friends, Anil, Venky and Vinod. I would like to thank all the members of the Speech and Vision Lab who helped me in one way or the other in completing this work.

I am extremely thankful to my mother for her affection and care. I am bereft of words to express my appreciation for the understanding and cooperation of my wife.

-Gupta

ABSTRACT

Keywords: *Speaker recognition; autoassociative neural networks; speaker-specific source information; higher order relations; glottal closure instants.*

Speaker recognition is the task of identifying a person by his/her voice. Speech signal carries information related to not only the message to be conveyed, but also about speaker, language, emotional status of the speaker, environment and so on. In a speaker recognition task the speech signal is processed to extract speaker-specific information. Speech is produced by exciting the time varying vocal tract system with a time varying excitation. Each sound is produced by a specific combination of excitation and vocal tract dynamics. The time varying filter characteristics capture the variations in the shape of vocal tract system in the form of resonances, antiresonances, and spectral roll-off characteristics. Since the vocal tract shape and its dynamics are unique for a given speaker, time varying filter representation has been exploited for developing speaker recognition systems. There is yet another component in speech, which is largely ignored in most speech analysis techniques. It is the residual of the speech signal after the vocal tract characteristics are suppressed from it. No specific attempt has been made in so far in exploring the speaker-specific information present in the residual, which mostly contains the excitation source information.

In this work, issues involved in developing a speaker recognition system using source features from linear prediction residual are addressed. An autoassociative neural network model is used for capturing the speaker-specific source characteristics. The speaker recognition studies show that the residual indeed contains speaker-specific information. Effect of parameters such as data size and network structure on the performance of the speaker recognition is discussed. All the sounds in speech are not equally important for speaker recognition. Therefore, performance of the speaker recognition for different types of sound units is also examined.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

ANN	- Artificial Neural Network
MLFFNN	- MultiLayer FeedForward Neural Network
DTW	- Dynamic Time Warping
AANN	- Autoassociative Neural Network
GMM	- Gaussian Mixture Model
HMM	- Hidden Markov Model
LP	- Linear Prediction
LPC	- Linear Prediction Coefficients
NIST	- National Institute of Standards and Technology
VQ	- Vector Quantization

CHAPTER 1

Introduction to Speaker Recognition

Within the past decade, technological advances such as telebanking and remote collaborative data processing over large computer networks have increased the demand for improved methods of information security. For personal information including medical records, bank accounts and credit history, the ability to verify the identity of individuals attempting to access such data is critical. To date, low-cost methods such as passwords, personal identification numbers (pins) and magnetic cards have been widely used. More advanced security measures have also been developed (e.g., face recognizers, retinal scanners, as well as automatic finger print analyzers). The use of these procedures has been limited by both cost and ease of use. In recent years, speaker recognition (recognizing a person from his/her voice by a machine) and verification algorithms have also received considerable attention. There are several reasons for this interest. In particular, speech provides a convenient and natural form of input, conveys a significant amount of speaker dependent information, and it is inexpensive to collect and analyze.

1.1 PRINCIPLES OF SPEAKER RECOGNITION

1.1.1 Speaker Recognition by Humans

People can reliably identify familiar voices. About 2-3 seconds of speech is sufficient to identify a voice, although performance decreases for unfamiliar voices [1]. Even if duration of the utterances was increased, but played backward (which distorts timing and articulatory cues), the accuracy decreased drastically. Widely varying performance on this backward task suggested that cues to voice recognition vary from voice to voice,

and that voice patterns may consist of a set of acoustic cues from which listeners select a subset to use in identifying individual voices.

Recognition often falls sharply when speakers attempt to disguise their voices [2]. This is reflected in machines, where accuracy decreases when mimics act as impostors. Humans appear to handle mimics better than machines do, easily perceiving when a voice is being mimicked [3]. If the target (intended) voice is familiar to the listener, he often associates the mimic voice with it. Certain voices are more easily mimicked than others, which lends evidence to the theory that different acoustic cues are used to distinguish different voices.

Speaker recognition is one area of artificial intelligence where machine performance can exceed human performance - using short test utterances and a large number of speakers, machine accuracy often exceeds that of humans [3]. This is especially true for unfamiliar speakers, where the training time for humans to learn a new voice is normally very long compared to that for machines. Human performance in adverse conditions was also reviewed in [4], where it was reported that human listeners are adept at using various cues to verify speakers in the presence of acoustic mismatch.

1.1.2 Issues in Speaker Recognition

Speaker recognition by a machine involves three stages. They are: (1) Extraction of features to represent the speaker information present in the speech signal. (2) Modeling of speaker features. (3) Decision logic to implement the identification or verification task. The issues involved in each of these stages are discussed below.

The primary task in a speaker recognition system is to extract features capable of representing the speaker information present in the speech signal. It is known that human beings use high-level features such as speaker dialect, style of speech and verbal mannerisms (for example, use of particular words and idioms, or a particular kind of a laugh) to recognize speakers. Intuitively, it is clear that these features constitute important speaker information. Difficulty arises due to limitations of the existing feature extraction techniques [4]. Current speaker recognition systems use

segmental features such as the shape of the vocal tract to represent the speaker-specific information. These features show significant variations across speakers, but they also show considerable variations from time to time for a single speaker. In addition to this, the characteristics of the recording equipment and transmission channel are also reflected in these features [4].

Once a proper set of feature vectors is obtained, the next task in speaker recognition is to develop a model (prototype) for each speaker. The development of speaker modeling is called the training phase. Feature vectors representing the voice characteristics of the speaker are extracted and used for building the reference models. The performance of a speaker recognition system depends primarily on the effectiveness of the models in capturing the speaker-specific information, and hence this phase plays a major role in determining the performance of a speaker recognition system.

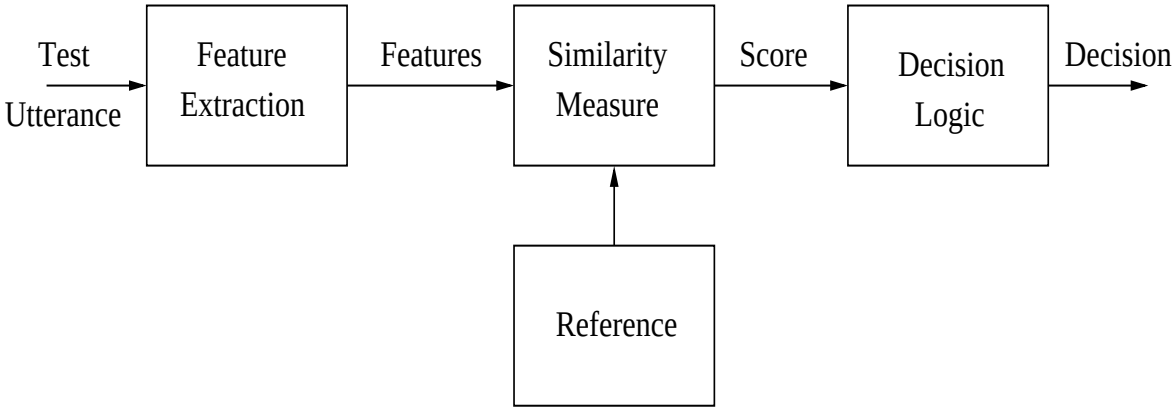

Fig. 1.1: Testing phase of the process.

The final stage in the development of a speaker recognition system is the decision logic stage, where a decision to either accept or reject the claim of a speaker is taken based on the result of matching technique used. Matching techniques are of two types, template matching and probabilistic modeling. Matching generally gives a score which will be a measure of how well the test feature vector matches with the reference feature vector. A decision can be taken based on these scores by fixing some threshold

appropriately. The block diagram of testing phase and decision logic is shown in the Fig. 1.1.

1.1.3 Categories of Automatic Speaker Recognition

Speaker recognition is the general term used to include many different ways of discriminating people based on their voices. The main categories are: Speaker identification and speaker verification.

In speaker identification, a speech utterance from an unknown speaker is analyzed and compared with models of all known speakers. The unknown speaker is identified as the speaker whose model best matches the input utterance. Fig. 1.2 shows the basic structure of speaker identification system. Speaker identification can be a closed set identification or an open set identification. In a closed set identification, it is assumed that the test utterance belongs to one of N enrolled speakers (N decisions). In the case of open set identification, there is an additional decision to be made to determine whether the test utterance was uttered by one of the N enrolled speakers or not, that is, there are $N + 1$ decision levels.

Speaker verification aims to accept or reject the claim of the speaker based on the samples of his speech. If the match between test and reference is above a certain threshold, the claim is accepted. A high threshold makes it difficult for impostors to be accepted by the system, but at the risk of rejecting the genuine person. Conversely, a low threshold ensures that the genuine person is accepted consistently, but at the risk of accepting impostors. Fig. 1.3 shows the basic structure of a speaker verification system.

1.2 APPROACHES FOR SPEAKER RECOGNITION

Early studies on text-independent speaker recognition used averaging of the feature vectors to create reference templates [5] [6]. In [7] the correlation matrices derived from the spectra of relatively long duration of speech signals are used to specify speaker

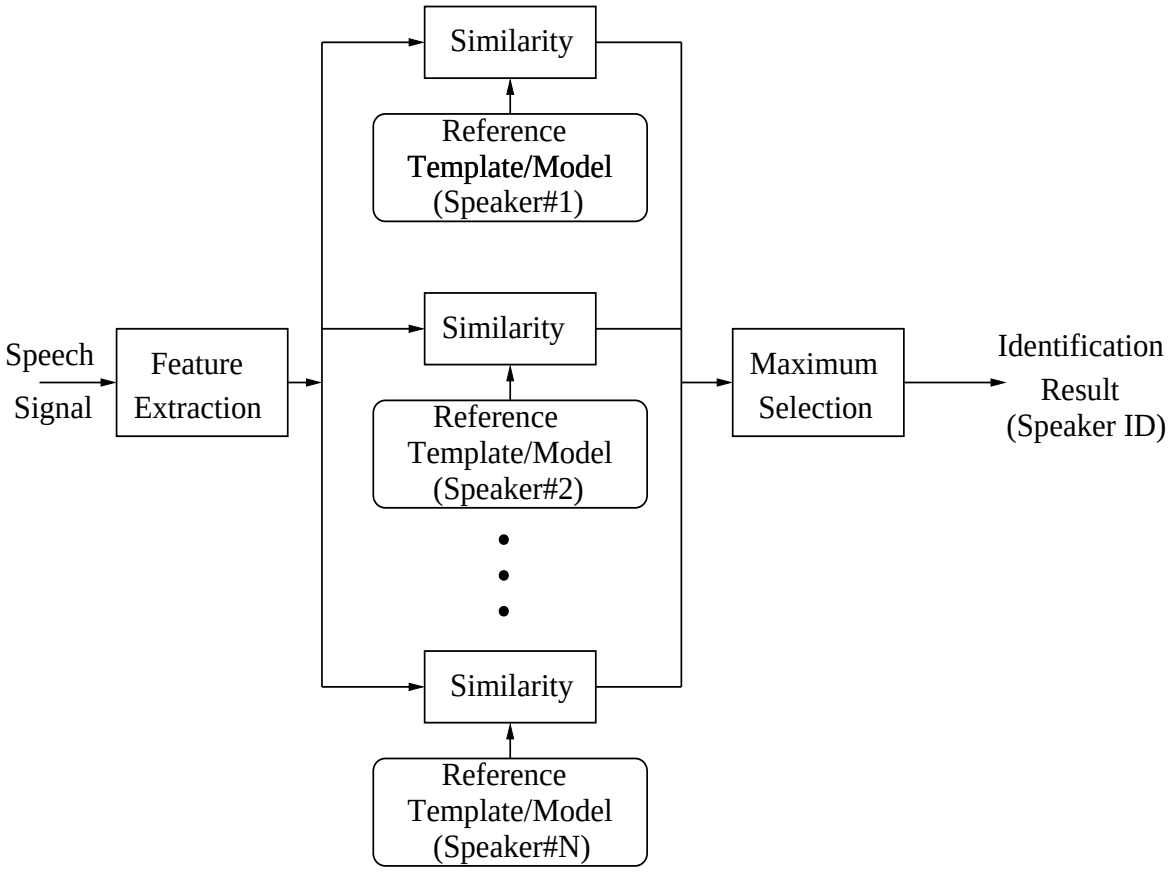

Fig. 1.2: Basic structure of a closed set identification system.

differences. Such methods may not adequately represent the distribution of feature vectors. Hence, the probability distribution of feature vectors are modeled by parametric or nonparametric methods. Models which assume a probability density function are termed parametric. In nonparametric modeling, minimal or no assumptions are made regarding the probability distribution of feature vectors. In this section, we briefly review Gaussian Mixture Model (GMM), Hidden Markov Model (HMM), Vector Quantization (VQ), and neural network based approaches for speaker recognition. GMM and HMM are parametric models. VQ and neural network models are treated as nonparametric models.

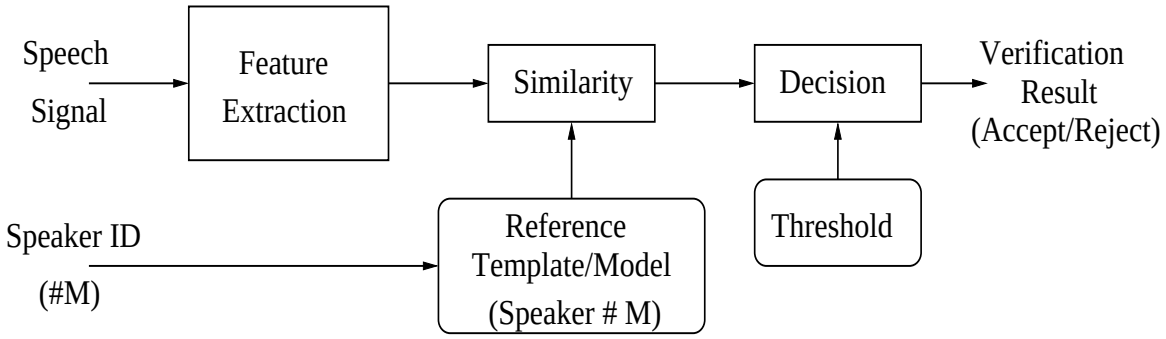

Fig. 1.3: Basic structure of a speaker verification system.

1.2.1 Parametric Approaches

Models which assume a probability density function are termed as parametric. We briefly review the Gaussian Mixture Model and Hidden Markov Models in parametric approaches.

1.2.1.1 Gaussian Mixture Models

The basis for using GMM is that the distribution of feature vectors extracted from an individual's speech data can be modeled by a Gaussian mixture density. For an N-dimensional feature vector denoted as $\boldsymbol{x}$, the mixture density function for speaker s is defined as

$$p(\boldsymbol{x}/\lambda^s) = \sum_{i=1}^{M} \alpha_i^s f_i^s(\ \boldsymbol{x})$$

The mixture density function is a weighted linear combination of M component unimodal Gaussian densities $f_i^s(.)$. Each Gaussian density function $f_i^s(.)$ is parameterized by the mean vector $\boldsymbol{\mu}_i^s$ and the covariance matrix C_i^s using

$$f_i^s(\boldsymbol{x}) = \frac{1}{(2\pi)^{(n/2)}|C_i^s|}\ exp(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_i^s)^T[C_i^s]^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_i^s)),$$

where $[C_i^s]^{-1}$ and $\mid C_i^s \mid$ denote the inverse and determinant of the covariance matrix C_i^s, respectively. The mixture weights α_i^s satisfy the constraint $\sum_{i=1}^{M} \alpha_i^s = 1$. Collectively the parameters of the speaker model λ^s are denoted as $\lambda^s = \{\alpha_i^s, \boldsymbol{\mu}_i^s, C_i^s\}$, $i = 1, 2 \cdots M$. The number of mixture components is chosen empirically for a given data set. The parameters of the GMM are estimated using the iterative expectation-maximization algorithm [8] [9].

A speaker verification system using GMM can be developed as follows [10]: Feature vectors of a given speaker are used to build a speaker-specific GMM (λ^s). The speaker-specific GMM is derived from a universal background model (λ^b), which is trained with feature vectors of several speakers. During the verification phase, the test utterance is given to the universal background model, and a few mixture components which contribute significantly to the likelihood value are noted. The likelihood of the speaker-specific GMM is then computed by considering only the selected mixture components. The claim is rejected or accepted by comparing the log likelihood ratio with the threshold θ using

$$ln \frac{p(\boldsymbol{x}/\lambda^s)}{p(\boldsymbol{x}/\lambda^b)} \begin{array}{c} reject \\ < \\ > \\ accept \end{array} \theta$$

This likelihood ratio is viewed as a means of normalizing the likelihood for the target speaker. It is observed that the performance of GMM-based speaker verification system is dependent on the number of mixture components and also on the database used to build the universal background model [11] [12].

1.2.1.2 Hidden Markov Models

The Hidden Markov Models (HMM) is a doubly embedded stochastic process where the underlying stochastic process is not directly observable. HMMs can be used as probabilistic speaker models for both text-dependent and text-independent speaker recognition [13] [14] [15]. An HMM not only models the underlying speech sounds but also the temporal sequencing of the sounds. This temporal modeling is advantageous

for text-dependent tasks. For text-dependent speaker recognition task, HMM-based methods have achieved significantly better recognition accuracies than Dynamic Time Warping-based methods [16] [17]. But this temporal modeling does not aid in the case of text-independent system, because the sequence of sounds in the test utterance need not be the same as that in the training utterance.

In training phase, an HMM for each speaker is obtained by estimating the parameters of model using feature vectors from the training data. The parameters of HMM are [18]:

- State-transition probability distribution: It is represented by $A = [a_{ij}]$, where

$$a_{ij} = P(q_{t+1} = j | q_t = i) \quad 1 \leq i, j \leq N$$

  defines the probability of transition from state i to j at time t.

- Observation symbol probability distribution: It is given by $B = b_j(k)$, in which

$$b_j(k) = P(\vec{o}_t = \vec{v}_k | q_t = j) \quad 1 \leq k \leq M$$

  defines the symbol distribution in state j, $\quad j = 1, 2, ...N$

- The initial state distribution: It is given by $\prod = [\pi]$, where

$$\pi_i = P(q_1 = i) \quad 1 \leq i \leq N$$

Here, N is the total number of states, and q_t is the state at time t. M is the number of distinct observation symbols per state, and o_t is the observation symbol at time t.

In the testing phase, $P(O|\lambda)$ for each model is calculated, where $O = (o_1 o_2 o_3 .... o_T)$ is the sequence of the test feature vectors. The goal is to find the probability, given the model, that the test utterance belongs to that particular model. The speaker model that gives the highest score is declared as the identified speaker. GMM corresponds to the single-state continuous ergodic HMM [19].

1.2.2 Non-parametric Approaches

1.2.2.1 Vector Quantization

In the Vector Quantization-based method, codebooks are generated for each speaker. Codebooks consists of a small number of representative feature vectors to characterize a speaker [20] [21] [22] [23]. In the training phase, the feature vectors are grouped into certain fixed number of clusters. The representative data is the centroid vector of the cluster. A set of such centroid vectors is known as codebook. Assuming that each speaker's feature vectors will have different distributions, the codebook generated for each speaker will be unique to him. A separate codebook is generated for each speaker. In the testing phase, the test utterance is vector quantized using the codebook of each reference speaker. The VQ distortion, which is nothing but the distance of a test vector from the codebook element that is closest to it, is calculated for each speaker's codebook. This distortion is accumulated over the entire test utterance. The accumulated distance is used to arrive at a decision for speaker recognition. In [24] weights are given to the distortion score for individual codebook elements, assuming that different codebook elements encode different levels of speaker-specific information. Moreover, there can be more than one codebook for each speaker [25] [26].

VQ-based speaker recognition systems are easy to build and are shown to give good results. VQ based speaker recognition system can be evaluated in both text-dependent and text-independent modes [27]. VQ can be considered as a degenerate case of single-state HMM with observation probability being replaced by the distance measure [19].

1.2.2.2 Artificial Neural Network Models

An Artificial Neural Network (ANN) has an input layer, output layer, and one or more hidden layers in between. Each layer consists of processing units, where each unit represents the model of an artificial neuron, and the interconnection between two units has a weight associated with it. ANN models with different topologies perform

different pattern recognition tasks [28] [29] [30]. The capability of a neural network model to discriminate between patterns of different classes is exploited for speaker recognition task [31]. A global classifier for N speakers may perform poorly, as the complexity of the classification task increases with the increase in the value of N [32]. Oglesby *et al.,* [33] proposed one network model for each speaker. The model is trained to discriminate between speech data of a particular speaker and a small set of other speakers (impostors). Recent studies exploiting the mapping capability of neural network models for speaker recognition task can be found in [34] [35] [36] [37].

Theoretical analysis of multilayer perceptron using sigmoidal activation function suggests that these networks may not draw separating hyperplanes in the feature space [38]. This conclusion exposes the inadequacy of the neural network models using sigmoidal activation function for classification task. Some of the other ANN models (radial basis function networks [39], autoassociative networks [40]) were also explored for speaker recognition. It is also to be noted that neural network models are studied to capture the distribution of the data (feature vectors of a speaker) [41].

1.3 ISSUES ADDRESSED IN THIS THESIS

In this thesis, we illustrate the significance of source features for text-independent speaker recognition task. Speech is used to convey a message through a sequence of sound units, which are produced by exciting the time varying vocal tract system with time varying excitation. Each sound is produced by a specific combination of excitation and vocal tract dynamics. The time varying filter characteristics capture the variations in the shape of the vocal tract system in the form of resonances, antiresonances and spectral roll-off characteristics. Since the vocal tract shape and its dynamics are unique for a given speaker, time varying filter representation has been exploited for developing speaker recognition systems. There is yet another component in speech, which is largely ignored in most speech analysis techniques. It is the residual of the speech signal obtained after the vocal tract characteristics are suppressed from the signal. No

specific attempt has been made in exploring the speaker information present in the residual, which contains mostly information about the source of excitation.

The issues involved in developing a text-independent speaker recognition system using source features from the residual are addressed. The speaker characteristics present in the short segments of the Linear Prediction (LP) residual are explored, and it is showed that the residual indeed contains speaker-specific information. Significance of different parameters on the performance of speaker recognition is presented. All the sounds of speech are not equally important for speaker recognition. Some specific sounds tend to be more useful than others. Therefore, speaker recognition performance for different types of sounds is addressed.

All the studies reported in this thesis are performed on six sets of data, each set contains 20 speakers.

1.4 ORGANIZATION OF THE THESIS

This thesis is organized as follows:

Chapter 2 describes features used for speaker recognition, and reviews different systems based on them. To have a better understanding of the features, the speech production mechanism is described. Database used for the recognition studies in this thesis is discussed.

Chapter 3 describes the baseline speaker recognition system using source features. The procedure used to extract the source characteristics from the speech signal and modeling these source characteristics using Autoassociative Neural Networks (AANN) are explained. Performance of the speaker recognition system is discussed.

Chapter 4 explores the speaker characteristics present in the short segments of the LP residual. For comparison, speaker recognition studies using system and source features for different LP orders are made.

Chapter 5 focuses on the effect of different parameters on the performance of the speaker recognition system based on source features. The significance of size of data

in training and testing phases of system is illustrated. The significance of network structure on the performance of the system is explored for different number of units in the hidden layers of AANN. The significance of blocks, containing glottal closure instants on speaker recognition performance is examined.

Chapter 6 deals with the study of effectiveness of different sound units for speaker recognition. Five vowels are considered for this study.

Chapter 7 concludes the thesis by summarizing the work. Scope for future work is also discussed.

CHAPTER 2

System and Source Features for Speaker Recognition

Identity of a speaker may exist in the physiological and behavioral characteristics. The physiological characteristics correspond to the characteristics of the vocal tract system and that of the voice source. The behavioral characteristics are due to the manner in which speakers have learnt to use their speech production apparatus. Automatic speaker recognition systems rely mainly on features derived from the physiological characteristics of the speaker. To have a better understanding of the features derived from the physiological characteristics of the speaker, it is necessary to have knowledge of the speech production mechanism. The following section discusses the speech production mechanism.

2.1 SPEECH PRODUCTION MECHANISM

Speech signal is produced as a result of time varying excitation of the time varying vocal tract system [18]. The schematic diagram is as shown in Fig. 2.1.

Speech production mechanism essentially consists of a vibrating source of sound coupled to a resonating system. For a majority of the sounds produced, the larynx acts as the vibrating source and the air column from larynx to the lips, referred to as the vocal tract acts as the system. But to produce some special sounds called nasal sounds, along with the vocal tract the nasal tract also plays as an important role. The nasal tract begins at the velum and ends at nostrils. When the velum is lowered, the nasal tract is acoustically coupled to the vocal tract to produce the nasal sounds of speech. But it is a known fact that no sound can be produced without a supply of force or energy. It is the breathing mechanism, constituting of the lungs and muscles of the chest and abdomen, that constitute the energy supply. By the use of laryngeal

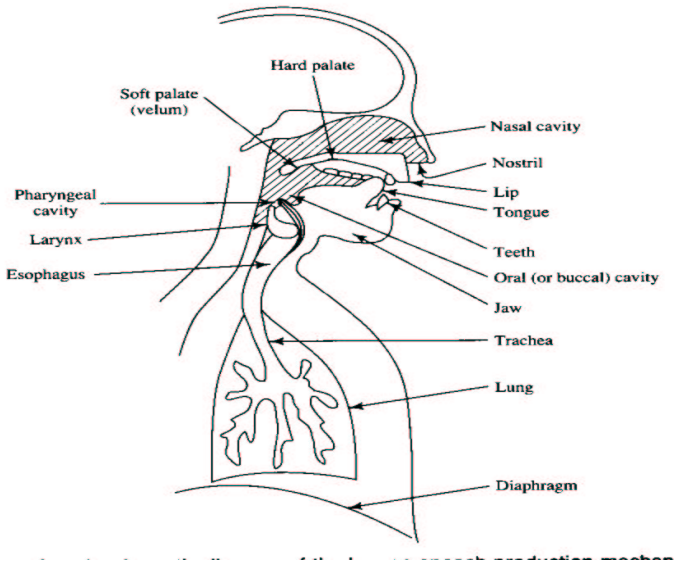

Fig. 2.1: The speech production mechanism.

muscles the vocal chords can be brought together so as to form a shelf across the air way, which leads from the lungs through trachea to pharynx and the mouth. There is a steady flow of air from the lungs into the trachea. While the edges of chords are held together, pressure on the underside of the shelf rises. When it reaches a certain level, it is sufficient to overcome the resistance offered by the obstruction and so the vocal chords open. The ligaments and muscle fibers that make up the vocal chords have a degree of elasticity and having been forced out of position, they tend to return as rapidly as possible to their initial disposition. The pressure rises again and the cycle of opening and closing is repeated.

The studies on musical wind instruments have shown that the dimensions of the air column involved are important in determining the frequency at which resonance would occur. The same principle applies equally in the case of the vocal tract also. Speech is produced as a sequence of sounds. To produce a particular sound, the articulators have to be positioned in a particular way. Hence the state of the vocal chords, shape and size of various articulators change over time to reflect the sound being produced. But when different speakers try to produce same sound, though their vocal tracts are

positioned in a similar manner, the actual shapes will be different due to differences in the anatomical structure of the vocal tract. The main objective in all the studies made on automatic speaker recognition, is to effectively capture the variability due to the anatomical structure of the vocal tract.

Speech signals, as any other real world signals, are produced by exciting a system with a source [42]. From signal processing point of view, the speech production mechanism can be represented as shown in the Fig. 2.2. Vibrations of the vocal folds, powered by air coming from the lungs during exhalation, is the sound source for speech. It sets up a pulse wave in which the pulses are roughly triangular. The amplitude, fundamental frequency, and the shape of the waveform can be modified by the action of the laryngeal muscles. The sound generated in the larynx does not transmit linguistic information. It acts as the source for the information which is imposed upon it by modifications introduced by the vocal tract. Hence, as can be seen from the Fig. 2.2, the glottal excitation forms the source, and the vocal tract forms the system. Speech is produced by exciting the vocal tract by the glottal excitation. The vocal tract is replaced with filter, and the filter coefficients depend on the physical dimensions of the vocal tract. Glottal excitation is replaced with two types of signal generators, impulse train generator for voiced sounds and random number generator for unvoiced and fricative sounds.

In signal processing terms, the speech signal is produced by convolving the source characteristics with the system characteristics. In the frequency domain, this convolution becomes multiplication. If the Fourier Transform of the source is denoted as $G(f)$ and the vocal tract by a time-invariant linear system, represented as $H(f)$, and lip radiation as $R(f)$, then the Fourier Transform of the output signal produced for voiced speech is given by

$$S(f) = G(f)H(f)R(f) \tag{2.1}$$

In general the spectral envelope of the source function $G(f)$ is smooth. The transfer

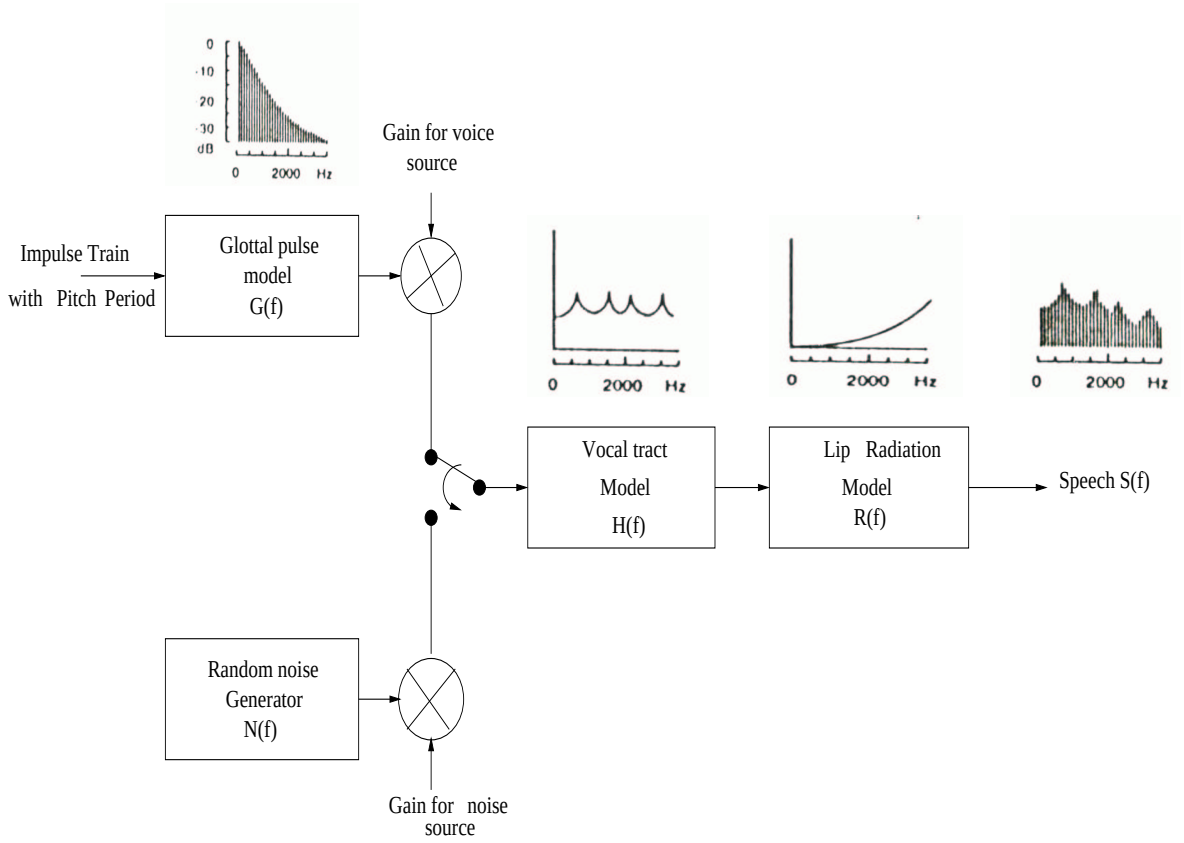

Fig. 2.2: The representation of the speech production mechanism.

function of the vocal tract system $H(f)$, however, is usually characterized by several peaks corresponding to resonances of the acoustic cavities that form the vocal tract. The spectra of the source, system, and the output signal for the production of a voiced sound are as shown in the Fig. 2.2.

2.2 SYSTEM AND SOURCE FEATURES FOR SPEAKER RECOGNITION

Speech is produced as a sequence of sounds. Hence the state of vocal folds, shape and size of various articulators, change over time to reflect the sound being produced. To produce a particular sound the articulators have to be positioned in a particular way. When different speakers try to produce same sound, though their vocal tracts

are positioned in a similar manner, the actual vocal tract shapes will be different due to differences in the anatomical structure of the vocal tract. System features represent the structure of vocal tract. The movements of vocal folds vary from one speaker to another. The manner and speed in which the vocal folds close also varies across speakers. Hence different voices are produced. Source features represent these variations in the vibrations of the vocal folds.

2.2.1 System Features

Speaker recognition systems have been developed mostly using the features of the vocal tract for capturing speaker-specific information. Some of the system features used for speaker recognition task are formants, linear prediction coefficients, and cepstral coefficients.

2.2.1.1 Formants

Formants may be described as the resonances of the vocal tract system. They vary in frequency, relative amplitude and bandwidth according to speech and speaker. Extraction of formant frequencies is a difficult problem in speech processing [43] [44]. Their presence in the spectrum envelope as peaks may be masked by the harmonics of the excitation signal, and thus smoothing is required prior to the use of any peak picking algorithms. Formants and their contours have been used for text-dependent speaker recognition studies [45] [46]. Nasal consonants are found to be effective for speaker recognition [47] [48]. Su *et al.,* used coarticulation between nasal and the following vowel as an acoustic cue for identifying speakers [49]. However, comparative studies on efficiency of different features indicate that distances based on formant frequencies contribute little towards discriminating impostors [50].

2.2.1.2 Linear Prediction Coefficients

The theory of Linear Prediction (LP) is closely linked to modeling of the vocal tract system, and relies upon the fact that a particular speech sample may be predicted by a linear combination of previous samples. The number of previous samples used for prediction is known as the *order* of the prediction. The weights applied to each of the previous speech samples are known as Linear Prediction Coefficients (LPC). They are calculated so as to minimize the prediction error. As a byproduct of the LP analysis, reflection coefficients and log area coefficients are also obtained [51].

A study into the use of LPC for speaker recognition was carried out by Atal [52]. These coefficients are highly correlated, and the use of all prediction coefficients may not be necessary for speaker recognition task [53]. Sambur [54] used a method called orthogonal linear prediction. It is shown that only a small subset of the resulting orthogonal coefficients exhibits significant variation over the duration of an utterance. It is also shown that reflection coefficients are as good as the other feature sets. Naik *et al.,* [55] used principal spectral components derived from linear prediction coefficients for speaker verification task.

2.2.1.3 Cepstral Coefficients

In many applications, Euclidean distance is used as a measure of similarity/dissimilarity. The sharp peaks of the LP spectrum may produce large errors in a similarity test, even for a slight shift in the position of the peaks. Hence, linear prediction coefficients are converted into cepstral coefficients using a recursive relation [56]. Cepstral coefficients represent the log magnitude spectrum, and the first few coefficients model the smooth envelope of the log spectrum [18]. These coefficients can be obtained either from linear prediction coefficients or from the Inverse Fast Fourier Transform (IFFT) of log magnitude spectrum of the speech signal. In both cases, the process results in estimating the vocal tract system characteristics from the speech signal [42].

In an early study, Luck [57] used FFT-based cepstral coefficients for speaker verification. Atal [52] explored the LPC-derived cepstral coefficients, and proved their effectiveness over LPC and other features such as pitch and intensity contours. Furui [56] observed a similar performance of speaker verification for LPC-derived and FFT-based cepstral coefficients. LPC-derived cepstral coefficients take less computation time, and are used even in recent studies for speaker recognition task [58] [59] [60].

2.2.1.4 Mel-Frequency Cepstral Coefficients

The FFT-based cepstral coefficients are computed by taking IFFT of the log magnitude spectrum of the speech signal. The mel-warped cepstrum is obtained by inserting a intermediate step of transforming the frequency scale to place less emphasis on higher frequencies before taking the IFFT. The mel-scale is based on human perception of frequency of sounds [18]. Most of the current speaker verification systems use mel-frequency cepstral coefficients to represent the speaker information present in the speech signal [58] [61] [62].

2.2.2 Source Features

It is interesting to note that human beings recognize people mostly from the source characteristics such as glottal vibrations, and prosodic features such as intonation and duration.

2.2.2.1 Pitch

Pitch information also contributes to the uniqueness of the speaker's voice at the segmental and suprasegmental level. Pitch frequency is the acoustic correlate of the rate of vibration of the vocal folds. The uniqueness of the rate of vibration of the vocal folds is due to the differences in the size of the vocal folds among the speakers, and also due to the speaking style or the accent imposed by the speaker. The physiological

constraints determine the average pitch of the speaker over the entire utterance. In general the average pitch of female speakers will be higher than those of male speakers. This is due to fact that, the female vocal folds are much thinner in comparison with those of male speakers. Pitch information can be extracted from the speech signal using various methods such as zero-crossing, cepstral methods, group delay functions [63] [64] [65]. A discussion of various algorithms for pitch extraction is given in [66]. It has been mentioned in the literature that the pitch features are insensitive to channel variations [67]. They can be reliably extracted from the speech signal recorded when the speaker to microphone distance is large and even from noisy speech data.

Several methods exist in the literature, which discuss attempts made to discriminate between speakers using the average pitch. Attempts have been made to use statistical methods which assume Gaussian distribution to the frame level pitch frequencies [68]. Other methods use the raw pitch frequency at certain anchor points of the utterance as input to a recognition model such as in a neural network model or a statistical model [69] [70]. These methods may fail, as locating the anchor points in the utterance such as syllable nuclei are prone to errors, depending on the noise level of the speech data or the type of text used. In [5], the long-term average value of the fundamental frequency is used for text-independent speaker recognition. Studies have also been made to investigate the usefulness of combining pitch information with spectral features [57] [26]. In [71] a multistage pattern recognition approach is proposed for speaker identification. A two stage classifier with pitch and autocorrelation coefficients is shown to perform better than a single stage classifier using these features together.

2.2.2.2 Intonation

The speaking style determines the pitch pattern of the utterance or the variation of the pitch frequency as a function of time, called intonation. The local variations of the pitch contour is more representative of the speaker than the average pitch of the utterance [72]. Hence, though a speaker's average pitch can be mimicked, it is

indeed difficult for an impostor speaker to reproduce the pitch pattern of the utterance. Intonation is used in text-dependent speaker verification [73]. In this, the similarities of the intonation pattern of the reference and test utterances are captured by using the DTW algorithm. For text-independent speaker verification, the speaker's pitch frequency (f_0) variations were modeled by fitting a piecewise linear model to the f_0 track to obtain a stylized f_0 contour. Parameters of the model are used as statistical features for speaker recognition.

2.2.2.3 Jitter

Jitter is defined as the perturbation of pitch or fundamental frequency. Jitter values are expressed as a percentage deviation of the pitch period. Large values for jitter may be encountered in pathological voices. However, jitter in normal voices is generally less than 1% of the pitch period. Jitter appears a very significant source of aperiodicity in the speech signal. It is generally known that the effect of jitter on the spectra of voiced speech is to widen the harmonic peaks.

2.2.2.4 Shimmer

Shimmer represents the variation in peak amplitudes of the signal in successive pitch periods. Large values for the shimmer variation may be encountered in pathological voices. However, shimmer in normal voices is generally less than about 0.7 dB. The effect of shimmer appears less important than the effect of jitter on the spectrum, and on the perceived aperiodicity.

2.2.2.5 Glottal Flow Derivative

Glottal flow derivative is one of the glottal source feature, which is important to generate natural sounding synthetic speech. It is also useful for characterizing different voices. Both parametric and physical models of glottal source have been developed. One of the most widely used parametric models is the Liljencrants and Fant (LF)

model. This model characterizes one cycle of the derivative of the glottal flow using seven parameters. It has been proved to be very flexible in generating variety of different voices [74]. These parameters can also be used for recognition. An automatic technique for estimating and modeling the glottal flow derivative from speech, and applying the model parameters for speaker identification was reported in [75]. The glottal flow derivative is decomposed into coarse structure, representing the general shape of the glottal flow waveform, and the fine structure comprising aspiration and other perturbations in the flow. The glottal flow derivative is estimated using an inverse filter, which is determined within the time interval of the closure of the vocal folds with in a pitch period. The glottal flow derivative estimate is modeled using LF model to capture its coarse structure, while the fine structure of the flow derivative is represented through energy and perturbation measures. The model parameters were used to develop a Gaussian mixture model for speaker identification.

2.2.2.6 Linear Prediction Residual

The individuality of a speaker associated with the excitation of the vocal tract has been a subject of interest in speaker recognition studies. Linear prediction analysis models the parameters of the vocal tract system and hence the information about the excitation source of the vocal tract is present in the residual signal.

In [76] it has been shown that humans can recognize people by listening to the LP residual signal. Several authors have used pitch frequency present in the residual signal for speaker recognition. LP residual carries much more information than the fundamental frequency alone. This residual is largely ignored. Wakita reported an experiment using LP residual energy for vowel recognition and also for speaker recognition [77]. The author found that most of the time the vowel produced by a speaker is close to one of his/her vowel productions than to the vowels from the other speakers. It was shown that the combination of LPCC coefficients and energy of the LP residual resulted in improvement of performance over the method which used only the Linear Prediction Cepstral Coefficients (LPCC) for speaker recognition [78]. The use of a cep-

strum computed over the LP residual signal was also proposed for speaker recognition [79]. Combination of LP cepstrum and LP residual cepstrum produced a reduction in error rate. In [80] the LP residual is converted into an one-sided autocorrelation sequence, and the DFT (Discrete Fourier Transform) based cepstral coefficients are computed. The cepstral coefficients representing the residual are suitably scaled and appended with the LP cepstral coefficients representing the system features for each frame. The combined feature vectors are used for the speaker identification studies.

In all the studies reported so far, no specific attempt was made to explore the speaker-specific information present in the residual, which contains mostly information about the source of excitation of the vocal tract system. When one listens to the LP residual, one can clearly make out the speaker characteristics present both at the segmental (10-30 ms) level and at the suprasegmental level (1-3 sec). In this work an approach for extracting speaker-specific information present in the short segments of the LP residual is developed.

2.3 SUMMARY

Although speaker-specific information is present in the vocal tract system and the excitation source, speaker recognition systems have been developed mostly using the vocal tract system features. In source features only the pitch frequency have been used. But the LP residual also carries significant information about the speaker. Due to difficulty in extracting the speaker information in the LP residual, no specific attempts has been made to use this information. In this work speaker recognition system using the source information from LP residual is developed. The issues involved in the speaker recognition system along with the effect of various parameters on the performance of the system are discussed in the next chapters.

CHAPTER 3

Baseline Speaker Recognition System Using Source Features

In the previous chapters, we have discussed the concept of source and system features and different techniques for modeling these features. The objective of this chapter is to describe the baseline speaker recognition system using source features and to demonstrate that the source of excitation in speech production contains information useful for speaker recognition. The source information can be derived from the speech signal by removing the spectral features corresponding to the shape of the vocal tract system. The source features of voiced speech for a given speaker are captured from the LP residual by training an Autoassociative Neural Network model.

3.1 LP RESIDUAL FOR SPEAKER RECOGNITION

In the linear prediction analysis of speech each sample is predicted as a linear weighted sum of the past p samples, where p represents the order of prediction [51].

If $s(n)$ is the present sample, then it is predicted by the past p samples as

$$\hat{s}(n) = -\sum_{k=1}^{p} a_k s(n-k) \tag{3.1}$$

The difference between the actual and the predicted sample value is termed as the prediction error or residual, and is given by

$$e(n) = s(n) - \hat{s}(n) = s(n) + \sum_{k=1}^{p} a_k s(n-k) \tag{3.2}$$

The linear prediction coefficients $\{a_k\}$ are determined by minimizing the mean squared error over an analysis frame. The coefficients are obtained by solving the set of p normal

equations

$$\sum_{k=1}^{p} a_k R(n-k) = -R(n), \qquad n = 1, 2, ...p \tag{3.3}$$

where

$$R(k) = \sum_{n=0}^{N-(p-1)} s(n)s(n-k), \qquad k = 0, 1, ...p \tag{3.4}$$

are the autocorrelation coefficients and $\{s(n)\}$ are the speech samples.

The minimum residual in Eq.(3.2) is obtained by passing the speech signal through an inverse filter $A(z)$, which is given by,

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

The LP spectrum $|H(w)|^2$ is given by

$$|H(w)|^2 = \left| \frac{G}{1 + \sum_{k=1}^{p} a_k e^{-jwk}} \right|^2 \tag{3.6}$$

Fig.3.1 shows a segment of voiced speech, its LP residual, short-time spectrum and the 8^{th} order LP spectrum.

The envelope of the short-time spectrum corresponds to the frequency response of the vocal tract shape, thus reflecting the characteristics of the vocal tract system. Typically the vocal tract system is characterized by a maximum of five resonances in the 0-4 kHz range. Therefore an LP order of about 8-14 seems to be most appropriate for a speech signal sampled at 8 kHz. Residual obtained from LP analysis of order 8 may not contain significant information about the vocal tract system. Among the categories of excitation source, voiced excitation contains speaker-specific information, and the corresponding glottal vibrations may be unique for a given speaker. Therefore the residual of voiced speech only is considered for extracting the source information.

LP analysis extracts the second order statistical features through the autocorrelation matrix. Hence, the LP residual does not contain any significant second order statistics corresponding to the vocal tract system. That is why the autocorrelation

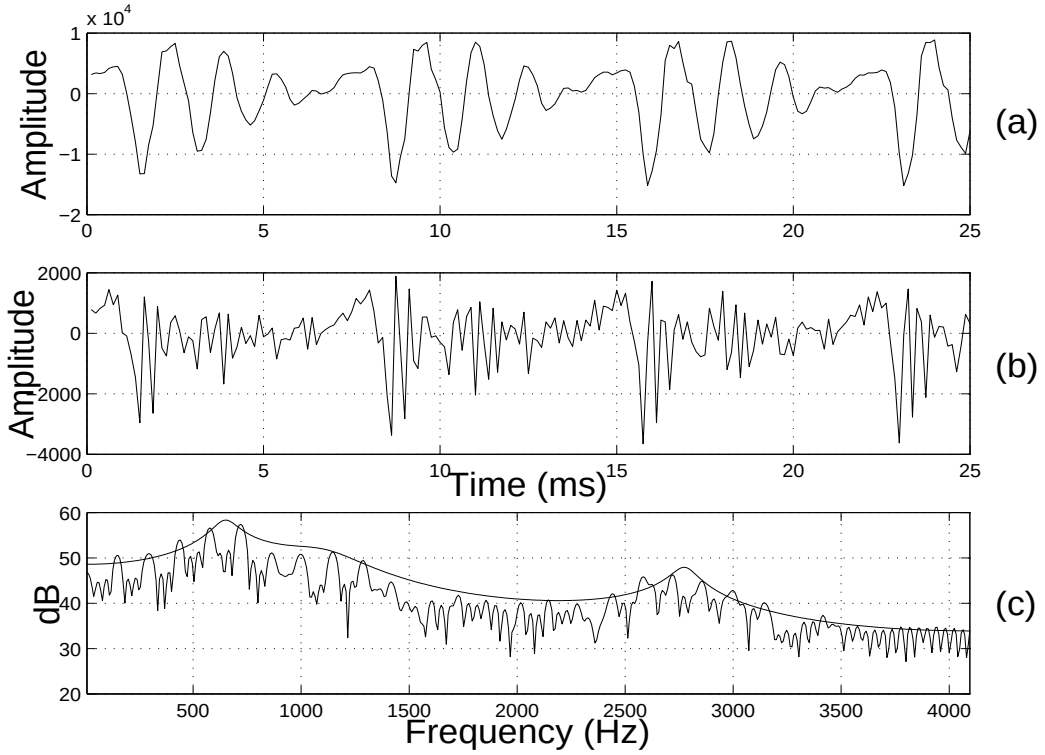

Fig. 3.1: (a) Segment of voiced speech, (b) LP residual and (c) short-time spectrum along with the LP spectrum. Smooth curve in (c) indicates LP spectrum.

function of the LP residual is almost like that for a white noise. However, the source characteristics are still preserved in the LP residual. It is conjectured that the source features may be present in some higher order relations among the samples of the residual signal. Since it is not clear which specific set of parameters need to be extracted from the residual signal to represent the speaker-specific source information, the extraction of such an information may required nonlinear processing. Neural network models are explored to capture the speaker-specific information from the residual are proposed [81].

3.2 AANN MODELS FOR SPEAKER RECOGNITION

AANN models are basically FeedForward Neural Network (FFNN) models which try to map an input vector onto itself, and hence the name autoassociation or identity mapping [82] [83]. It consists of an input layer, an output layer and one or more hidden layers. The number of units in the input and output layers are equal to the size of the input vectors. Number of nodes in the middle hidden layer is less than the number of units in the input and output layers, and this layer is called dimension compression hidden layer. The activation function of the units in the input and output layers are linear, whereas the activation function of the units in the hidden layer can be either linear or nonlinear.

AANN models capture the distribution of the input spectral feature vectors in the feature space effectively [83] [41]. However, when an AANN is presented with raw data, such as samples of speech or LP residual, the interpretation of the behavior of AANN in terms of capturing the distribution of feature vectors is not appropriate. This is because the blocks of residual samples used as input to the network are not feature vectors. The adjacent overlapping blocks of the residual samples may have similar features, corresponding to some nonlinear (higher order) relations among the samples.

Speech signal contains both the second (autocorrelation) and higher order relations among the samples. If the signal is given directly to the network, then the second order correlations present among the samples, called system features, dominate the training of the network. If the second order correlations are removed through the LP analysis, then the network is expected to capture the implicit higher order relations in the LP residual signal. These relations may correspond to the desired speaker-specific features in the excitation source.

For effectively capturing the speaker-specific information present in the LP residual signal, a five layer AANN model with the structure shown in Fig. 3.2 is used. The structure of the network used in our study is $xL48N12N48NxL$, where x refers to the

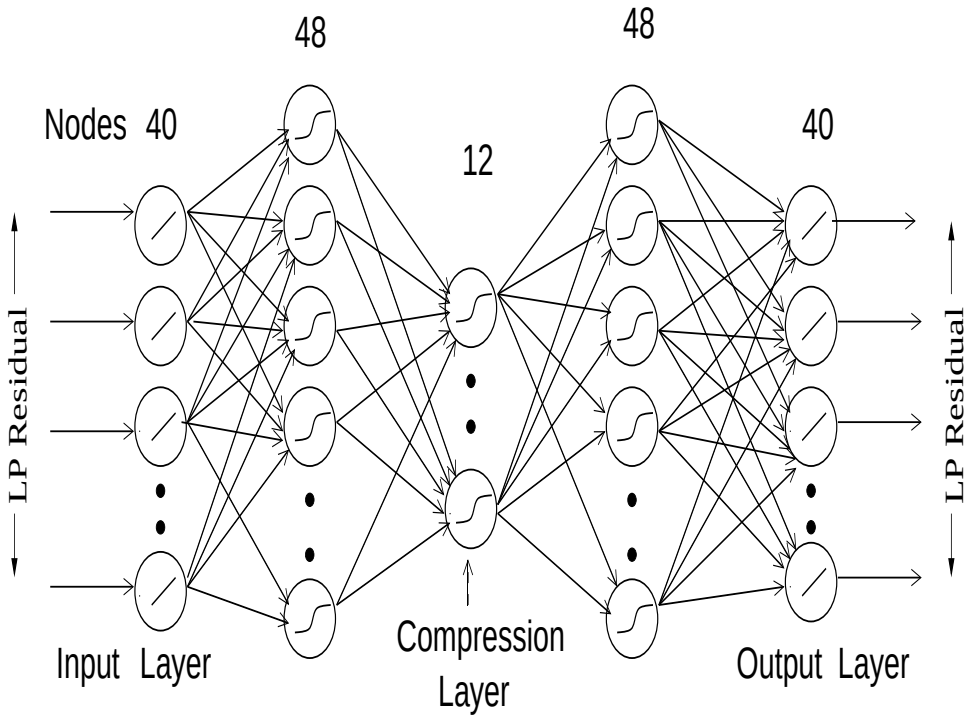

Fig. 3.2: Structure of AANN model used for capturing speaker-specific source features.

number of LP residual samples per frame, L refers to linear units and N to nonlinear units. A $tanh(.)$ is used as the nonlinear activation function. when the input to an AANN consists of samples of random noise, then the network weights will not converge, indicating that the error between the input (desired output) and the actual output is reduced during training, only when there is some relation among the samples. Note also that as the number (x) of LP residual samples per block is increased, then the relations over longer length of the block are captured, and hence the network is expected to capture the speaker-specific information better [84]. But, if the length of the block exceeds a pitch period, then the effect of pitch period may also influence the training of the network. Therefore in the study the number of samples per block are limited to less than a pitch period. The structure of the network used in our study is $40L48N12N48N40L$.

3.3 SPEAKER RECOGNITION SYSTEM BASED ON SOURCE FEATURES

Speaker recognition is a pattern recognition task which involves three phases namely, feature extraction, training and testing. In the feature extraction stage, features representing speaker information are extracted from the speech signal. In the present study the normalized LP residual derived from the speech data is used for training and testing. In the training phase, AANN models are built, one for each speaker, using the training data of the speaker. During the testing phase, the models are tested with the test data. Based on the results with the test data, decision is made about the identity of the speaker.

Only high energy voiced speech data is used in these studies. The algorithm used to detect the voiced frames is based on the amplitude of the speech signal in the time-domain. It also assumes Gaussian distribution for the amplitudes. The speech signal is blocked into frames using specified frame size (20ms) and frame shift (10ms). The maximum positive amplitude in each frame is determined. The sum of mean and a fraction (0.1) of the standard deviation of these positive amplitudes is considered as the maximum amplitude value in the speech signal. Ten percent of the maximum amplitude is taken as the threshold for a frame to be considered as a voiced frame. There is also condition that atleast 30% of the frames should be unvoiced frames. Hence, when the number of unvoiced frames are less than this percentage of the total number of frames, the threshold is progressively increased till the minimum specified number of frames are obtained. The frames above the threshold are considered and they correspond mostly to voiced frames. Since only the high energy voiced part of speech is used, only about 70% of the training and test speech data is available for these studies.

LP residual of the speech signal is computed using an 8^{th} order LP analysis. Blocks or frames of 40 samples of the LP residual, corresponding to 5 ms of data, are used as input to the AANN. Successive blocks are formed with a shift of one sample. Each

block of these 40 samples is normalized to unit magnitude before giving as input to the network. The weights of the network are initialized to random values in the range -1 to 1. The network is trained for 60 epochs using the backpropagation learning algorithm [30] [29]. The choice of the number of epochs was mostly dictated by the time taken for computation of the weights of the AANN. More number of epochs may help in training the model better. One model is created for each speaker. Note that, since the block size is less than a pitch period, only the characteristics of the source within a glottal pulse are used. It is likely that within a pitch period, the blocks contain the region of the glottal closure may be more useful than the blocks in the other regions.

For testing, the LP residual is derived from the high energy voiced segments of the test speech data. Blocks of 40 samples of the LP residual, normalized to unit magnitude, are given as input. The output of each model is compared with its input to compute the squared error for each block. The error (E_i) for the i^{th} block is transformed into a confidence value using $C_i = exp(-\lambda E_i)$, where the constant value $\lambda = 1$ is used throughout this study. The confidence value will be larger for smaller values of the error, that is for blocks matching with the corresponding models. The value of C_i will be low for large error value, thus giving less emphasis to blocks not matching with their respective models. A given test utterance is compared with each of the claimant models to obtain the average confidence value $C = (1/N) \sum_{i=1}^{N} C_i$ for each model, where N is number of blocks in the test utterance. The average confidence value is used to evaluate the performance of the test utterances with respect to a given model.

3.4 SPEECH DATABASE

This study uses speech data collected over three different channels, namely, microphone, telephone and cellular phone. The objective of using speech collected over different channels is to understand the degradation in the quality of data, and also to study the robustness of the proposed approach under different channel degradations.

The microphone data was collected in the laboratory environment from 20 speakers, who volunteered for the task. Two one minute speech utterances were collected over the same channel, but in different sessions for each speaker. One set of utterances was identified for training, and the other set for testing. This set of 20 speakers is termed as MIC.

The telephone channel data was selected from the NIST 99 evaluation development database [58]. The database contains 230 male and 309 female speakers. Among these, 80 male speakers were chosen at random, and four sets TEL1, TEL2, TEL3 and TEL4, each of 20 speakers, were formed. Each speaker is having two minutes speech collected over the same channel in different sessions. Data (1min) for one of the sessions was identified for training, and the data (1min) from the another session for testing.

Cellular phone data was chosen from the NIST 2001 evaluation development database [85]. Among the total 45 male speakers, 20 speakers were chosen at random to form CEL set. One minute of speech data was identified for training and one minute of data for testing. In all the cases the speech was sampled at 8 kHz sampling frequency.

3.5 PERFORMANCE EVALUATION OF SPEAKER RECOGNITION SYSTEM

A model for each speaker is generated as explained above. During testing, the test utterance of each of the 20 speakers belonging to a particular set is tested against all the 20 speakers models of the set. The average confidence value for each of the 20 models for each test utterance is computed, and this confidence value is used to rank the speaker. Ideally, a genuine speaker should have highest confidence value, and thus have rank one. The performance of the system is summarized in Table 3.1. From the table it is evident that source features seem to give good performance for speaker recognition.

One can observe the variation in performance from one data set to the other.

Performance of the speaker recognition is in correspondence with the speech quality. As microphone data is collected in lab environment, noise present in the signal is low. In telephone data speech signal is effected by the characteristics of the transmission path. In cellular mobile standard, speech coding algorithms are used. The purpose of these coders is to compress the speech signal before transmission to reduce the number of bits needed in its digital representation, while keeping an acceptable perceptual quality of the decoded output. For instance, there exist three Global System for Mobile communications (GSM) speech coders, which are referred to as the full rate, half rate and enhanced full rate GSM coders. Their corresponding telecommunication standards [86] [87] are the GSM 06.10, GSM 06.20 and GSM 06.60, respectively. The process of coding and decoding modifies the speech signal, together with other perturbations introduced by the mobile cellular network (channel errors, back ground noise). In the coding process, LP analysis is performed to give set of LPC coefficients, and excitation is selected from a codebook of random sequences [88]. These components are coded to transmit the signal in digital form. In this process of excitation selection from the codebook, excitation signal near to the actual excitation signal is selected. Due to this coding and decoding process, the performance of speaker recognition system is poor for cellular data.

3.6 SUMMARY

In this chapter, issues involved in developing a speaker recognition system based on source features was described and the performance of the system was studied. Speaker recognition performance shows the LP residual indeed contains some speaker-specific information. The performance of the speaker recognition system varies from one data set to another data set. This performance is in accordance with quality of speech in the data set. The significance of source features in LP residual for speaker recognition is verified in the next chapter.

Table 3.1: Performance of speaker recognition system on different data sets. Performance of the system indicates number of speakers identified correctly out of 20 speakers. The values in the parenthesis are percentage recognition.

Data Set	Performance of baseline system
MIC	19 (95%)
TEL1	15 (75%)
TEL2	14 (70%)
TEL3	16 (80%)
TEL4	14 (70%)
CEL	13 (65%)

CHAPTER 4

Significance of Source Features for Speaker Recognition

A crucial parameter in linear prediction analysis of speech is the order of the predictor. There are two usual rules of thumb for estimating the order: (1) Twice the expected number of formants of the vocal tract system, plus two: Ideally, each formant corresponds to a damped sinusoid that can be captured by a pair of roots with correct frequency and damping. The two extra coefficients are to take care of glottal roll-off and radiation. (2) The sampling frequency in kHz: if $F_s = 10$ kHz, for example, then one would use 10 LP coefficients. The rationale is that it takes sound approximately 1 ms to travel from the glottis to lips, and so the statistical structure among 1 ms duration of samples is sufficient to capture the vocal tract resonances.

In the speaker recognition studies discussed in the previous chapter, an 8^{th} order LP analysis was used. One may attribute this performance to the vocal tract information that may be present in the LP residual due to approximate modeling of the vocal tract by LP analysis. Significance of the source features in the LP residual for speaker recognition is verified by studying speaker characteristics present in the LP residual, for different orders of LP analysis. This study also helps in arriving at the optimal LP order for speaker recognition using source features.

4.1 SOURCE INFORMATION IN THE LP RESIDUAL FOR DIFFERENT LP ORDERS

As the envelope of the short-time spectrum corresponds to the frequency response of the vocal tract shape, one can observe the short-time spectrum of the LP residual for different LP orders and the corresponding signal LP spectra to determine the extent of

the vocal tract information present in the LP residual. As the order of the LP analysis is increased, the LP spectrum approximates the short-time spectral envelope better. The envelope of the short-time spectrum corresponds to the frequency response of the vocal tract shape, thus reflecting the vocal tract system characteristics. Typically the vocal tract system is characterized by a maximum of five resonances in the 0-4 kHz range. Therefore an LP order of about 10-14 seems to be most appropriate for a speech signal sampled at 8 kHz. For a low order, say 2, as shown in Fig.4.1(a), the LP spectrum may pick up only the prominent resonance, and hence the residual will still have a large amount of information about the vocal tract system. Thus the spectrum of the residual (Fig.4.1(b)) contains most of the information of the spectral envelope, except for the prominent resonance. On the other hand, if a large order, say 30 is used, then the LP spectrum may pick up spurious peaks as shown in Fig.4.1(e). These spurious peaks influence the corresponding LP residual obtained by passing the speech signal through the inverse filter. The source characteristics in the LP residual may be affected due to the influence of these spurious nulls in the spectrum of the inverse filter.

From above discussion, it is evident that LP residual does not contain any significant features of the vocal tract shape for LP orders in the range 8-20. The LP residual may contain mostly the source information. In the next section we verify the speaker-specific information present in the LP residual for different LP orders.

4.2 EFFECT OF LP ORDER ON SPEAKER RECOGNITION

The extent of speaker information in the LP residual depends on the order of LP analysis. Hence, speaker recognition studies are conducted for different LP orders from 1 to 40.

The LP residual is extracted from the speech signal for a given LP order with frame size 20 ms and frame shift 10 ms. To capture the source characteristics, a block of 40 samples of the residual with one sample shift is used as input to the AANN. Each block

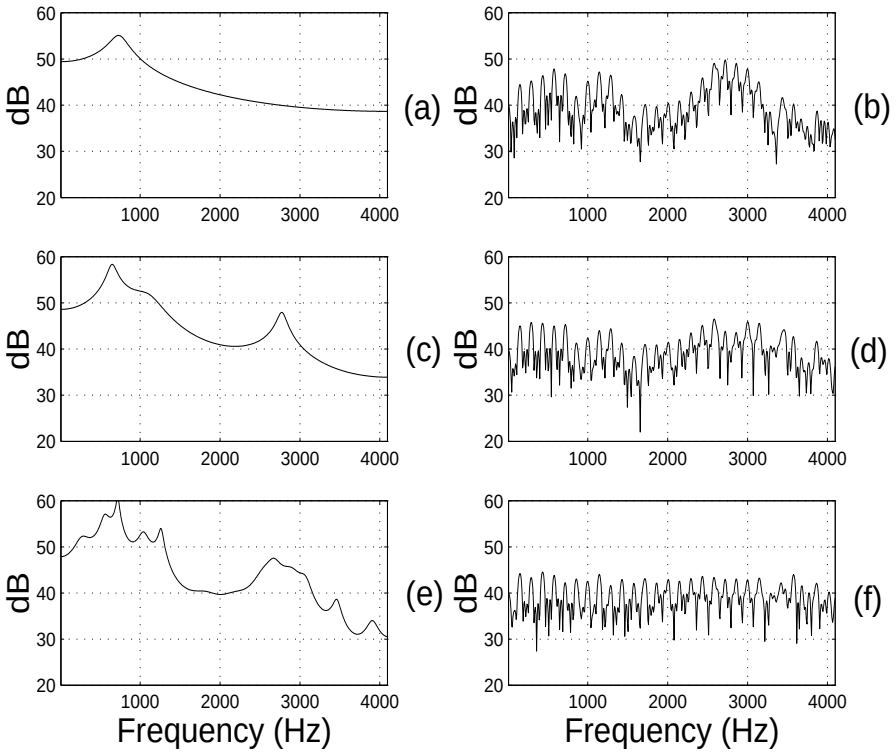

Fig. 4.1: (a) LP spectrum and (b) residual spectrum for LP order 2. (c) LP spectrum and (d) residual spectrum for LP order 8. (e) LP spectrum and (f) residual spectrum for LP order 30.

of 40 samples is normalized to unit magnitude before giving it as input to the network. The model is trained for 60 epochs using the backpropagation learning algorithm [29] [30]. One AANN model is trained for each speaker. The extent of speaker-specific source information in the LP residual may be explained from the learning of the AANN models. The training error curves of AANN models for a speaker are shown in Fig. 4.2 for LP residuals of different LP orders. The training error curve for random noise sequence is also shown in the figure. For low LP orders (< 8), the training error values are low. This is because, in these cases the LP residual is dominated by the speech information, and hence the network tries to capture mostly the speech components, ignoring the speaker-specific information. For LP orders in the range 8-20, the LP residual has mostly the the information about the excitation source, and the network

thus tries to capture the speaker-specific information in the source. For very high LP orders(> 30), the training error is again high due to effect of spurious spectral nulls in the inverse filter obtained in the LP analysis on the speaker-specific information in the resulting LP residual. When the AANN model is trained with random noise sequence, the training error is high and also flat, indicating that no information is present in the data for the network to learn. Thus one can attribute the low training errors achieved for LP orders in the range 8-20 is mainly due to the speaker-specific source information presented in the LP residual.

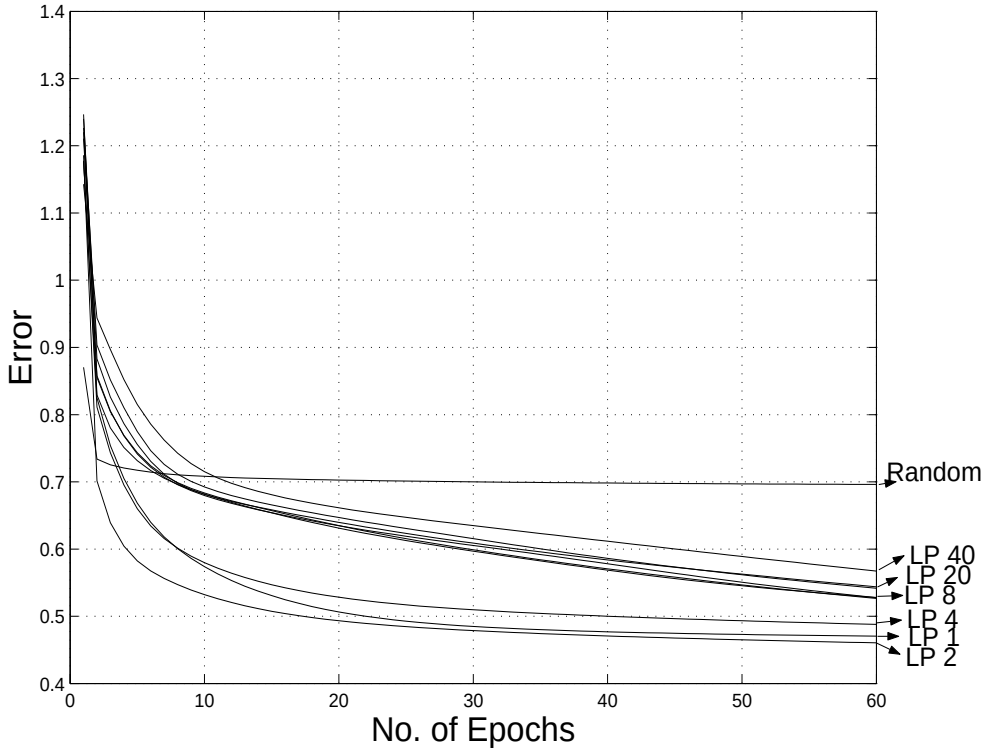

Fig. 4.2: Training error curves of AANN models for LP residuals extracted for different LP orders and random noise.

The presence of speaker-specific source information in the LP residual, and capturing of the same by the proposed AANN models, can also be verified by testing the trained AANN models for speaker recognition. For testing, blocks of 40 samples of the LP residual, normalized to unit magnitude, are given as input. The output of each

model is compared with its input to compute the squared error for each block. The error E_i for the i^{th} block is transformed into a confidence value using $C_i = exp(-\lambda E_i)$, where the constant value $\lambda = 1$ is used. This confidence value will be larger for smaller values of the error, that is for blocks matching with the corresponding models. A given test utterance is compared with each of the claimant models to obtain the average confidence value $C = (1/N) \sum_{i=1}^{N} C_i$ for each model, where N is number of blocks in the test utterance. The average confidence value is used to evaluate the performance of the test utterances with respect to a given model. Testing was conducted for each set as indicated previously. The recognition performance is shown in Fig. 4.3. For lower LP orders (< 8) the performance of the recognition system is low. This is due to the fact that the LP residual has significant spectral envelope information as explained earlier, both in section 4.1 and this section. For the LP orders in the range 8-20, the recognition system gives good performance, as most of the spectral information corresponding to the shape of the vocal tract system is removed. For LP orders greater than 30, the speaker-specific information in the LP residual is masked due to the effects of the spurious spectral nulls in the inverse filter, and also due to noise in the high frequency range, and hence the results are poor.

From the above studies, we may conclude that the optimal range of the LP order for speaker recognition is in the range of 8-20 for speech signals sampled at 8 kHz. We also note that the LP residual indeed has significant speaker-specific source information. The variation of peak performance for each set in Fig. 4.3 is due to the quality of data in the set, the confusability of the speakers and also may be due to inadequate training.

4.3 EFFECT OF LP ORDER ON SPEAKER RECOGNITION BASED ON VOCAL TRACT SYSTEM FEATURES

A study was conducted to understand the presence of speaker information in the vocal tract system features for different LP orders. For this, a separate AANN based speaker

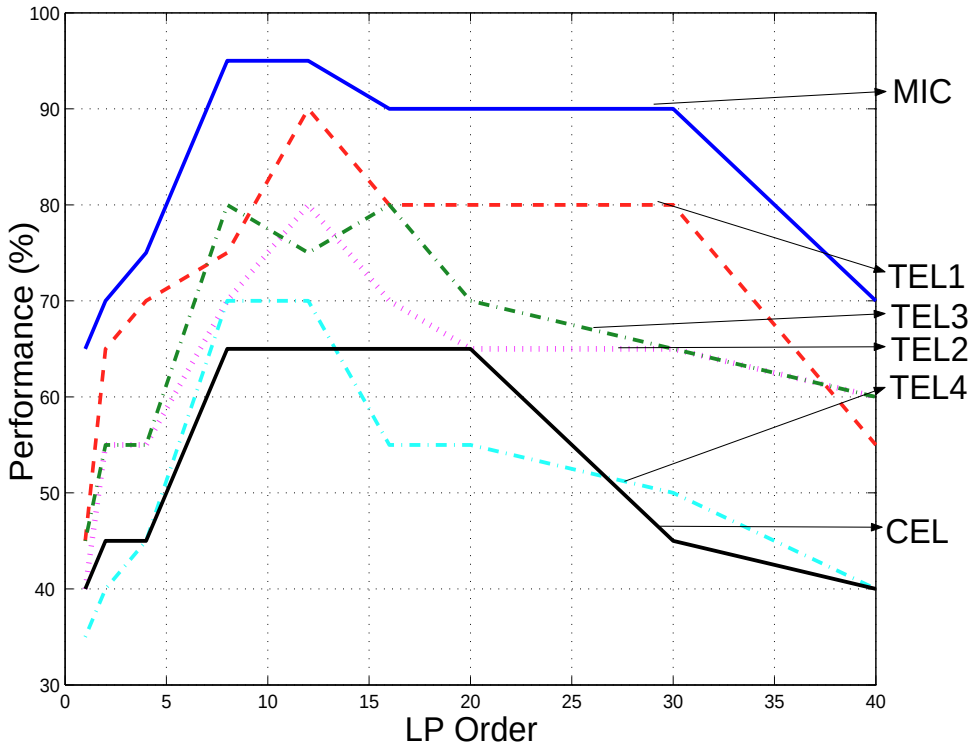

Fig. 4.3: Performance of speaker recognition system for different LP orders. Performance of the system indicates number of speakers identified correctly out of 20 speakers. MIC is microphone data, TEL1 to TEL4 are telephone data and CEL is cellular data.

recognition system using features representing the vocal tract system characteristics is developed [89]. 19 weighted LPCCs are used as the feature vector. The corresponding AANN model is shown in Fig. 4.4. The model captures the distribution of the vocal tract system features of a given speaker [41]. The distributions are usually different for different speakers. Thus, the AANN model trained with the feature vectors of a speaker captures the distribution for that speaker. Each model is trained with feature vectors derived from one minute of speaker data. The feature vectors are computed for every 20 ms frame, separated by 10 ms. The model is trained using backpropagation learning algorithm for 60 epochs. One such model is generated for each speaker.

For testing, the feature vectors extracted from the test utterance are given to the speaker model. The output of each model is compared with its input to compute

the error for each feature vector. The error E_i is defined as $\frac{1}{\ell}\sum_{i=1}^{\ell}\frac{\|\boldsymbol{x}_i-\boldsymbol{y}_i\|^2}{\|\boldsymbol{x}_i\|^2}$, where $\boldsymbol{x}_i$ is the input vector of the model, $\boldsymbol{y}_i$ is the output given by the model, and ℓ is the dimension of feature vector. The error (E_i) for the i^{th} feature vector is transformed into a confidence value using $C_i = exp(-\lambda E_i)$, where the constant value $\lambda = 1$ is used throughout this study. A given test utterance is compared with each of the claimant models to obtain the average confidence value $C = (1/N)\sum_{i=1}^{N} C_i$ for each model, where N is number of feature vectors in the test utterance. The average confidence value is used to evaluate the performance of the test utterances with respect to a given model.

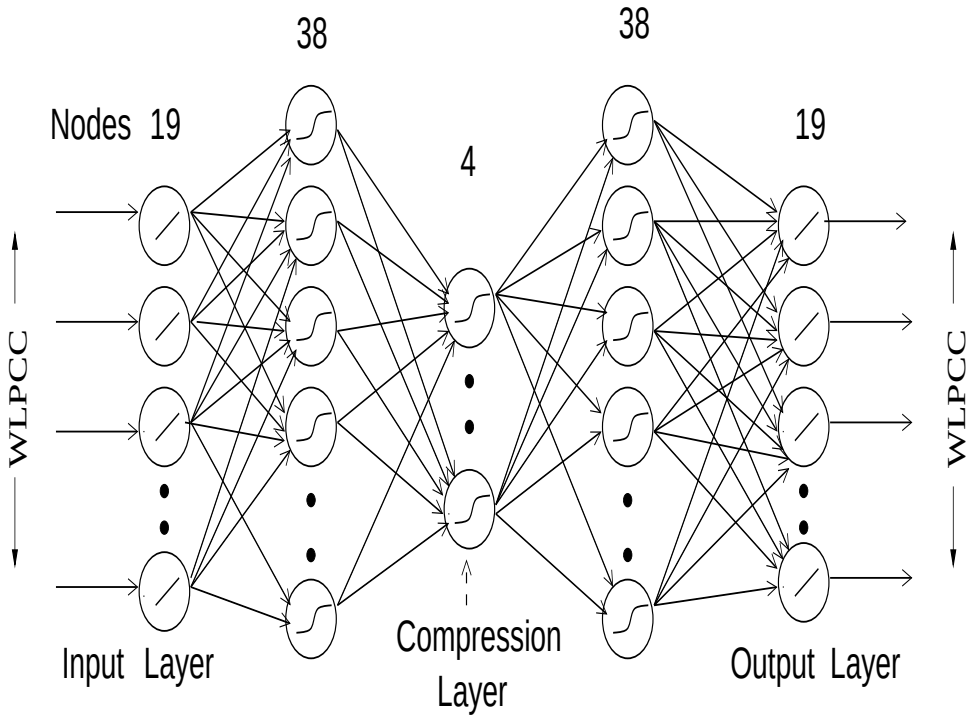

Fig. 4.4: Structure of AANN Model used for capturing speaker specific system features.

As shown in the network structure, 19 weighted LPCCs are used as feature vectors, and in cases where the order of LP analysis is greater than 19, only the first 19 LPCCs are used for the study. The speaker models built using the LPCC features for different LP orders are tested as described earlier. The performance of the system for different

LP orders is shown in Fig. 4.5. For low LP orders (< 8), the performance of the system is low, as it cannot capture speaker information present in all the formants of the vocal tract system. For LP orders in the range 8-20, the speaker-specific vocal tract information is best represented in the LPCC features, and hence the performance is high. For high LP orders (> 30), the performance is again low due to spurious peaks introduced in the LP spectrum.

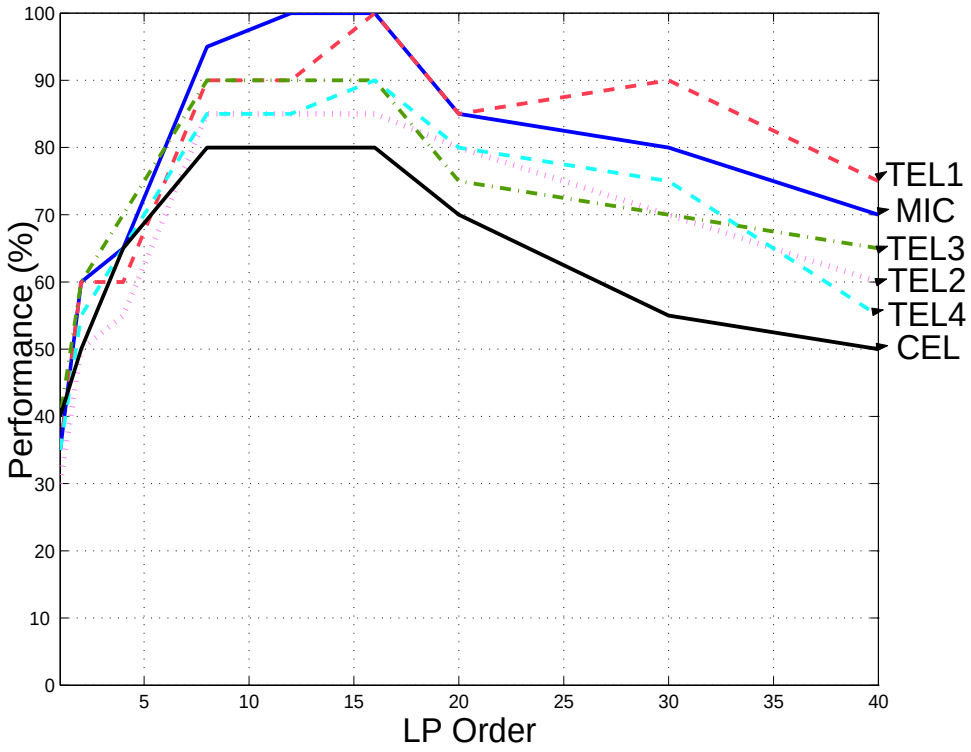

Fig. 4.5: Performance of speaker recognition system based on vocal tract system information for different LP orders. Performance of the system indicates number of speakers identified correctly out of 20 speakers. MIC is microphone data, TEL1 to TEL4 are telephone data and CEL is cellular data.

Thus from both these studies on source and system features, we may conclude that the optimal range of LP order for speaker recognition using speech signals sampled at 8 kHz is 8-20. It is interesting to note that intuitively we feel that for low LP orders (< 8), the missing speaker-specific information due to vocal tract system information is present in the LP residual, and hence can be captured of the speaker recognition system

using the LP residual. But in fact the presence of the vocal tract information in the LP residual degrades the performance. The speaker information is best represented either in the vocal tract system features or in the excitation source features, when the LP order is in the optimal range of 8-20. It is also important to note that the degradation in speech data used for training and testing can affect the performance of the speaker recognition system based on either source features or system features. Hence the performance is poor for noisy CEL data set, compared to MIC data set.

4.4 SUMMARY

Speaker recognition studies using system and source features for different LP orders showed almost similar variations in performance. For low LP orders (< 8) the recognition performance is low. For LP orders in the range 8-20, the systems give good performance. For higher orders, the performance of the systems is again low. The performance variation of both the systems follow the characterization of the vocal tract system. Even though, the vocal tract system is characterized well, the recognition performance due to source features is still comparable to that due to the vocal tract system features. This demonstrates the significance of speaker-specific source information present in the LP residual. Note that, since the block of LP residual is less than a pitch period, only the glottal pulse characteristics presented in the LP residual are captured for speaker recognition.

Performance of the speaker recognition may be improved by optimizing the design parameters. Computation time involved to develop the model for the each speaker and to test is high. This is due to large number of blockss in a given data due to shift of one sample for each block. This can be reduced by optimizing parameters. The effect of different parameters on recognition is examined in the next chapter.

CHAPTER 5

Data and Network Structure for Speaker Recognition

Various parameters used in the recognition system based on source features, described in the previous section can be modified to improve the performance of speaker recognition. For this, a study on the dependence of the speaker recognition system on various parameters is required. The issue of duration of speech data for training and testing is often very important. The structure of the neural network also plays a important role. These issues are discussed in this chapter.

5.1 SIZE OF DATA FOR SPEAKER RECOGNITION

Traditionally speaker recognition systems based on spectral features follow statistical approach [90] [91] [92]. The statistical methods capture the speaker variability in terms of the Probability Density Function (PDF) of the feature vectors of the speaker in the feature space. The performance of these systems depends on the amount of data available for both training and testing. If the data available is small, the distribution of the feature vectors in the feature space is sparse, and hence the recognition performance is poor during testing. In the proposed speaker recognition system based on source features, the speaker-specific information is captured in terms of the higher order relations present among the samples of the residual signal, and not in terms of the PDF of the feature vectors of the speaker. Hence the recognition system based on source features may not require large amount of data for training and testing.

In the speaker recognition studies discussed so far, one minute of speech data was used for generating the speaker models. Using different amount of training data for generating the speaker models, the effect of size of the training data on the performance

of the system can be studied. All the systems are evaluated independently using the one minute test data as explained in the section 3.2. The results are shown in Fig. 5.1. It is evident from the figure that about 6 seconds of data is enough for capturing the speaker-specific information. This is because the speaker-specific information in the LP residual depends less critically on the type of sound unit, as the vocal tract shape information corresponding the sound unit is removed in the residual.

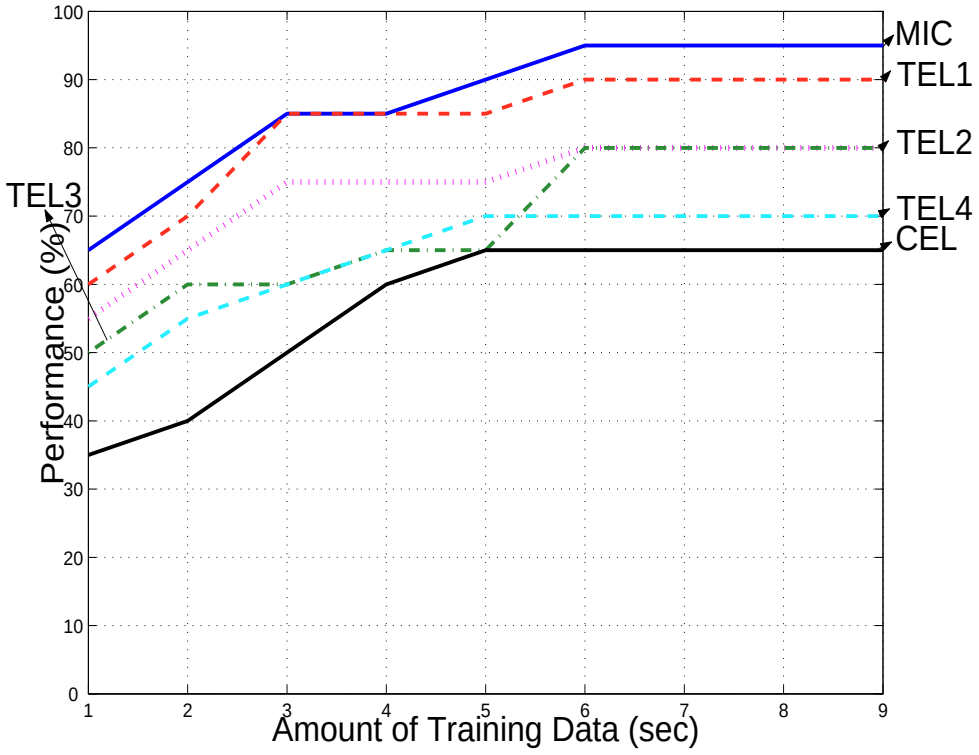

Fig. 5.1: Performance of proposed speaker recognition system based on source features with respect to amount of training data. Performance of the system indicates number of speakers identified correctly out of 20 speakers. MIC is microphone data, TEL1 to TEL4 are telephone data, and CEL is cellular data.

To examine the effect of size of the test data on the performance of the system, different cases are considered, each case using different amount of test data. Models trained with 6 seconds of speaker's data are used in the study. The performance variation with respect to the amount of the test data is shown in Fig. 5.2. From the figure it can be seen that about 5 seconds of data may be sufficient for testing the

models.

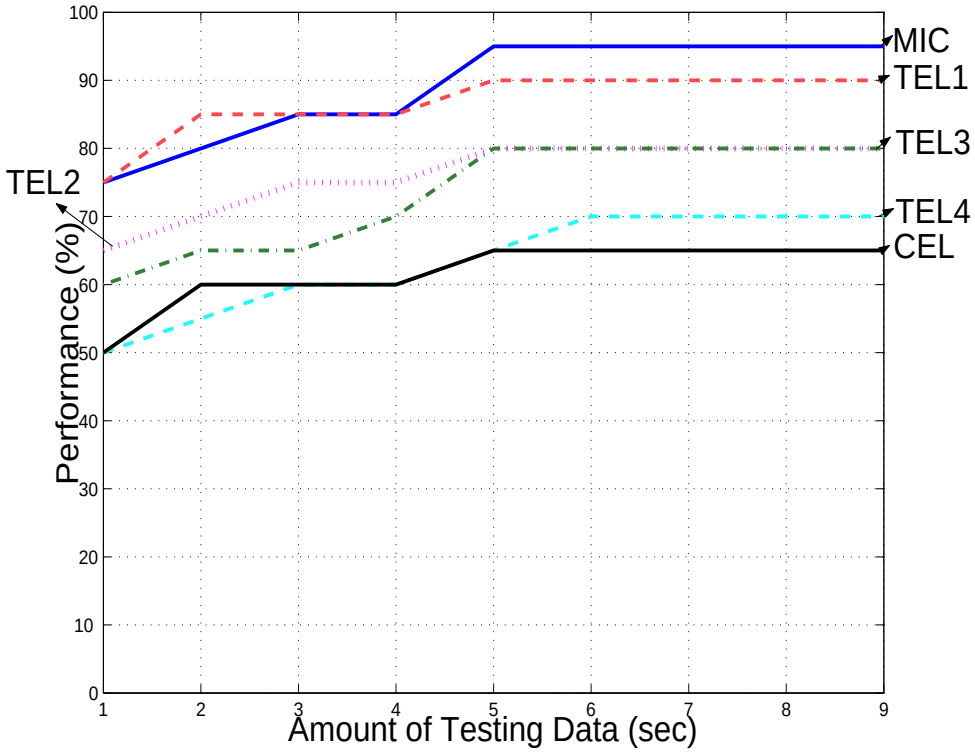

Fig. 5.2: Performance of proposed speaker recognition system based on source features with respect to amount of testing data. Performance of the system indicates number of speakers identified correctly out of 20 speakers. MIC is microphone data, TEL1 to TEL4 are telephone data, and CEL is cellular data.

Results obtained for different amount of training and test data for speaker recognition based on system features are shown in Figure 5.3 and 5.4. The results show significant reduction in the performance when the quantity of data is reduced. This is because the distribution of the system features of a speaker can be captured well only when there is sufficient amount of data representing all the different types of units both during training and testing.

From this study we can conclude that the proposed speaker recognition system based on the source features requires significantly less amount of data. The performance of the system can be improved by generating more than one model for each speaker for a given amount of data, and likewise more tests can be made using differ-

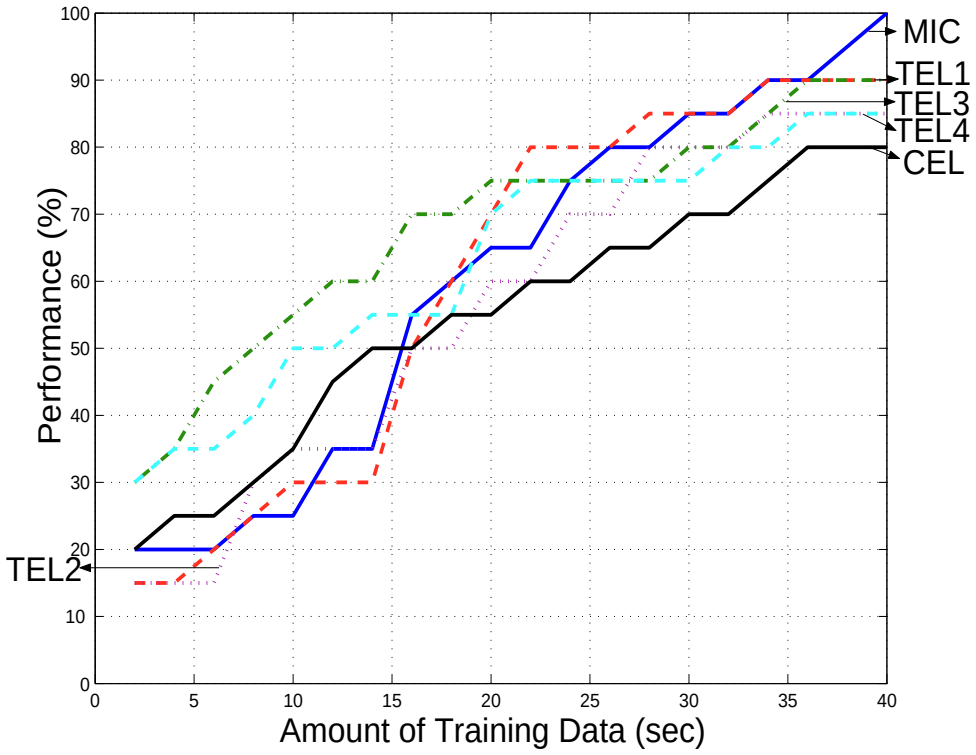

Fig. 5.3: Performance of proposed speaker recognition system based on system features with respect to amount of training data. Performance of the system indicates number of speakers identified correctly out of 20 speakers. MIC is microphone data, TEL1 to TEL4 are telephone data, and CEL is cellular data.

ent test segments from a given test data. It may be possible to combine the multiple evidences for taking a decision.

5.2 NETWORK STRUCTURE FOR SPEAKER RECOGNITION

In the present approach, AANN model is used for capturing the higher order relations that may be present among the samples of the given residual signal. The role played by the number of units in the compression and expansion layers on the performance of the system is examined. The optimization of the network structure also depends on

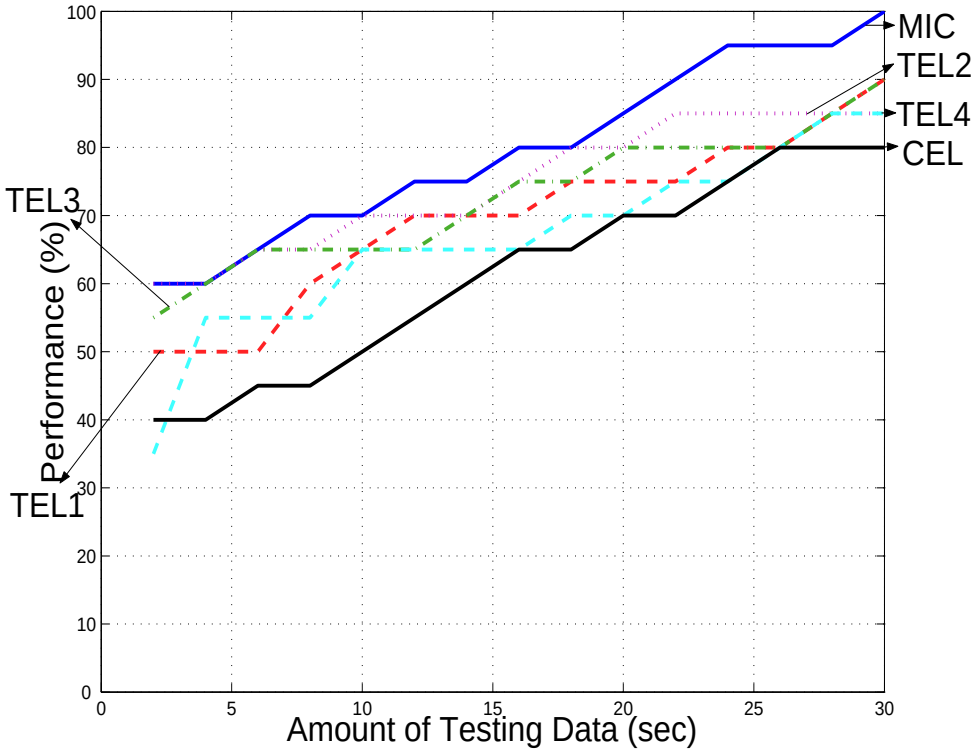

Fig. 5.4: Performance of proposed speaker recognition system based on system features with respect to amount of testing data. Performance of the system indicates number of speakers identified correctly out of 20 speakers. MIC is microphone data, TEL1 to TEL4 are telephone data, and CEL is cellular data.

the type of data used. To analyze the effect of the system on the number of nodes in the compression layer on the performance of the system, the performance is evaluated by varying number of nodes in the compression layer. The results are shown in the Fig. 5.5.

From Fig. 5.5, it is interesting to note that the performance of the system improves even after reducing the number of nodes in the compression layer from 12 to 8 and thereafter the performance is reduced. The optimal number of nodes in compression layer is 8.

To analyze the behavior of the number of nodes in the expansion layer, the per-

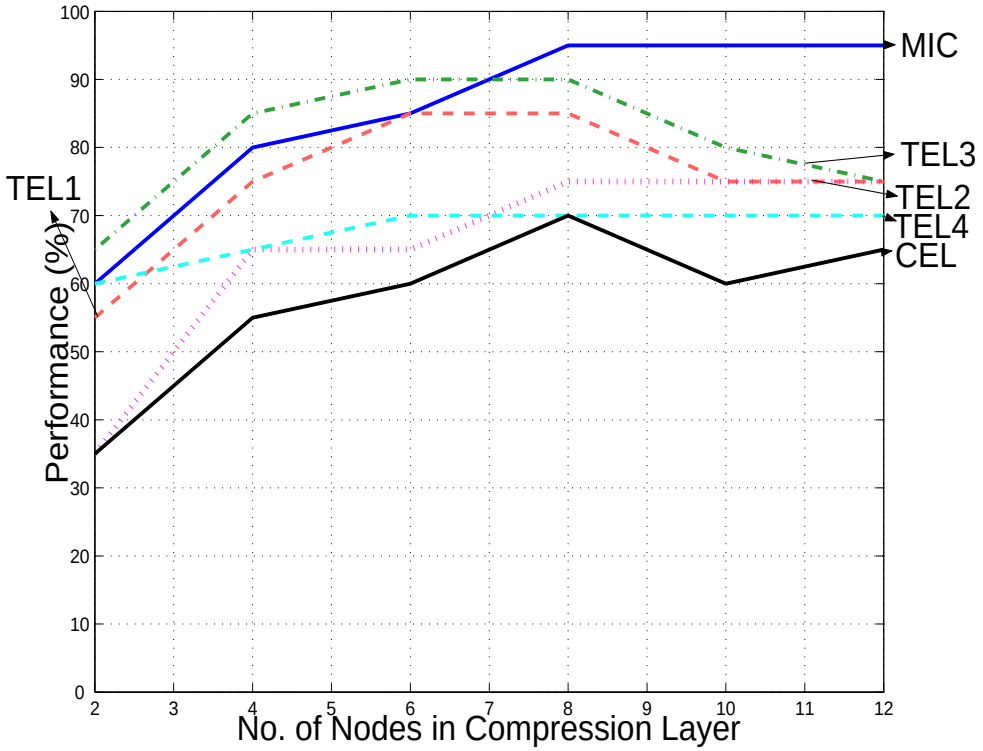

Fig. 5.5: Performance of speaker recognition system for the different number of nodes in compression layer. Performance of the system indicates number of speakers identified correctly out of 20 speakers. MIC is micro phone data, TEL1 to TEL4 are telephone data, and CEL is cellular data.

formance is obtained with varying number of nodes keeping number of nodes in the compression layer to 8. The performance variations is shown in Fig. 5.6. From Fig. 5.6, one can note that the performance is retained even after reducing the number of nodes from 48 to 32, and thereafter the performance deteriorates. From the above two experiments, it can be concluded that the optimal network structure is $40L32N8N32N40L$. By this optimization of the network structure, 30% of computaion time for generating speaker model is reduced as compared to the baseline system. At present, for testing a given claimant model, confidence values of all the blocks are considered. It is possible to evolve a block selection criterion, which may improve the performance. In the next section, a study on block selection criterion is presented.

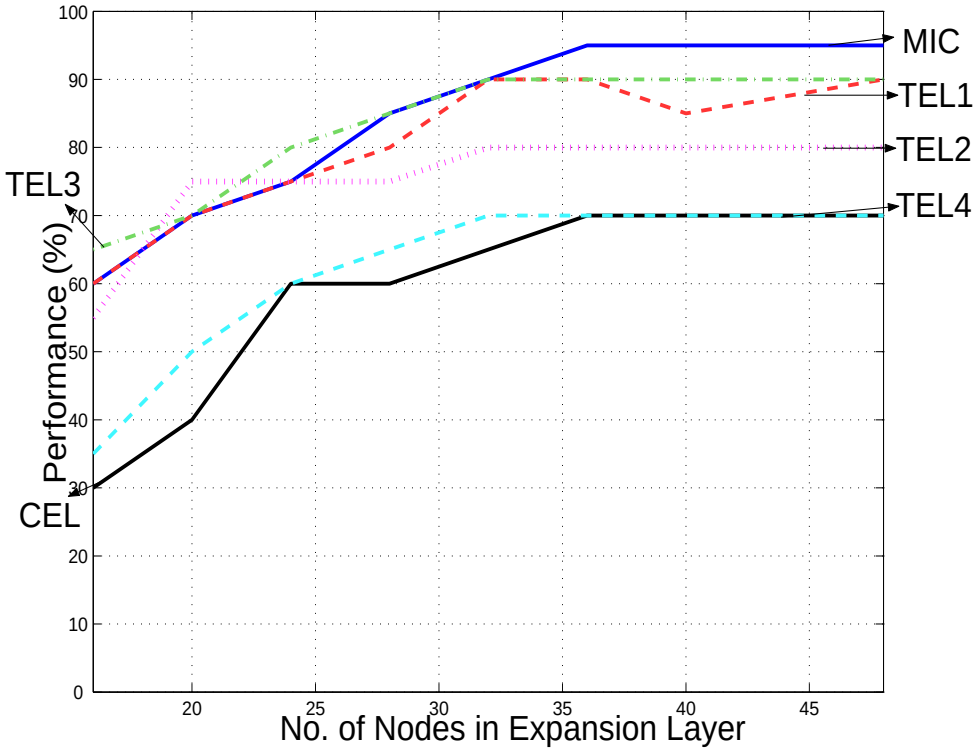

Fig. 5.6: Performance of speaker recognition system for the different number of nodes in expansion layer. Performance of the system indicates number of speakers identified correctly out of 20 speakers. MIC is micro phone data, TEL1 to TEL4 are telephone data, and CEL is cellular data.

5.3 REGION AROUND GLOTTAL CLOSURE INSTANTS FOR SPEAKER RECOGNITION

Speaker information varies from one sound unit to another sound unit. This can be observed in the confidence values for various blocks. To select blocks where speaker-specific source information is better reflected, knowledge of different voice sources is helpful. In the next section, we discuss differences in vocal fold vibrations which make voices different.

5.3.1 Differences in Vocal Fold Vibrations

Videos of vocal fold vibration show large variations in the movement of the vocal folds from one individual to another [75] [93]. For certain speakers, the vocal folds may close completely, while for others, the vocal folds may never reach full closure. The manner and speed with which the vocal folds close also vary across speakers. For example, the vocal folds may close in zipper-like fashion, or may close along the length of the vocal folds at approximately same time. Differences in the vibration of the vocal folds correspond to differences in voices. These voices vary from *soft* voice to *hard* voice depending on closing of vocal folds.

If the vocal folds are held together only loosely, a breathy voice will be produced. There are many variations of breathy voice. It is sometimes made with the vocal folds fairly far apart, so that it sounds like voice produced while sighing. It is as if the vocal folds were flapping in the breeze. At other times the vocal folds are only slightly further apart than in regular voice, producing a kind of murmured sound. The difference between breathy voice and regular voice becomes clearer when one looks at the waveforms for the word, $b^h al$ as shown Fig. 5.7. The Fig. 5.7 (b) is regular voice and (a) is breathy voice. One can see not only the onset of each wave produced by a vocal fold pulse (marked by solid points) but also the prominent peaks within each pulse (marked by open circles) that correspond to the primary resonance of the vocal tract, namely the first formant. The breathy voiced wave in (a) has a slightly lower amplitude and far less well-defined structure within each repetition. The vocal fold pulses are still visible, but the waves corresponding to the formants are not so obvious. This is because the resonances of the vocal tract are excited to a lesser extent by breathy voice. To get well-defined formants, which can be seen as subsidiary waves within each major peak in (b), sharp pulses from the vocal folds are needed. Breathy-voiced waveforms often show little more than the fundamental frequency with a few extra variations superimposed in air pressure.

The vocal folds vibrate more stiff when they are held somewhat tightly together,

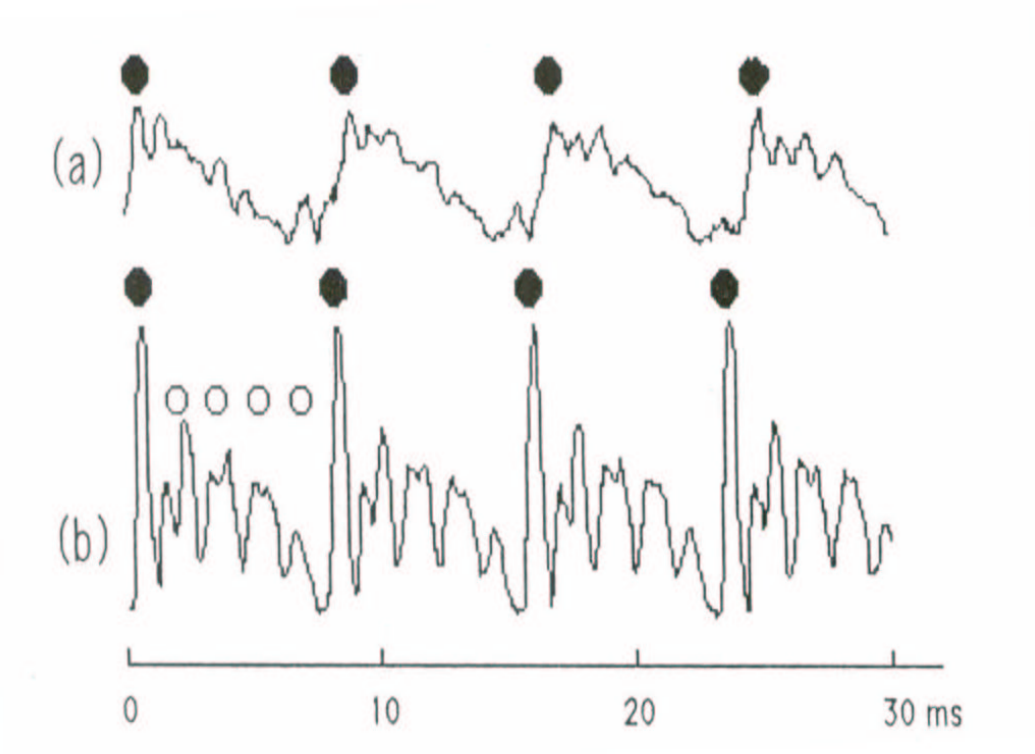

Fig. 5.7: (a) Breathy voice (b) Regular voice.

producing a creaky voice. The difference between creaky voice and regular voice becomes clearer when one look at the waveforms shown in Fig. 5.8 more closely. Regular voice in the Fig. 5.8(a) has vocal fold pulses (marked with solid points) and a wave corresponding to the first formant (marked with open circles). Fig. 5.8(b) shows creaky voice. The vocal fold pulses are slightly further apart. Within each pulse the wave corresponding to the first formant has a greater amplitude than in regular voicing.

From the previous discussion, it is evident that voice differences occur in the vibrations of the vocal folds and a wave corresponding to the first formant. These differences are well manifested in LP residual obtained from inverse filtering of speech. Ideally, the output of the inverse filter for voiced speech should consists of impulses separated by pitch periods. However, such an output is seldom observed. The vocal fold excitation event and error in estimating the first formant wave after excitation is manifested in the LP residual. We hypothesize that speaker-specific information in the excitation is

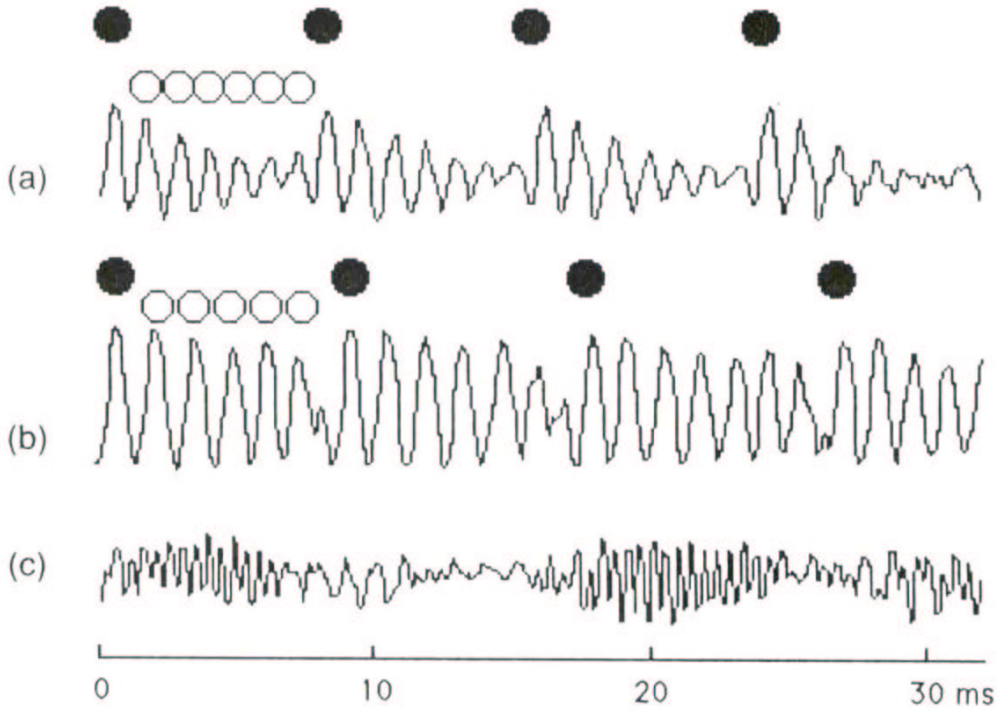

Fig. 5.8: (a) Regular voice (b) Creaky voice and (c) Breathy voice.

centered around the glottal closure instants. This can also be observed in the variation of the confidence value for each block, as shown in Fig. 5.9.

Blocks which contain glottal closure instants have higher confidence value than other blocks. For blocks that contain the instants of glottal closure, impostor models have lower confidence values than for the genuine speaker. Confidence value represents probability that the block belongs to that speaker, and hence the regions around the glottal closure instants have significant speaker-specific information. Performance of speaker recognition can be improved by suitably employing the knowledge of the glottal closure instants. This can be done in two ways. One way is to assign higher weightage to the blocks containing the glottal closure instants over other blocks. The other way is to consider only the blocks containing glottal closure instants for modeling the speaker, and also for testing. The latter one is more advantageous. It requires lesser computation time because of the reduced number of blocks. Hence, algorithms for

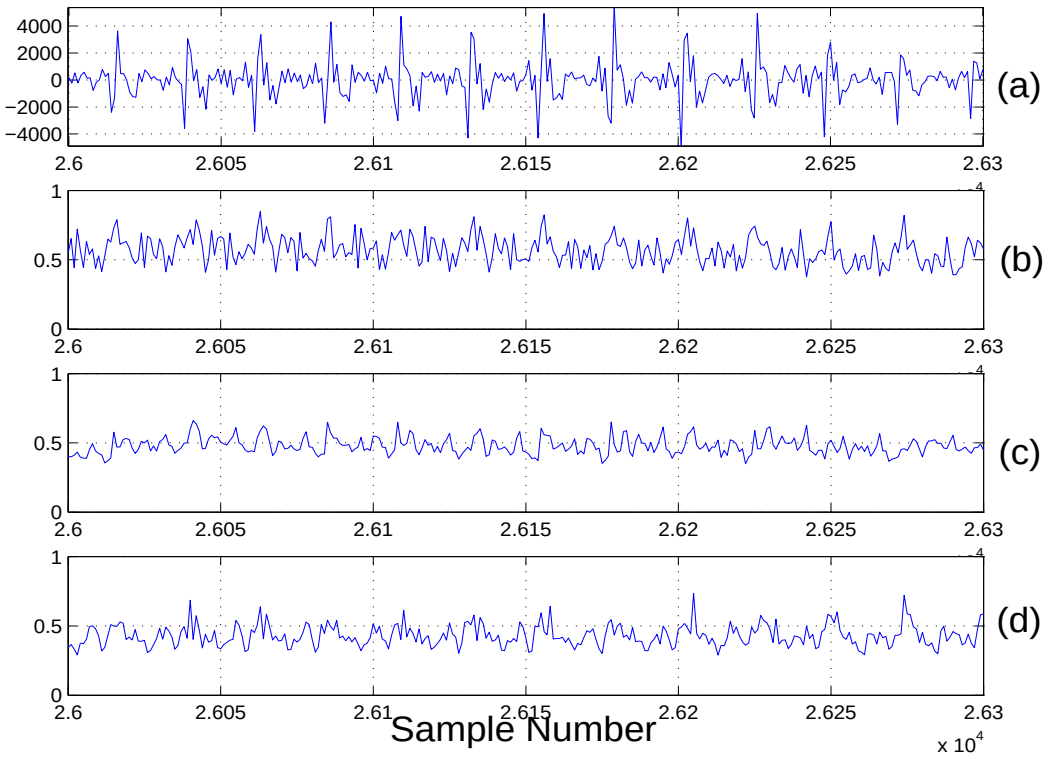

Fig. 5.9: (a) LP residual for voiced speech (b) Block confidences of genuine speaker, (c), and (d) are impostor block confidences.

identifying glottal closure instants are needed.

5.3.2 Algorithm for Identifying Glottal Closure Instants

An algorithm for finding glottal closure instants is proposed. Speech analysis of voiced speech consists of determining the frequency response of the vocal tract system and the glottal pulses representing the voice source. Although the source of excitation for voiced speech is a sequence of glottal pulses, the significant excitation of the vocal tract system can be considered, to a first approximation, to be at discrete instants of time called epochs [65] [94]. There can be more than one epoch within a pitch period but the significant excitation, which coincides with glottal closure, can be treated as an important epoch. Thus, epochs can be effectively used in identifying the glottal closure instants.

A large value in the LP residual is supposed to indicate the region around the epoch location. Speech samples following the estimated epoch or those belonging to an interval with low values of LP residual are assumed to belong to the closed glottis interval. However, direct use of the LP residual for extracting epochal information is not very effective owing to the occurrence of samples of either polarity of large values around the instant of significant excitation. Unambiguous identification of epochs from the LP residual consists of computing the Hilbert envelope of LP residual [94]. The Hilbert envelope $h(n)$ of the residual signal $r(n)$ is obtained as [95]

$$h(n) = \sqrt{r^2(n) + r_h^2(n)} \tag{5.1}$$

where $r_h(n)$ is Hilbert transform of $r(n)$. The Hilbert transform of a signal $r(n)$ is obtained by exchanging the real and imaginary parts of the DFT of $r(n)$, and then computing the IDFT. Peaks in the Hilbert envelope are then identified. Though most of the peaks coincide with glottal closure instants, some spurious peaks do exist. These are eliminated based on the hypothesis that the time gap between two successive glottal closure instants is not likely to vary much in the adjacent pitch periods.

The following steps outline the method used for identifying the glottal closure instants.

1. Down sample the 8 kHz sampled signal by a factor two.
2. Compute the LP residual using 6^{th} order LP analysis, with a frame size of 20 ms and a frame shift of 10 ms.
3. Compute the Hilbert envelope of the LP residual.
4. Identify the peaks in Hilbert envelope.
5. Eliminate spurious peaks by hypothesizing that time gap between two successive glottal closure instants is not likely to vary much in the adjacent pitch periods.
6. Hypothesize the remaining peaks as glottal closure instants.

LP residual is extracted from the down sampled signal. Down sampling improves the signal to noise ratio and reduces the data for compuation. LP order is also reduced

correspondingly to six. Fig. 5.10 shows a sample of the result of the algorithm to detect the instants of glottal closure.

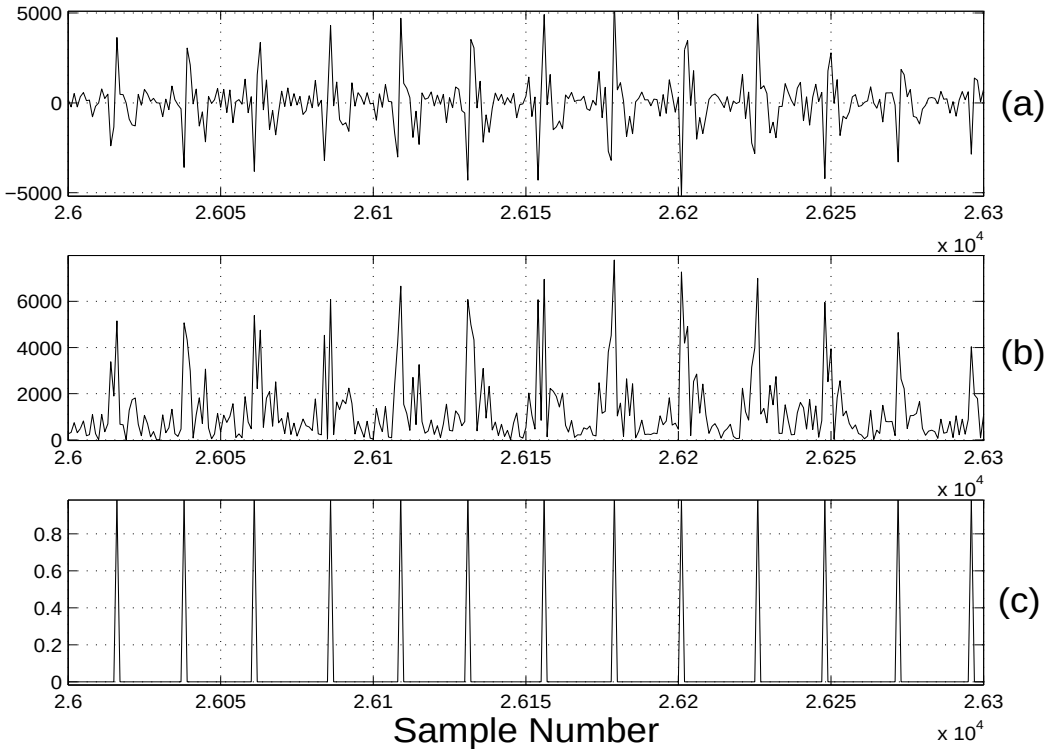

Fig. 5.10: (a) LP residual for voiced speech (b) Hilbert envelope of (a), and (c) glottal closure.

5.3.3 Results

The extent of speaker-specific information present around the glottal closure instants is verified by speaker recognition studies. The LP residual is extracted from the down-sampled signal. Downsampling improves the signal-to-noise ratio of the LP residual. Glottal closures are identified using the above algorithm. Speaker modeling and testing are performed by considering 11 blocks (5 blocks before and after glottal closure) around the glottal closure instant. Confidence value measurement and decision logic are not modified. The network structure is scaled down to $20L16N5N16N20L$, because, each block has 20 samples now, due to downsampling. The performance of the

recognition system before and after the block selection is shown in Table 5.1.

Table 5.1: Performance of speaker recognition system for before and after block selection on various data sets. Performance of the system indicates number of speakers identified correctly out of 20 speakers. The values in the parenthesis are percentage recognition.

Data Set	Performance of system before block selection	Performance of System after block selection
MIC	19 (95%)	19 (95%)
TEL1	18 (90%)	18 (90%)
TEL2	16 (80%)	16 (80%)
TEL3	18 (90%)	18 (90%)
TEL4	14 (70%)	14 (70%)
CEL	13 (65%)	14 (70%)

One can notice that the performance is similar in almost all sets of data, and slightly improved in CEL. Glottal closure regions are high SNR (Signal to Noise Ratio) regions in the LP residual. Considering these high SNR regions in the LP residual enhances the confidence value of the genuine speaker. Cellular data is more noisy compared to other sets. Hence the improvement in the performance of the speaker recognition is more visible in CEL set. By considering only the glottal closure regions, the system based on block selection clearly retains the performance of the previous system. Hence, significant speaker specific information is available at glottal closure instants. Computation time is reduced significantly by downsampling and considering only the region around the glottal closure instants. The computation overhead for finding the instants of glottal closure is relatively small.

5.4 SUMMARY

In this chapter, the effect of various parameters on the performance of speaker recognition system was presented. The study made on the size of data for training and testing showed that about 6 seconds of data was enough to capture speaker variability in terms of source characteristics. The amount of training as well as testing data required in the case of speaker recognition system based on source features is very less compared to the existing systems based on the vocal tract system features. Hence for the same amount of data, one can have multiple models and multiple test segments. Giving multiple test segments to all the models and combining the evidences of these models may improve the performance of the system. The study made on the effect of number of units in the dimension compression layer and expansion layer on the performance showed the robustness of the speaker recognition system to variations in network structure. It provides the advantage of a small network structure. The study on the glottal closure instants for speaker recognition showed the importance of regions around these instants for speaker recognition.

The present system gives equal importance for all sounds of speech. Some specific sounds may have more speaker information than others. The performance can be improved by selecting sound units which contain more speaker-specific information. The speaker information in the LP residual in different sound units is analyzed in the next chapter.

CHAPTER 6

Significance of Sound Units for Speaker Recognition

6.1 INTRODUCTION

All sounds uttered by a speaker do not carry the same amount of information about the identity of the person. Therefore, speaker recognition performance can be improved by selecting sound units useful for speaker recognition. In the context of speech perception we assume that the vowels (or steady voiced regions) carry more information related to the speaker than consonants (especially stop consonants). This can be justified as follows. Consonants are dynamic sounds and their duration is less compared to the duration of the vowels. Hence while perceiving a consonant, the listener will have to pay a lot of attention in comprehending the message (i.e., recognizing the consonant). This causes the listener to ignore the speaker characteristics embedded in a consonant. In the case of vowels, since the duration is relatively large, the perceiving will be easy so that the listener can pay attention to the voice characteristics also. But this argument need not be valid for consonants like laterals and nasals which carry a lot of speaker characteristics. Nasals have been of particular interest because the nasal cavities of different speakers are distinct, and not easily modified. Sachin *et.al.,* have shown that vowels, diphthongs, glides, nasals, fricatives and stops have speaker information in decreasing order [96]. In this chapter speaker-specific information present in the excitation source of different sound units will be examined.

6.2 SIGNIFICANCE OF DIFFERENT EXCITATION SOURCES

The sources of excitation for speech production are plosive, fricative and glottal vibration. Plosive excitation is due to total closure and sudden release at some point along the vocal tract system, and it results in the production of stop consonants. Fricative excitation is due to narrow constriction somewhere along the length of vocal tract system, which results in the production of fricative sounds. Glottal vibration produces voiced sounds like vowels, nasals and semivowels. Glottal vibration is the major excitation of speech, as more than 70% of the speech is voiced. Moreover, if voicing is replaced by random noise excitation to produce whispered type of speech, one notices that most of the speaker's identity is lost. Thus it appears that significant speaker-specific information may be present in the nature of vibration of the vocal folds. Among the voiced sounds, speaker information may be significant in the case of vowels. Hence we consider the source information for the case of five vowels /*a*/, /*i*/, /*u*/, /*e*/ and /*o*/ in this study.

The excitation source characteristics are different for different speakers. This is illustrated in Fig. 6.1, where the LP residuals for segments of vowel /*a*/ are shown for five different speakers. As shown in the figure, the rate of vibration of the vocal folds and the strength of excitation are different for different speakers.

The five vowels considered in the present study may be grouped into three categories depending on the position of tongue hump as, front vowels (/*i*/), mid vowels (/*a*/ and /*e*/) and back vowels (/*u*/ and /*o*/). The vowels are also classified depending on the lip rounding as rounded (/*u*/ and /*o*/) and unrounded (/*a*/, /*i*/ and /*e*/). Even though the source of excitation is glottal vibration in all the cases, the characteristics of the excitation source will be different for different vowels due to the position of the tongue hump and lips. This can be seen in Fig. 6.2, where segments of the LP residuals for the five vowels are given for a speaker. As can be seen in the figure, the excitation in the case of vowels /*u*/ and /*o*/ are not as sharp as for the other vowels. Perceptually also speaker characteristics seem to be manifested well for unrounded vowels

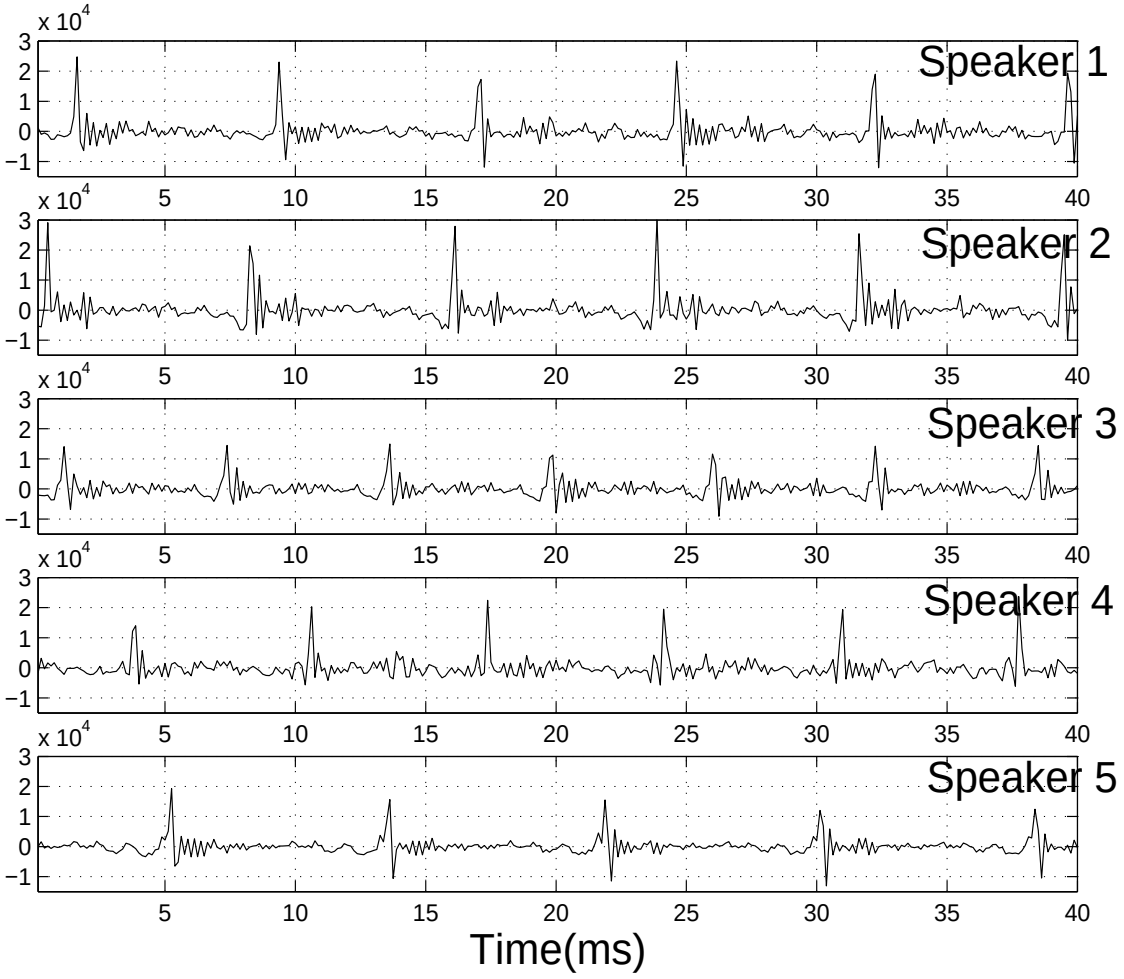

Fig. 6.1: LP residuals for vowel /a/ for five different speakers.

compared to rounded vowels. Thus the extent of speaker information manifested in the excitation source may be different for different vowels. This is also confirmed by the experimental studies to be discussed in the next section.

6.3 EXPERIMENTAL STUDY ON THE SIGNIFICANCE OF DIFFERENT EXCITATION SOURCES

To study the effectiveness of the speaker-specific source information for each of the vowels, we conducted recognition experiments separately for each vowel. The data for

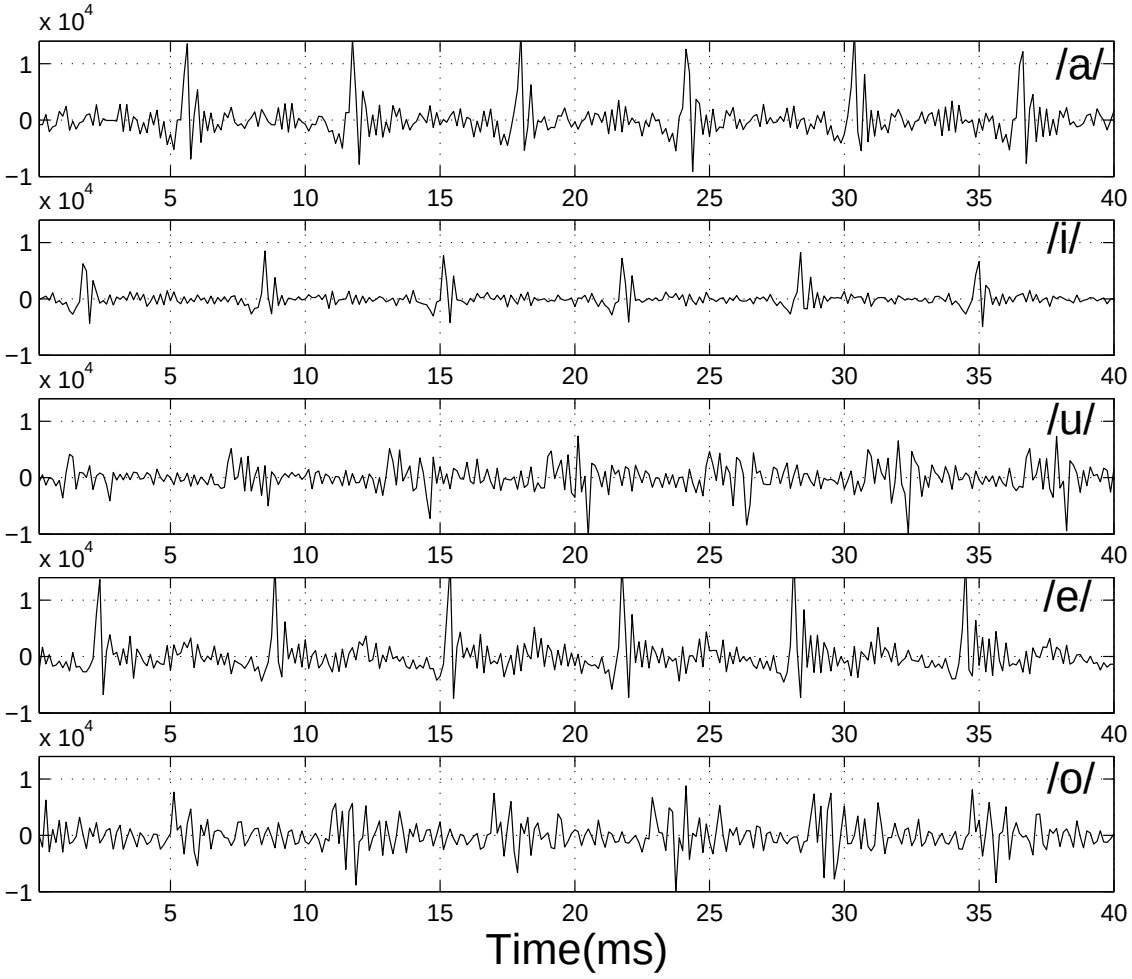

Fig. 6.2: LP residuals for five different vowels of the same speaker.

the recognition experiments is collected from 20 cooperative speakers. For building speaker models, we collected vowels of duration 1-3 sec. The speech signal is collected by a microphone in the laboratory environment. The signal is sampled at 8 kHz, and is stored as 16 bit integers. LP residual is extracted from the speech signal using a 12^{th} order LP analysis, and the residual is normalized to unit magnitude before feeding it to the AANN models. Residual samples are given in blocks of 40 samples with one sample shift. The speaker models are trained for 60 epochs using backpropagation learning algorithm [29]. The training error curves for all the five vowels of a speaker are given in Fig. 6.3. The low training error values for vowels $/a/$, $/i/$ and $/e/$ shows

that speaker-specific information may be represented better in the case of unrounded vowels. Higher training error values for vowels /u/ and /o/ may be attributed to poor representation of speaker-specific information. One model is built for each vowel for each speaker.

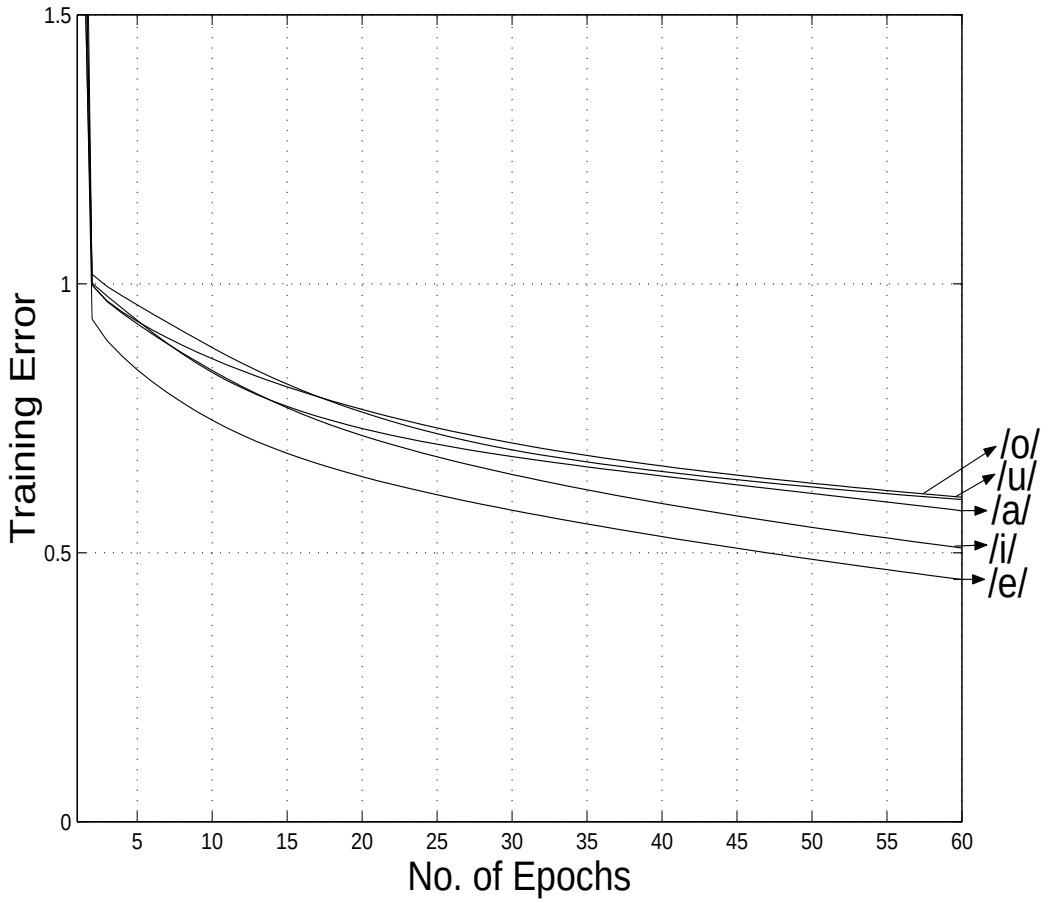

Fig. 6.3: Training error curves for the five vowels /a/, /i/, /u/, /e/, and /o/ for a speaker.

During verification, a test utterance of typically 0.5 sec duration is used. The LP residual is computed using a 12^{th} order LP analysis, and is normalized to unit magnitude for each block of 40 samples. The blocks are presented with one sample shift to all the models. The output of each model is compared with its input to compute

the squared error for each block. The error (E_i) for the i^{th} block is transformed into a confidence value using $C_i = exp(-\lambda E_i)$, where the constant $\lambda = 1$. The block confidences for a segment of all the vowels, for both the genuine as well as an impostor speaker are shown in Figs. 6.4 to 6.8. As shown in the figures, the confidence values for genuine speakers in the case of vowels /a/, /i/ and /e/ have high discrimination compared to the confidence values of the impostors. The average confidence value for the genuine speaker is around 0.6, whereas that for the impostor speaker it is around 0.5. For /u/ and /o/, the discrimination between the confidence values of genuine and impostor speakers is very low. As shown in the figure, the average confidences for both the cases are around 0.5. The average of all the block confidences for a given test utterance is given by $C = (1/N) \sum_{i=1}^{N} C_i$, where N is number of blocks in the test utterance. The average confidence value is used to decide how best the test utterance matches the given model.

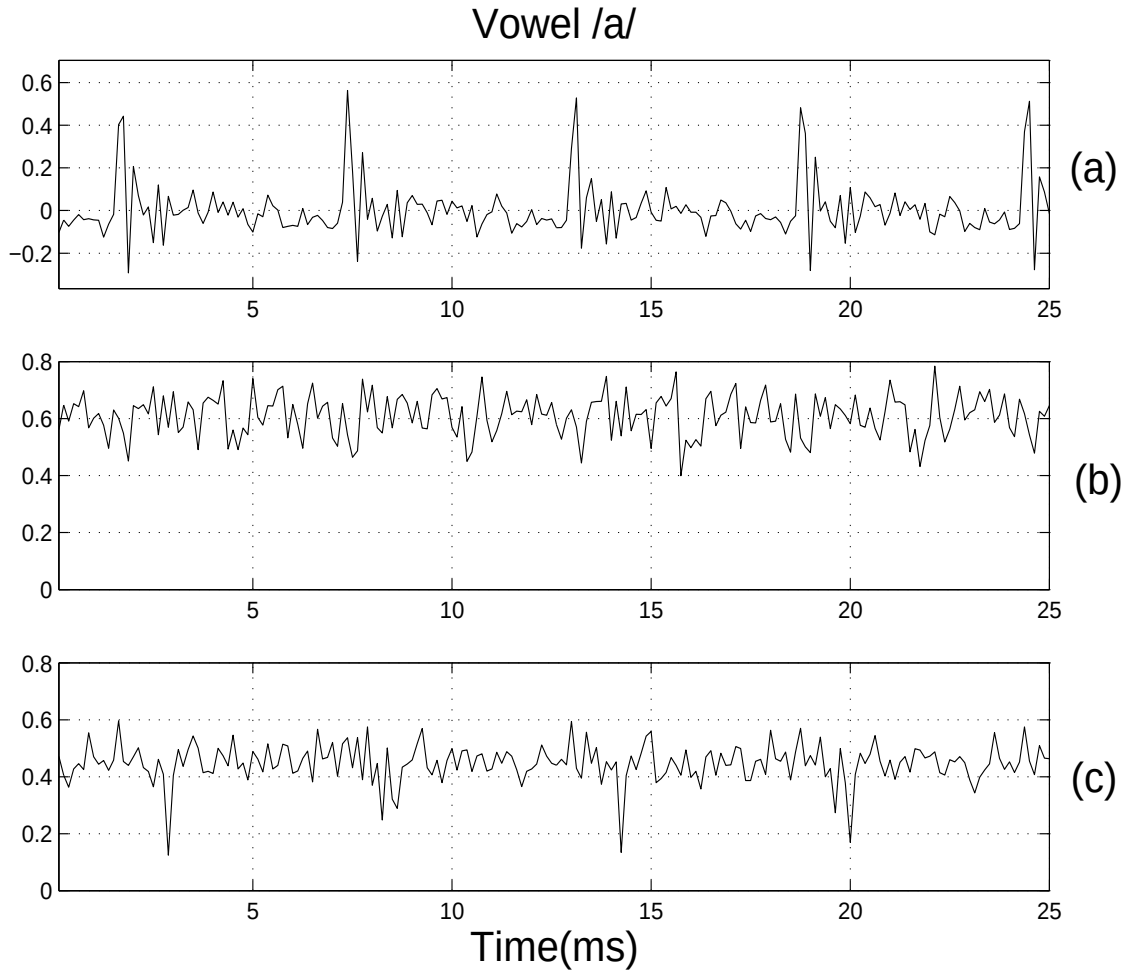

Fig. 6.4: For a segment of vowel $/a/$, (a) Normalized LP residual, (b) Block confidences for genuine speaker, and (c) Block confidences for an impostor speaker.

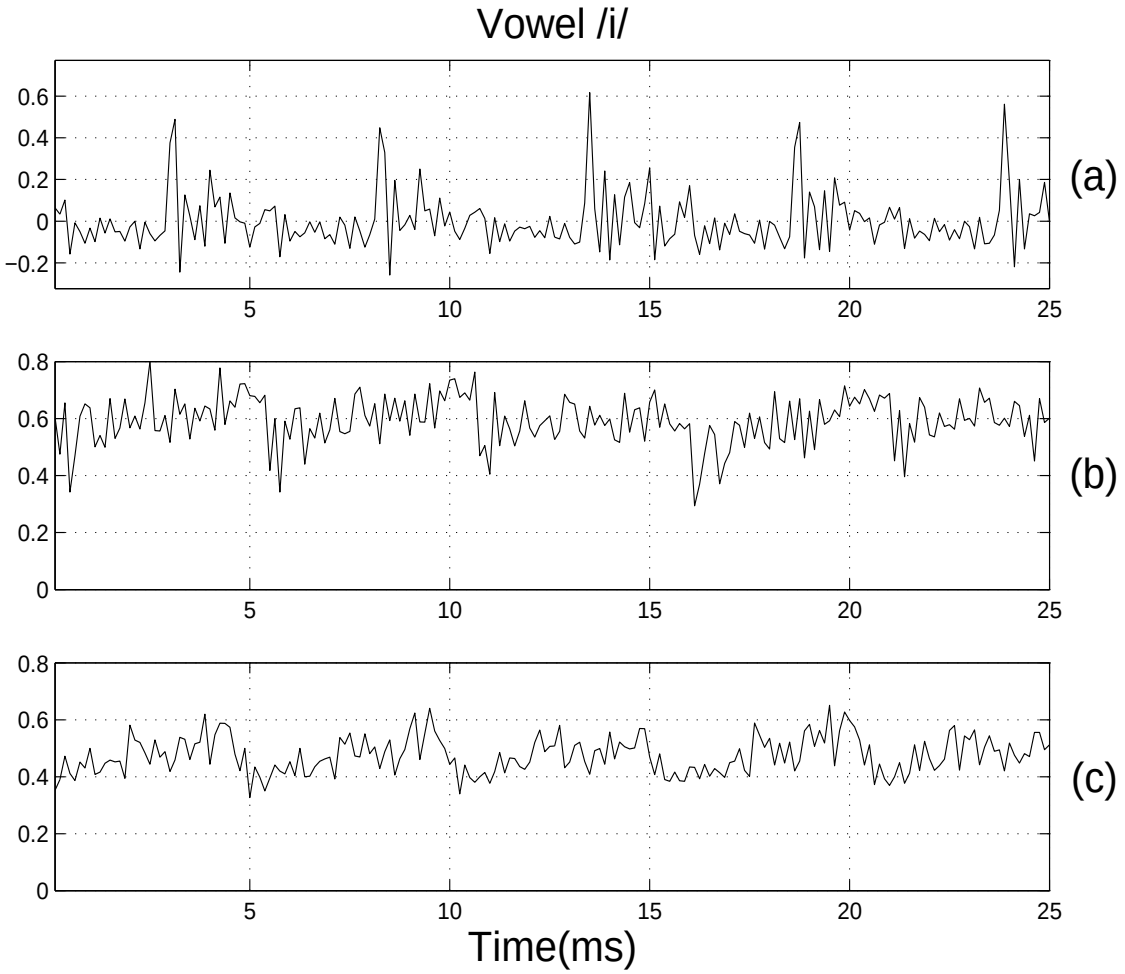

Fig. 6.5: For a segment of vowel $/i/$, (a) Normalized LP residual, (b) Block confidences for genuine speaker, and (c) Block confidences for an impostor speaker.

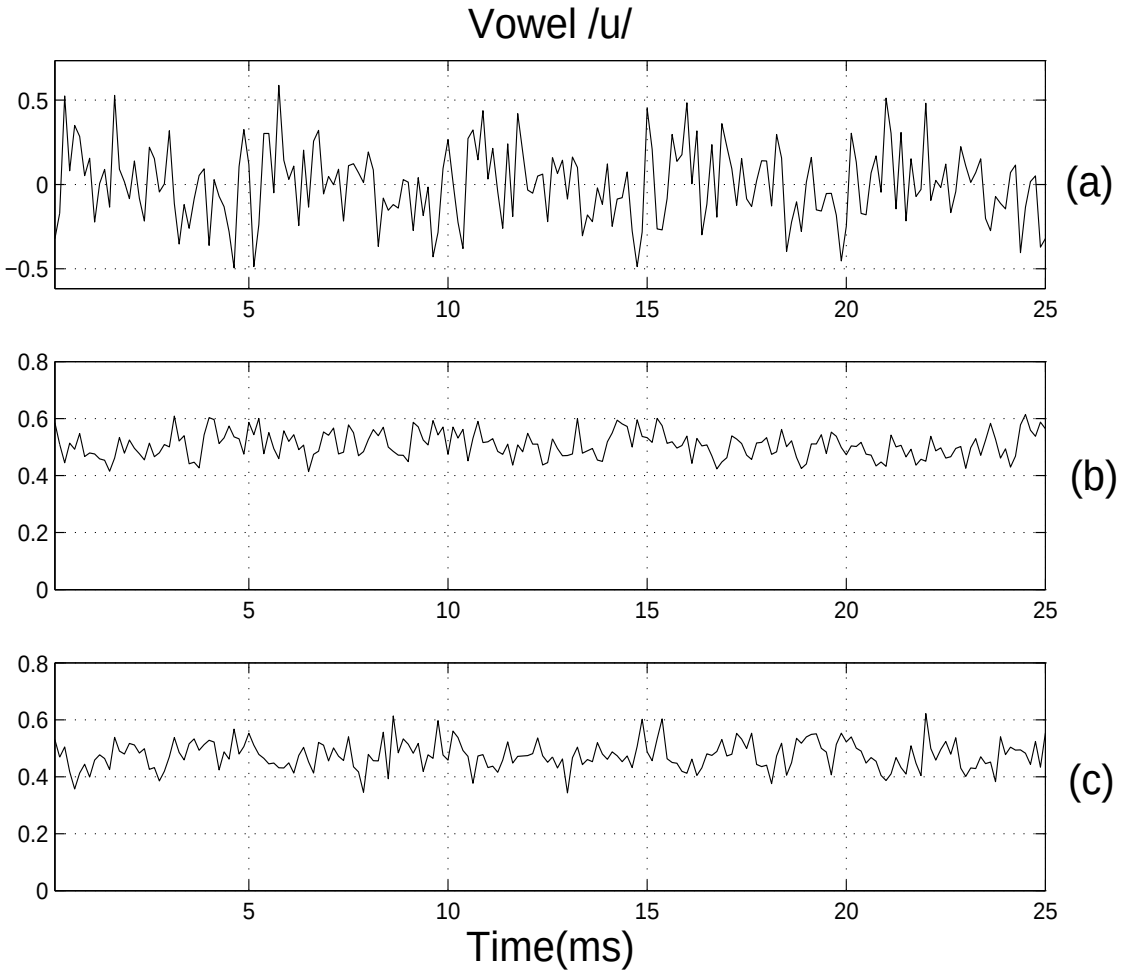

Fig. 6.6: For a segment of vowel $/u/$, (a) Normalized LP residual, (b) Block confidences for genuine speaker, and (c) Block confidences for an impostor speaker.

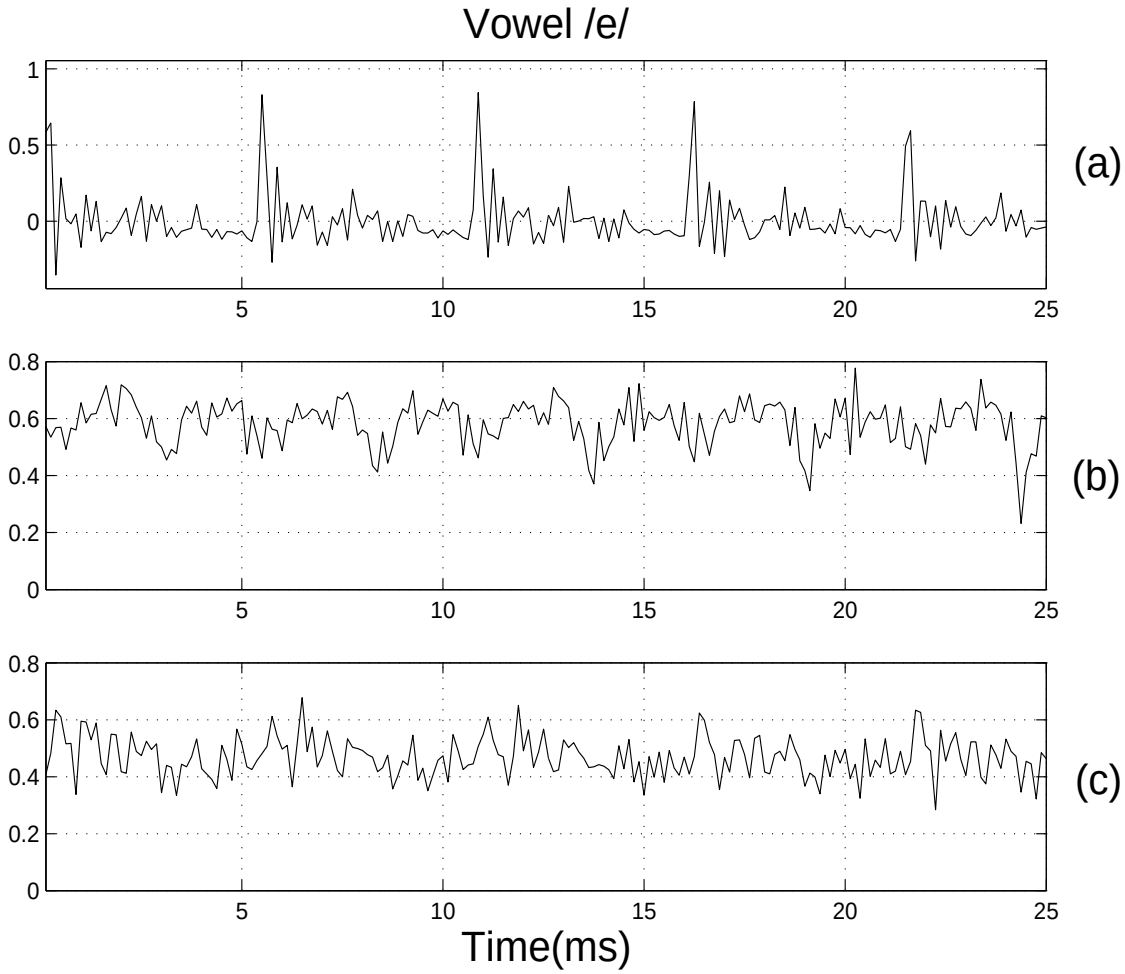

Fig. 6.7: For a segment of vowel $/e/$, (a) Normalized LP residual, (b) Block confidences for genuine speaker, and (c) Block confidences for an impostor speaker.

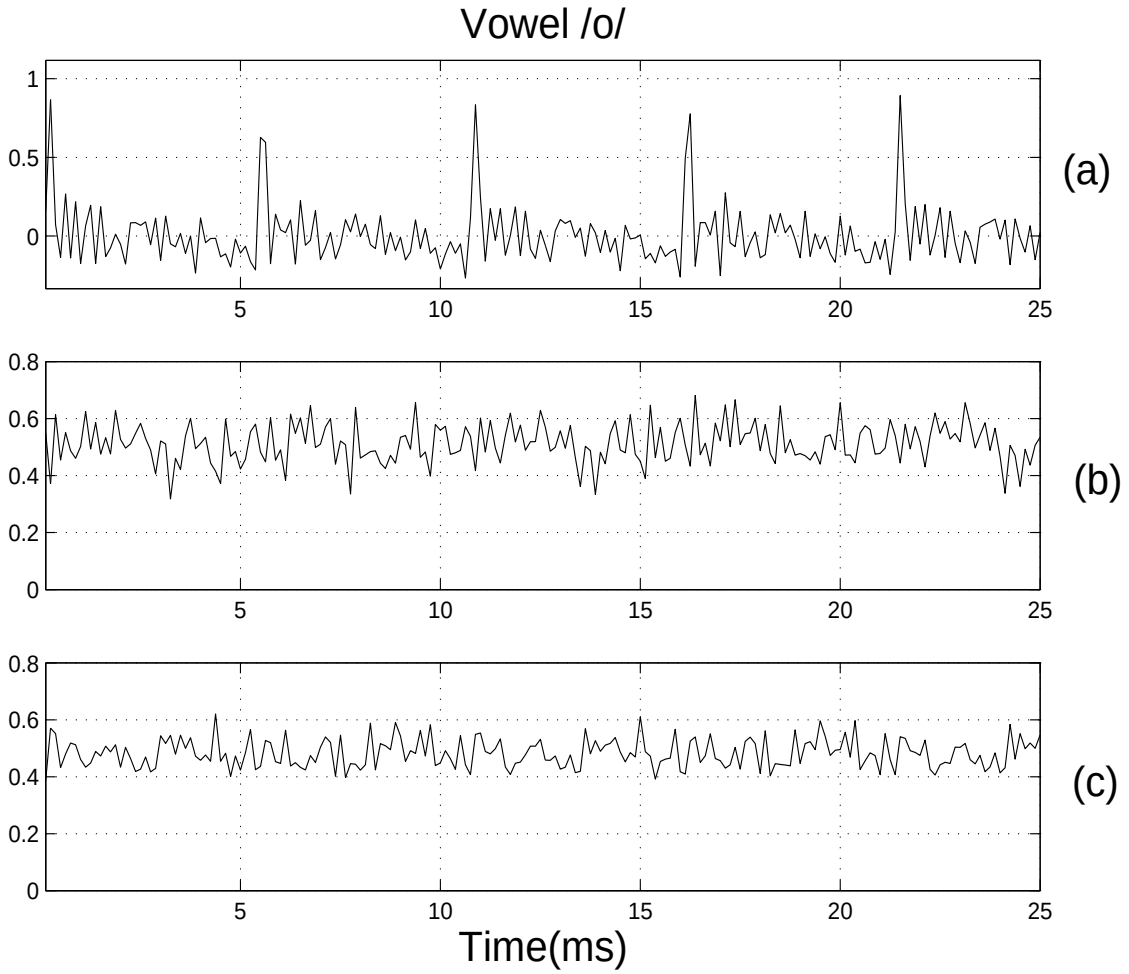

Fig. 6.8: For a segment of vowel /o/, (a) Normalized LP residual, (b) Block confidences for genuine speaker, and (c) Block confidences for an impostor speaker.

Person authentication is a typical two-class problem: either the claimant is the genuine speaker, or is an impostor. When dealing with this type of two-class problem, the decision module can make two kinds of errors. These two errors are:
False Rejection (FR): when a genuine claimant is rejected;
False Acceptance (FA): when an impostor is accepted.
The performances of the speaker verification system are usually given in terms of global error rates computed during tests: False Rejection Rate (FRR) and False Acceptance Rate(FAR). These error rates are defined as follows:

$$FRR = \frac{number \quad of \quad False \quad Rejections}{number \quad of \quad genuine \quad trials} \tag{6.1}$$

$$FAR = \frac{number \quad of \quad False \quad Acceptances}{number \quad of \quad impostor \quad trials} \tag{6.2}$$

A perfect identity verification (FAR=0 and FRR=0) is in practice unachievable. However, any of the two (FAR and FRR) can be reduced to an arbitrary small value by changing the decision threshold, with the drawback of increasing the other one.

To evaluate the recognition performance, 50 genuine trials and 50 impostor trials for each of the vowels are conducted. The performance is expressed in terms of FAR and FRR, and both are expressed in percentage. The results of the testing for all the vowels are given in Table 6.1. The high false rejection in the case of vowels /u/ and /o/ indicates the poor presence of speaker-specific information in these vowels.

6.4 SUMMARY

The speaker-specific source information present in different sound units was explored. The observation is that the excitation source characteristics in vowel sound units reveal sharpness differences in different vowels. The excitation in rounded vowels is not as sharp as unrounded vowels. The experimental studies also confirm that the speaker characteristics are better reflected in unrounded vowels compared to rounded vowels.

The amount of training as well as testing data required in the case of source-feature-based speaker recognition system is very less compared to the existing systems

Table 6.1: Performance of speaker verification system using each of the five vowels. False acceptance and false rejection are expressed in percentage out of total 50 trials conducted for each case.

Vowel	FAR in %	FRR in %
/a/	2	40
/i/	4	36
/u/	0	62
/e/	2	18
/o/	4	60

based on the vocal tract system features. Hence for the same amount of data, multiple models and multiple test segments can be made, providing multiple evidences for decision about the speaker, which may improve the performance of the system. It is however possible to divide the speech sounds into categories and perform the modeling independently for each of them. Ideally, this should lead to better modeling of the speaker and hence better performance of the system. Furthermore performance may be improved by combining evidences from several categories. To get good results categorization of sound units is needed.

CHAPTER 7

Summary and Conclusions

The objective of this work is to illustrate the significance of source features for text-independent speaker recognition task.

Speech signal carries with it both message and speaker information. Speech is used to convey the message through a sequence of sound units, which are produced by exciting the time varying vocal tract system with time varying excitation. Each sound unit is produced by a specific combination of excitation and vocal tract dynamics. For representation of speech message information, the vocal tract system is modeled as a time varying filter, and the excitation as voiced or unvoiced or plosive or combination of these types. The time varying filter characteristics capture the variations in the shape of the vocal tract system in the form of resonances, antiresonances and spectral roll-off characteristics. This representation of speech has been very effective for developing speech recognition systems. Since the vocal tract shape and its dynamics are also unique for a given speaker, the same time varying filter representation has been exploited for developing speaker recognition systems as well.

Speech and speaker characteristics are also present in the suprasegmental features of a speech signal such as duration and intonation. These suprasegmental features are higher level production features, and are difficult to characterize. Moreover, these features vary significantly depending on the manner in which the speech is uttered by a given speaker. Therefore it is difficult to extract and represent the duration and intonation knowledge present in a speech signal for applications in speech and speaker recognition systems.

There is yet another component in speech, which is largely ignored in most speech analysis techniques. It is the residual of the speech signal obtained after the vocal

tract characteristics are suppressed from the signal. No specific attempt has been made earlier in exploring the speaker information present only in the residual, which mostly contains the excitation source information. When one listens to the LP residual, one can clearly make out the speaker characteristics present both at the segmental (10-30 ms) level and at the suprasegmental level (1-3 sec). In this work, an approach for extracting speaker-specific information present in the LP residual by using AANN models was proposed. Short segments of the linear prediction residual can be considered to belong to one of the four broad categories, namely, silence, unvoiced, plosive and voiced. The voiced category is dominant, as more than 70% of speech is voiced. Moreover, if voicing is replaced by random noise excitation to produce whispered type of speech, one notices that most of the speaker's information is lost. Thus it appears that significant speaker-specific information may be present in the segmental and suprasegmental features of the residual speech. Suprasegmental information is subjected to large variations by the speaker's manner of production of speech.

In this work the speaker characteristics present in the short segments of the LP residual was explored. It was shown that the residual indeed contains speaker-specific information. Speaker recognition studies using system and source features for different LP orders show almost similar variation in performance. For low LP orders (< 8) the performance of recognition is low. For LP orders in the range 8-20, the recognition systems give good performance. For higher orders, the performance of recognition is low. The performance variation of both the systems followed the characterization of the vocal tract system. When the vocal tract system is characterized well, the recognition performance using the residual source features is also good. This demonstrates the significance of speaker-specific source information present in the LP residual. Note that, since the block is less than a pitch period, only the glottal pulse characteristics presented in the LP residual is used.

The effect of various parameters on the performance of speaker recognition system was presented. The study made on the size of data for training and testing showed that about 6 seconds of data was enough to capture speaker variability in terms of source

characteristics. The amount of training as well as testing data required in the case of speaker recognition system based on source features is significantly less compared to the existing systems based on the vocal tract system features. The study made on the effect of number of units in the dimension compression hidden layer and in the expansion layer on the recognition performance showed the robustness of the speaker recognition system based on source features to variations in the network structure. The study on regions around the glottal closure instants for speaker recognition showed the importance of these regions for speaker recognition. Performance has improved and computation time is reduced.

All the sounds of speech are not equally important for speaker recognition. Some specific sounds tend to be more useful. Therefore, speaker recognition performance can be improved by selecting the most useful sounds for speaker recognition. In the context of speech perception we assume that the vowels (or steady voiced regions) carry more information related to the speaker than consonants (especially stop consonants). Speaker-specific source information present in different sound units (five vowels /*a*/, /*i*/, /*u*/, /*e*/ and /*o*/) was explored. The vowels are classified depending on the lip rounding as rounded (/*u*/ and /*o*/) or unrounded (/*a*/, /*i*/ and /*e*/). Even though the source of excitation is glottal vibration in all the cases, the characteristics of the excitation source will be different for different vowels due to the position of the tongue hump and lips. The observation of segments of LP residuals for the five vowels for a given speaker, shows that the excitation in the case of vowels /*u*/ and /*o*/ is not as sharp as for the other vowels. Perceptually also speaker characteristics seem to be manifested well for unrounded vowels compared to rounded vowels. Thus the extent of speaker-specific information manifested in the excitation source may be high in the unrounded vowels. This was also confirmed by the experimental studies.

7.1 MAJOR CONTRIBUTIONS OF THE WORK

Significance of source features for text-independent speaker recognition has been illustrated.

The optimal LP order range for speaker recognition system based on source features is 8-20.

The effect of various parameters on the performance of speaker recognition system was presented. The study made on the size of data for training and testing showed that small size of data is adequate compared to the systems based on the vocal tract system features. The study made on the network structure showed robustness of the speaker recognition system based on source features to the variations in the network structure.

The importance of glottal closure information for speaker recognition has been illustrated.

The significance of different sound units for speaker recognition was illustrated for vowels. The studies showed the speaker-specific source information is present more in the unrounded vowels compared to the rounded vowels.

7.2 SCOPE FOR FUTURE WORK

Theoretical analysis is required to explain the speaker recognition characteristics in the LP residual and to clarify the higher order relation capturing capability of the AANN.

The amount of training as well as testing data required in the case of speaker recognition system based on source features is very less compared to the existing systems based on vocal tract system features. Hence for the same amount of data, we can have multiple models and multiple test segments. Giving multiple test segments to all the models, and combining the evidences of these models may improve the performance of the system. It is however possible to divide the speech sounds into categories and perform the modeling independently for each of them. Ideally, this

should lead to better modeling of the speaker and hence better performance of the system. Furthermore performance may be improved by combining evidences from several categories.

The proposed speaker recognition system uses gradient descent algorithm with momentum update for modeling of the speaker. Computation time can be reduced by using conjugate gradient algorithms. Using gradient descent method, it needs more number of epochs for network to converge in speaker modeling. Hence computation time is high. By conjugate gradient method network can converge from ten to twenty times faster than the gradient descent method. Hence conjugate gradient method may reduce the computation time. This will facilitate the development of speaker recognition system in real time applications. Conjugate gradient method may help for better speaker modeling. Hence the performance of the proposed speaker recognition may also improve.

The proposed speaker recognition system uses only speaker-specific information present in the short segments of the LP residual. Speaker-specific source information is also present in the suprasegmental level such as intonation and duration. Combining speaker-specific information in the short segments of the LP residual with suprasegmental information, pitch and jitter may improve the performance of the speaker recognition system.

Usefulness of speaker-specific source information in the LP residual can be explored for speaker segmentation task also. Speaker segmentation task is the recognition of the sequence of speakers engaged in conversation. In other words, the aim is to know who speaks and when. This task can be done by locating the points of speaker change and identifying the speaker in each segment. Speaker-specific source information in the LP residual can be useful in locating the points of speaker change as well as identifying the speaker in each segment.

BIBLIOGRAPHY

[1] D. Lancker, J. Kreiman, and K. Emmorey, "Familiar voice recognition: Patterns and parameters - recognition of backward voices," *Phonetics*, vol. 13, no. 1, pp. 19–38, 1985.

[2] A. Reich and J. Duke, "Effects of selected vocal disguises upon speaker identification by listening," *J. Acoust. Soc. Amer.*, vol. 66, no. 4, pp. 1023–1028, 1979.

[3] M. Sigmund, *Speaker recognition identifying people by their voices.* Habilitation thesis, Brno University of Technology, Institue of radio electronics, 2000.

[4] G. R. Doddington, "Speaker recognition–identifying people by their voices," *Proc. IEEE*, vol. 73, pp. 1651–1664, Nov. 1985.

[5] J. D. Markel, B. T. Oshika, and A. H. Gray, "Long-term feature averaging for speaker recognition," *IEEE Trans. Acoust. Speech, Signal Processing*, vol. 25, pp. 330–337, Aug. 1977.

[6] W. A. Hargreaves and J. A. Starkweather, "Recognition of speaker identity," *Language and Speech*, vol. 6, pp. 63–67, 1963.

[7] K. P. Li and G. W. Hughes, "Talker differences as they appear in correlation matrices of continuous speech spectra," *J. Acoust. Soc. Amer.*, vol. 55, no. 4, pp. 833–837, 1974.

[8] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Royal Statist. Soc. Ser. B. (methodological)*, vol. 39, pp. 1–38, 1977.

[9] R. A. Redner and H. F. Walker, "Mixture densities, maximum likelihood and the EM algorithm," *SIAM Review*, vol. 26, pp. 195–239, 1984.

[10] D. A. Reynolds, "Comparison of background normalization methods for text–independent speaker verification," in *Proc. EUROSPEECH*, (Greece), pp. 963–966, 1997.

[11] Sarel van Vuuren, *Speaker Recognition in a Time-Frequency Space.* PhD dissertation, Orgeon Graduate Institute of Science and Technology, Department of Electrical and Computer Engg., Portland, Mar. 1999.

[12] Yegnanarayana *et.al.*, "IITM speaker recognition system," in *Proc. NIST Speaker Recognition Workshop*, Jun. 2000.

[13] M. Forsyth and M. Jack, "Discriminating semi-continuous HMM for speaker verification," in *Proc. IEEE Int. Conf. Acoust. Speech, Signal Processing*, vol. 1, pp. 313–316, 1994.

[14] M. Forsyth, "Discriminating observation probability (DOP) HMM for speaker verification," *Speech Communication*, vol. 17, pp. 117–129, 1995.

[15] J. D. Veth, G. Gallopyn, and H. Bourlard, "Limited parameter HMMs for connected digit speaker verification over telephone channels," *IEEE Trans. Acoust. Speech, Signal Processing*, pp. 247–250, 1993.

[16] Y. C. Zhang and B. Z. Yuan, "Text-dependent speech identification using corcular HMMs," in *Proc. IEEE Int. Conf. Acoust. Speech, Signal Processing*, pp. 580–582, 1988.

[17] J. Naik, "Speaker verification over long distance telephone lines," in *Proc. IEEE Int. Conf. Acoust. Speech, Signal Processing*, pp. 524–527, 1989.

[18] L. R. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition.* Prentice-Hall, 1993.

[19] S. Furui, "An overview of speaker recognition technology," in *Automatic Speech and Speaker Recognition* (C.-H. Lee, F. K. Soong, and K. K. Paliwal, eds.), ch. 2, pp. 31–56, Boston: Kluwer Academic, 1996.

[20] K. P. Li and E. H. K. Jr., "An approach to text-independent speaker recognition with short utterances," in *Proc. IEEE Int. Conf. Acoust. Speech, Signal Processing*, pp. 555–558, 1983.

[21] K. Shikano, "Text-independent speaker recognition experiments using codebooks in vector quantization," *J. Acoust. Soc. Amer.*, vol. 77, p. S11 (A), 1985.

[22] A. E. Rosenberg and F. K. Soong, "Evaluation of a vector quantization talker recognition system in a text independent and text dependent modes," in *Proc. IEEE Int. Conf. Acoust. Speech, Signal Processing*, pp. 873–876, 1986.

[23] F. K. Soong, A. E. Rosenberg, L. R. Rabiner, and B. H. Juang, "A vector quantization approach to speaker recognition," in *Proc. IEEE Int. Conf. Acoust. Speech, Signal Processing*, pp. 387–390, 1985.

[24] I. Booth, M. Barlow, and B. Watson, "Enhancement to DTW and VQ decision algorithms for speaker recognition," *Speech Communication*, vol. 13, pp. 427–433, 1993.

[25] F. K. Soong and A. E. Rosenberg, "On the use of instantaneous and transitional spectral information in speaker recognition," in *Proc. IEEE Int. Conf. Acoust. Speech, Signal Processing*, pp. 877–890, 1986.

[26] T. Matsui and S. Furui, "Speaker clustering for speech recognition using the parameters characterizing vocal–tract dimensions," in *Proc. Int. Conf. Spoken Language Processing*, pp. 137–140, 1990.

[27] A. L. Higgins, L. G. Bahler, and J. E. Porter, "Voice identification using nearest-neighbour distance measure," in *Proc. IEEE Int. Conf. Acoust. Speech, Signal Processing*, vol. 2, pp. 375–378, 1993.

[28] R. P. Lippmann, "An introduction to computing with neural nets," *IEEE Trans. Acoust. Speech, Signal Processing*, vol. 4, pp. 4–22, Apr. 1989.

[29] B. Yegnanarayana, *Artificial neural networks.* New Delhi: Prentice-Hall of India, 1999.

[30] Simon Haykin, *Neural networks: A comprehensive foundation.* New Jersey: Prentice-Hall Inc., 1999.

[31] Y. Bennani and P. Gallinari, "Neural networks for discrimination and modelization of speakers," *Speech Communication*, vol. 17, pp. 159–175, 1995.

[32] Y. Bennani, "Speaker identification through a modular connectionist architecture: Evaluation on the TIMIT database," in *Proc. Int. Conf. Spoken Language Processing*, pp. 607–610, 1992.

[33] J. Oglesby and J. S. Mason, "Optimisation of neural models for speaker identification," in *Proc. IEEE Int. Conf. Acoust. Speech, Signal Processing*, pp. 261–264, 1990.

[34] Y. Gong and J. Haton, "Nonlinear vector interpolation for speaker recognition," in *Proc. IEEE Int. Conf. Acoust. Speech, Signal Processing*, vol. 2, (SanFrancisco, California, USA), pp. 173–176, Mar. 1992.

[35] H. Hermansky and N. Malaynath, "Speaker verification using speaker-specific mapping," in *RLA2C*, (Avigon,France), Apr. 1998.

[36] Hemant Misra, M. Shajith Ikbal, and B. Yegnanarayana, "Spectral mapping as a feature for speaker recognition," in *Proc. Fifth Nat. Conf. Communications*, (IIT, Kharagpur), pp. 151–156, Jan. 1999.

[37] Hemant Misra, *Development of a Mapping Feature for Speaker Recognition.* M.S Thesis, Indian Institute of Technology Madras, Department of Electrical Engg., chennai, May 1999.

[38] M. Gori and F. Scarselli, "Are multilayer perceptrons adequate for pattern recognition and verification," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 20, pp. 1121–1132, Nov. 1998.

[39] J. Oglesby and J. S. Mason, "Radial basis function networks for speaker recognition," in *Proc. IEEE Int. Conf. Acoust. Speech, Signal Processing*, pp. 393–396, 1991.

[40] M. Shajith Ikbal, Hemant Misra, and B. Yegnanarayana, "Analysis of autoassociative mapping neural networks," in *Proc. Int. Joint Conf. Neural Networks*, (Washington, USA), 1999.

[41] B. Yegnanarayana and S. P. Kishore, "AANN-An alternative to GMM for pattern recognition," *Neural Networks*, vol. 15, pp. 459–469, Apr. 2002.

[42] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals.* Prentice-Hall, 1978.

[43] B. Yegnanarayana, "Formant extraction from linear prediction phase," *J. Acoust. Soc. Amer.*, vol. 63, pp. 1638–1640, 1978.

[44] H. A. Murthy, K. V. Madhu Murthy, and B. Yegnanarayana, "Formant extraction from phase using weighted group delay functions," *Electron. Lett.*, vol. 25, pp. 1609–1611, 1989.

[45] S. K. Das and W. S. Mohn, "A scheme for speech processing in automatic speaker verification," *IEEE Trans. Acoust. Speech, Signal Processing*, vol. AU-19, pp. 32–43, 1971.

[46] G. Doddington, "A method of speaker verification," *J. Acoust. Soc. Amer.*, vol. 49, p. 139 (A), 1971.

[47] J. J. Wolf, "Efficient acoustic parameters for speaker recognition," *J. Acoust. Soc. Amer.*, vol. 52, no. 6, pp. 2044–2056, 1972.

[48] J. W. Glenn and N. Kleiner, "Speaker identification based on nasal phonation," *J. Acoust. Soc. Amer.*, vol. 43, no. 2, pp. 368–372, 1968.

[49] L. S. Su, K. P. Li, and K. S. Fu, "Identification of speakers by the use of nasal coarticulation," *J. Acoust. Soc. Amer.*, vol. 56, pp. 1876–1882, 1974.

[50] R. Lummis, "Speaker verification by computer using speech intensity for temporal registration," *IEEE Trans. Acoust. Speech, Signal Processing*, vol. AU-21, no. 2, pp. 80–89, 1973.

[51] J. Makhoul, "Linear prediction: A tutorial review," *Proc. IEEE*, vol. 63, pp. 561–580, Apr. 1975.

[52] B. S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *J. Acoust. Soc. Amer.*, vol. 55, pp. 1304–1312, Jun. 1974.

[53] A. E. Rosenberg and M. Sambur , "New techniques for automatic speaker verification," *IEEE Trans. Acoust. Speech, Signal Processing*, vol. 23, no. 2, pp. 169–175, 1975.

[54] M. R. Sambur, "Speaker recognition using orthogonal linear prediction," *IEEE Trans. Acoust. Speech, Signal Processing*, vol. 24, pp. 283–289, Aug. 1976.

[55] J. Naik and G. R. Doddington, "High performance speaker verification using principal spectral components," in *Proc. IEEE Int. Conf. Acoust. Speech, Signal Processing*, pp. 881–884, 1986.

[56] S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Trans. Acoust. Speech, Signal Processing*, vol. 29, pp. 254–272, Apr. 1981.

[57] J. E. Luck, "Automatic speaker verification using cepstral measurements," *J. Acoust. Soc. Amer.*, vol. 46, no. 4, pp. 1026–1032, 1969.

[58] Yegnanarayana *et.al.*, "IITM speaker recognition system," in *Proc. NIST Speaker Recognition Workshop*, (University of Maryland, MD,USA), Jun. 1999.

[59] R. J. Mammone, X. Zhang, and R. P. Ramachandran, "Robust speaker recognition: A feature–based approach," *IEEE Signal Processing Mag.*, vol. 13, pp. 58–71, Sept. 1996.

[60] M. M. Homayounpour and G. Chollet , "A comparison of some relevant parametric representations for speaker verification," in *ESCA Workshop on speaker Recognition, Identification, and Verification*, pp. 185–188, Apr. 1994.

[61] Gish and M. Schmidt, "Text–independent speaker identification," *IEEE Signal Processing Mag.*, pp. 18–32, Oct. 1994.

[62] G. R. Doddington, M. A. Pryzbocki, A. F. Martin, and D. A. Reynolds, "The NIST speaker recognition evaluation - Overview, methodology, systems, results, perspective," *Speech Communication*, vol. 31, pp. 225–254, Jun. 2000.

[63] M. M. Sondhi, "New methods of pitch extraction," *IEEE Trans. Speech, Audio Processing*, vol. AU-16, pp. 262–266, 1968.

[64] A. M. Noll, "Cepstrum pitch determination," *J. Acoust. Soc. Amer.*, vol. 41, pp. 293–309, 1970.

[65] T. V. Ananthapadmanabha and B. Yegnanarayana, "Epoch extraction of voiced speech," *IEEE Trans. Speech, Audio Processing*, vol. ASSP-27, pp. 562–570, 1975.

[66] W. Hess, *Pitch determination of speech signals, Algorithms and Devices.* Springer-Verlag, 1983.

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

[68] M. K. Sonmez, L. Heck, M. Weintraub, and E. Shriberg, "A log-normal tied mixture model of pitch for prosody-based speaker recognition," *Proc. EUROSPEECH'97*, vol. 3, pp. 1391–1394, Sept. 1997.

[69] K. Sonmez, E. Shriberg, L. Heck, and M. Wintraub, "Modeling dynamic prosodic variation for speaker verification," *Proc. EUROSPEECH'97*, vol. 7, pp. 3189–3192, Sept. 1997.

[70] M. Mathew, *Combining evidences from multiple classifiers for text-dependent speaker verification.* M.S Thesis, Indian Institute of Technology Madras, Department of Computer Science and Engg., Chennai, 1999.

[71] H. M. Dante and V. V. S. Sharma, "Automatic speaker recognition for a large population," *IEEE Trans. Acoust. Speech, Signal Processing*, vol. 27, pp. 255–263, Jun. 1979.

[72] A. S. Madhukumar, *Intonation knowledge for Speech Systems for an Indian Language.* Ph. D thesis, Indian Institute of Technology, Department of Computer Science and Engg., Madras, 1993.

[73] J. M. Zachariah, *Text-Dependent speaker verification using segmental, suprasegmental and source features.* M.S Thesis, Indian Institute of Technology Madras, Department of Computer Science and Engg., Chennai, 2002.

[74] D. G. Childers and C.Ahn, "Modeling the glottal volume-velocity waveform for three voice types," *J. Acoust. Soc. Amer.*, vol. 97, pp. 505–519, Jan. 1995.

[75] M. D. Plumpe, T. F. Quatieri, and D. A. Reynolds, "Modeling of the glottal flow derivative waveform with application to speaker identification," *IEEE Trans. Speech, Audio Processing*, vol. 7, pp. 569–586, Sept. 1999.

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

[77] H. Wakita, "Residual energy of linear prediction applied to vowel and speaker recognition," *IEEE Trans. Acoust. Speech, Signal Processing*, vol. 24, pp. 270–271, 1976.

[78] M. Faundez-zanuy and D. Rodriguez-Porcheron, "Speaker recognition using residual signal of linear and nonlinear prediction models," in *Proc. Int. Conf. Spoken Language Processing*, vol. 2, pp. 121–124, 1998.

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

[80] J. H. Li Lui and G. Palm, "On the use of features from prediction residual signal in speaker recognition," in *Proc. EUROSPEECH*, pp. 313–316, 1997.

[81] B. Yegnanarayana, K. Sharat Reddy, and S. P. Kishore, "Source and system features for speaker recognition using AANN models," in *Proc. IEEE Int. Conf. Acoust. Speech, Signal Processing*, 2001.

[82] C. M. Bishop, *Neural networks for pattern recognition.* New York: Oxford University Press Inc., 1995.

[83] K. I. Diamantaras and S. Y. Kung, *Principal component neural networks, theory and applications.* New York: John Wiley & Sons, Inc., 1996.

[84] K. Sharat Reddy, *Source and system features for speaker recognition.* M.S Thesis, Indian Institute of Technology Madras, Department of Computer Science and Engg., Chennai, Sept. 2001.

[85] Yegnanarayana *et.al.*, "IITM speaker recognition system," in *Proc. NIST Speaker Recognition Workshop*, (Linthicum, MD, USA), May 2001.

[86] I. Gerson and M. Jasiuk , "A 5600 bps VSELP speech coder candidate for half rate GSM," in *EUROSPEECH*, vol. 1, pp. 253–256, 1993.

[87] K. Jarvinen , "GSM enhanced full rate codec," in *Proc. IEEE Int. Conf. Acoust. Speech, Signal Processing*, vol. 2, pp. 771–774, 1997.

[88] A. M. Kondoz, "Analysis-by-synthesis coding of speech," in *Digital speech coding for low rate communication systems*, pp. 141–213, New york : John wiley & sons Inc, Nov. 1995.

[89] S. P. Kishore, *Speaker verification using autoassociative neural netwrok models.* M.S Thesis, Indian Institute of Technology Madras, Department of Computer Science and Engg., Chennai, Dec. 2000.

[90] S. Furui, "An overview of speaker recognition technology," in *Automatic Speech and Speaker Recognition (C.H. Lee, F. K. Soong, and K. K. Paliwal, eds.)*, ch. 2, pp. 31–56, Boston: Kluwer Academic, 1996.

[91] S. Furui, "Recent advances in speaker recognition," *Pattern Recognition Lett.*, vol. 18, pp. 859–872, 1997.

[92] D. O'Shaughnessy, "Speaker recognition," *IEEE Trans. Acoust. Speech, Signal Processing*, pp. 4–17, Oct. 1986.

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

[94] T. V. AnanthaPadmanabha and B. Yegnanarayana, "Epoch extraction from linear prediction residual for identification of closed glottis interval," *IEEE Trans. Acoust. Speech, Signal Processing*, vol. ASSP-27, pp. 309–319, Aug. 1979.

[95] B. Yegnanarayana, S. R. M. Prasanna, and K. S. Rao, "Speech enhancement using excitation source information," in *Proc. IEEE Int. Conf. Acoust. Speech, Signal Processing*, (Orlando, Florida), 2002.

[96] S. S. Kajrekar and H. Hermansky, "Speaker verification based on broad phonetic categories," in *Proceedings of 2001: A Speaker Odyssey*, (Crete, Greece), 2001.

LIST OF PUBLICATIONS

JOURNALS

1. S.R.Mahadeva Prasanna, Cheedella S. Gupta, and B. Yegnanarayana, "Source Information from Linear Prediction Residual for Speaker Recognition" communicated to *Journal of the Acoustical Society of America.*

2. Jinu Mariam Zachariah, B. Yegnanarayana, S. R. Mahadeva Prasanna, and Cheedella S. Gupta, "Combining Evidence from Source, Suprasegmental and Spectral features for a Text-Dependent Speaker Verification" communicated to *IEEE Transactions on Speech and Audio Processing.*

CONFERENCES

1. Cheedella S. Gupta, S.R.Mahadeva Prasanna, and B. Yegnanarayana, "Autoassociative Neural Network Models for Online Speaker Verification using Source Features from Vowels" in *International Joint Conference on Neural Networks*, (Honolulu), vol.2, pp. 1252-1257, 2002.

2. S.R.Mahadeva Prasanna, Cheedella S. Gupta, and B. Yegnanarayana, "Autoassociative Neural Network Models for Online Speaker Verification using Source Features" in *International Conference on Cognitive and Neural Systems*, (Boston), 2002.