SOURCE AND SYSTEM FEATURES FOR SPEAKER RECOGNITION

A THESIS

submitted by

K. SHARAT REDDY

for the award of the degree

of

MASTER OF SCIENCE

(by Research)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY, MADRAS.

SEPTEMBER 2001

THESIS CERTIFICATE

This is to certify that the thesis entitled **Source and System Features for Speaker Recognition** submitted by **K. Sharat Reddy** 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

Dr. Sukhendu Das

(Dept. of Computer Science and Engg.)

ACKNOWLEDGEMENTS

I am extremely thankful to my guides Prof. B. Yegnanarayana and Dr. Sukhendu Das, and to all the members of Speech and Vision Lab for making my stay in this lab very special. I have been very fortunate to have Prof. as my guide. I am indebted to him for allowing me to take up a challenging thesis topic and for supplying me with support and creative insight. I have been benefited greatly from his advice and support. He taught me the importance of self organization and value of research. He has been a great source of inspiration and has provided the right balance of suggestions, criticism, and freedom. My interactions with him have been a constant source of encouragement and enthusiasm. This thesis would not have been possible without his generous efforts.

I also would like to thank Dr. Das for taking the role of supervisor during the initial phase of my research and for the support during that period. Thanks to Hema mam, for all the discussions we had during various meetings. Special thanks to Prof. C. Pandurangan, Head of CSE Dept., for providing a wonderful working atmosphere in the Computer Science Department. It is hard to imagine a nicer place to do research than at IIT Madras, where the social interaction of the students as well as OAT movies, SAC courts, and splendid natural beauty provide the right balance for successful research.

I consider myself to be very fortunate for becoming a member of Speech Lab. I had real wonderful time in the lab with all my lab mates. The cooperation and coordination in the speech lab always made me forget my home. Many thanks to my predecessor, Kishore for providing the outstanding help and technical support which made my work much easier. I also have to thank Surya (system admin), for patiently moving my home whenever I had some trouble and for ensuring the smooth running of my programs. I can never forget the encouragement and inspiration given by JyoTsna

(BBC) and Prasanna (mahadeva) when I needed them the most. They were always there to motivate me, both in good as well as in bad times.

I am grateful to Surya, Prasanna, and Jinu for their critical reading of my thesis drafts and for their valuable comments and suggestions. Without their help, it would have been impossible for me to complete my thesis draft in time. I want to thank Kiran(K), Ramesh, Gupta, Vinod, Srinivas Rao, Pal, Mathew, Ikku, P. Kiran, Anil, Nayeem, Kamakshi, Nagaraj, Dhanu, and Guru for their extended cooperation during my stay in IIT. I am also thankful to Jayant, Prasad, Devaraj, and Ve(Mo)nky for spending their time hitting ball with me either in T.T room or in SAC. Jayant, I can never forget the wonderful run of ours to two back-to-back T.T doubles championships, ofcourse the third one with Venky is equally memorable. I would terribly miss my machines, fant and sanskrit.

I definitely consider myself to be lucky for having such a great parents. I have reached this stage only because of their whole hearted support and their belief in my abilities. I thank my brother for his love and guidance at various stages of my career. Words won't be enough to express my gratitude to my friend Pavan, who went out of way to help me at the hours of need. My life would have been very different but for his guidance and true friendship. I am so proud to have a friend like him. This thesis is dedicated to them, especially my parents who instilled in me the value of education and the rewards and opportunities it can generate.

I don't know why but I fell in love with this lab the very first day and all my lab mates made it even more enchanting. The very thought of myself without this lab and all of you guys is so scaring. No matter where and what I will be, I will certainly miss you all.

-Sharat.

ABSTRACT

Keywords: *automatic speaker verification; autoassociative neural networks; source characteristics; higher order statistics.*

Automatic speaker recognition is the task of recognizing a person from his/her voice by a machine. Depending upon the text used, speaker recognition can be performed either in a text-dependent or a text-independent mode. In this thesis, an approach for capturing the speaker variability is proposed for text-independent speaker recognition.

Speech signal is produced by time varying excitation (source) of time varying vocal tract mechanism (system). The generated speech signal carries information pertaining to both source as well as system. All the present day speaker recognition systems use mostly vocal tract system (spectral) information, thereby ignoring the source information. But source characteristics are also expected to carry speaker-specific information which can be used for recognizing people.

In this thesis, we study the effectiveness of features extracted from the source component of the speech production process for the purpose of speaker recognition. Here, our assumption is that the source characteristics might be present in some higher order correlations among the speech samples. These higher order correlations are also referred to as higher order statistics in this thesis. A performance comparison is made between systems developed using features extracted from the source and vocal tract system components of the speech production process. The source and system components are derived using linear prediction (LP) analysis of short segments of speech. The source component is attributed to the LP residual derived from the signal, and the system component to the set of weighted linear prediction cepstral coefficients. The features are captured by a feedforward autoassociative neural network (AANN). The

performance is evaluated on a conversational telephone speech database of 80 male speakers. The study demonstrates the complementary nature of the two components. Through a study on the effect of training phase on the performance of the system, it is shown that for better performance speaker-specific training has to be incorporated. It is shown that higher the size of the input vector, and hence that network structure, the better is the performance. Use of source features for speaker recognition reduces the speech data required for both training and testing. Studies show that data as little as 6 sec is enough for capturing source characteristics of a speaker. We have observed that the performance of the source feature based speaker recognition system is as good as that of the vocal tract system (spectral) feature based speaker recognition system.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

ASR	- Automatic Speaker Recognition
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
BG	- Back Ground
LP	- Linear Prediction
LPC	- Linear Prediction Coefficients
NIST	- National Institute of Standards and Technology
VQ	- Vector Quantization

CHAPTER 1

INTRODUCTION

1.1 OBJECTIVE OF THE WORK

Every person has a unique voice using which that person can be recognized [1]. Automatic Speaker Recognition (ASR) [2] [3] is the task of recognizing a person from his/her voice by a machine. The objective of the present work is to explore the possibility of using excitation source characteristics of the speech production in the given speech signal as features for speaker recognition. The main idea in this approach is to capture the source characteristics related to glottal vibrations.

In this thesis, we address the issues involved in developing a text-independent speaker recognition system using non-spectral features such as the source features. The issue of capturing the speaker-specific source information is addressed using Autoassociative Neural Network (AANN) models. The robustness of source features and the amount of speech data required for automatic speaker recognition task are also discussed. A conversational telephone speech database is used in this study.

1.2 BACKGROUND TO SPEAKER RECOGNITION

1.2.1 What is Speaker Recognition?

From every day experience, it is clear that speech signal carries information about the speaker. Very often we are able to recognize a speaker from his voice. Given this fact alone, scientific curiosity prompts us to investigate how the speech signal encodes information about its producer, and how reliably that information can be extracted. *Speaker Recognition* may be defined as any activity whereby a speech utterance is attributed to a person on the basis of its acoustic-phonetic or perceptual properties [4] [5] [6].

In naive speaker recognition, the recognition is performed by untrained observers, for instance when answering the telephone call or hearing a voice in the next room. The decision is based on what is heard, and no special techniques are used. The term *Automatic Speaker Recognition* (ASR) first brings to mind the use of machines. Number of issues have to be addressed to get the speaker recognition work done by a machine. ASR make it possible to verify the identity of a person trying to access systems by voice. Before addressing the various issues involved in ASR, the significance of this problem and the importance of such a system in our day-to-day life is discussed.

1.2.2 Importance of Speaker Recognition

Speaker Recognition is one of the active fields of research. Physical tokens such as identity cards, badges, and passwords are currently being used for person authentication purpose. The aim of any one of the above tokens is to restrict the illegal entry of a person into system. Recently, biometric methods, such as recognition by voice(speaker recognition) [4] [5], iris recognition [7], face recognition [8] [9] and finger-print matching [10] [11] are also being used for person authentication purpose.

Lately, biometric methods are gaining importance and are being used more in real world applications than physical tokens, as the former has several advantages over the later. Unlike physical tokens, biometric methods are more secure, in the sense that they cannot be lost or stolen. Among various biometric methods mentioned above, authentication by voice is superior over other techniques because a person cannot fool the system based on voice as one cannot imitate other person's speech completely. Another advantage with this technique is that authentication can be done remotely. In other words, the person need not be present physically near the system, but it is enough to give his/her voice even over a telephone, thus making this technique superior to other techniques. Voice can also be used as a password for bank transactions over telephone, tele-shopping, voice mail and remote access to computers. Another important application is in forensic sciences where voice information is used for identifying criminals [12].

The importance of speaker recognition systems is bound to increase as in future there will hardly be any system which will not have a speech interface. For such systems to have effective security, they can be embedded with a speaker recognition system. As human beings, we don't have any problem in identifying familiar people from their speech. We shall proceed further to see the various characteristic features that contribute to the variability in a speaker's speech signal.

1.3 SPEAKER'S INDIVIDUALITY

We tend to think of people having a *voice* which we associate with them but it is not yet clear about the features and characteristics that are speaker-specific and which makes each speaker's voice unique. This section deals with the factors which might be contributing to the individuality of a person's speech.

1.3.1 Individuality in the Physical Mechanism: Speech as anatomy made audible

In speech analysis, it is useful to adapt a *source-filter* model of speech production. This means considering the larynx as a source of acoustic energy and the supra-laryngeal vocal tract as a filter or resonator which shapes that energy. The range of frequencies at which the vocal chords vibrate is determined by their length and mass. The vocal tract filter also varies in size and shape. Since phonetic properties such as resonant frequencies depends on the dimensions of the vocal tract, one's physique clearly influences how one sounds. All this will contribute to the distinctiveness of a speaker's voice. Differences due to physical mechanism are also known as organic differences.

1.3.2 Individuality in the Linguistic Mechanism: Speech as behavior

An act of speaking usually conveys more than a bare message. What we tend to think of as the message will also be accompanied by signals that indicate the attitude of the speaker [13] and also presents the speaker's self image. To the extent that speakers intend to convey these, they exploit what might be broadly termed their linguistic mechanism. This includes the lexical, syntactic, and prosodic resources of their language.

Most obviously, an individual has an accent. In general, this means that his/her speech allows them to be identified within a group. Phonetically, accents are differentiated along segmental phonology, prosody, and aspects of voice quality. Because accents subdivide the population speaking a language, they bring us some way towards identifying an individual.

1.3.3 Individuality in Phonetic Implementation

Each speaker learns one or more native languages, but in doing so he acquires more than the linguistic system, which defines what we think of as a language. Speakers acquire a socially and regionally marked variety of pronunciation according to their environment, and these variations constitutes the learned differences. Each speaker has to realize the complex resources of the learned and socially shared linguistic mechanism within the constraints of his/her anatomy. There is scope for differences in the implementation of linguistic functions which allows to distinguish one speaker from another.

If we imagine the many parameters which characterize a speaker as defining a location in a multidimensional space, we see that it is not a static point which characterizes the speaker but an area of variation. The fundamental issue in speaker recognition is whether each individual in the population occupy a unique area, or whether there is overlap regardless of how many parameters are used to characterize the speakers. If such uniqueness exists, is it possible to capture that uniqueness automatically. The next section focuses on various issues that need to be addressed for capturing the speaker-specific information automatically.

1.4 ISSUES INVOLVED IN SPEAKER RECOGNITION

Automatic speaker recognition is an application of pattern recognition. Speaker recognition system, like any other pattern recognition system, can be represented as shown in Fig. 1.1. This task involves three phases, feature extraction phase, training phase and testing phase [4]. Training is the process of familiarizing the system with the voice characteristics of a speaker, whereas testing is the actual recognition task. The issues in speaker recognition are discussed in [14] [15] [16] [17] [18].

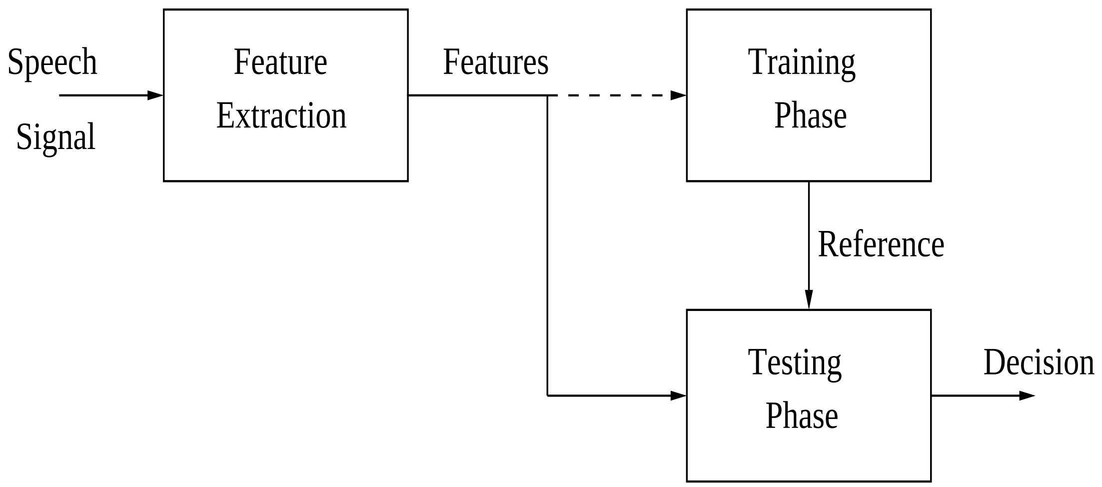


Fig. 1.1: Block diagram representation of a pattern recognition task.

For the automatic speaker recognition system to work, there are three major tasks to be accomplished. They are as given below:

1. Extracting appropriate speaker-specific features from the given speech data
2. Generating a model for each speaker

3. Developing matching algorithm and decision logic

We shall briefly discuss about each one of these three tasks.

1.4.1 Speaker-specific Feature Extraction

Feature vectors are parameter vectors extracted from the speech signal which captures the characteristics of the speaker. Feature extraction [19] is the key step in developing systems for automatic speaker recognition. Some of the desirable properties of the extracted parameters are:

1. High inter-speaker discriminability
2. Low intra-speaker variability
3. Robust to channel characteristics and noise

Normally, parameters used in the speech systems for recognition are some variations of the parameters used to represent the speech signal for coding and compression, which are derived based on spectral analysis. Feature vectors derived from spectral analysis [20] [21] satisfies the first two of the above mentioned desirable properties but not the last property.

1.4.2 Generating Speaker Models

Once proper set of feature vectors are obtained, the next task in speaker recognition is to develop a model (prototype) for each speaker. The development of speaker models comes under training phase. The block diagram is as shown in the Fig. 1.2.

Feature vectors representing the voice characteristics of the speaker are extracted in the feature extraction phase, and are used for building the reference models. The performance of ASR system depends primarily on the effectiveness of the models in capturing the speaker-specific information, and hence this phase plays a major role in the performance of a speaker recognition system.

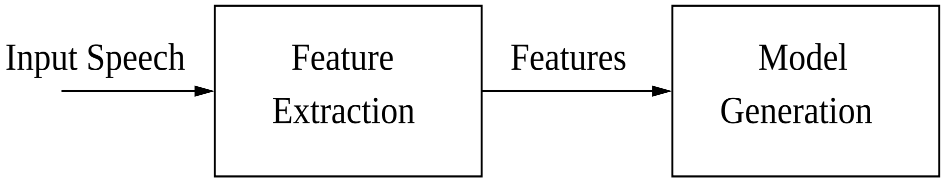


Fig. 1.2: Training phase of the process.

1.4.3 Matching Techniques

Matching techniques generally utilize the uniqueness in the speaker's feature vector distribution for discriminating the speakers. An effective matching technique should try to reduce the intra-speaker variability and increase the inter-speaker variability. The matching technique and decision logic comes under the testing phase of a pattern recognition system.

Matching techniques are of two types, template matching and probabilistic modeling. In template matching, the test feature vector is compared against the stored reference feature vector using a suitable distance measure. Probabilistic modeling involves modeling of speakers by probability distribution and deriving the classification decision based on the probability or likelihood.

The final stage in the speaker recognition system is the decision logic stage, where a decision, either to accept or reject the claim of a speaker is taken based on the result of matching technique used. 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. 1.3.

1.5 LIMITING FACTORS IN SPEAKER RECOGNITION

The primary limiting factors in ASR are:

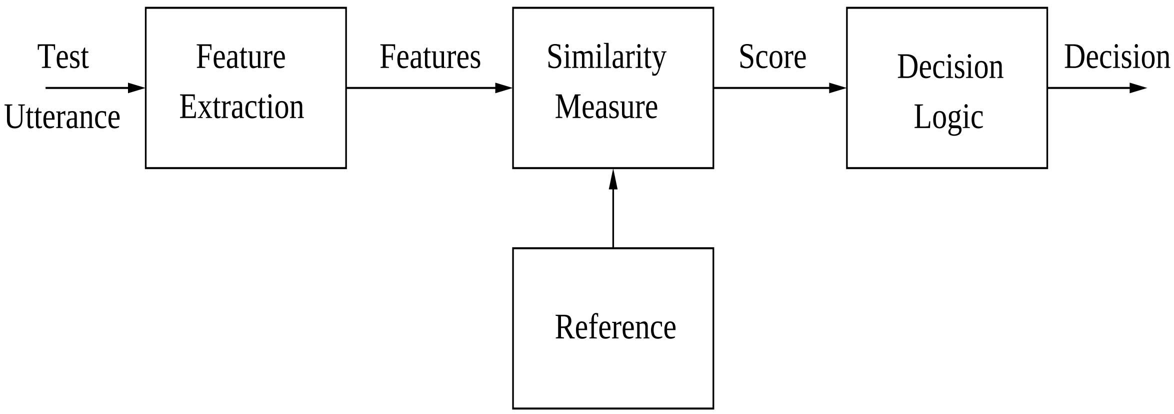


Fig. 1.3: Testing phase of the process.

- **Insufficient Data:** Large amount of speech data is required to capture the speaker-specific information effectively. Too small a data may not be sufficient to give a complete enough representation of the speaker in the higher dimensional feature space.
- **Quality of Data:** The performance of the system also depends upon the quality of the speech data being fed as input to the system [22]. If the quality of recording is poor (background noise, distortion imposed by telephone transmission or tape recording), the performance might be poor.
- **Voice Disguise:** There is always the possibility that a person is deliberately using voice disguise.

Present work is motivated by the first two limiting factors. Attempt to develop an approach which can work with small amount of speech data and also which is robust to noisy conditions is proposed in this work.

1.6 MOTIVATION FOR THE PRESENT WORK

For any pattern recognition task like ASR, the relevant information has to be captured in terms of suitable feature vectors. In speaker recognition, the feature vectors are in general some parameter vectors extracted from frames of the speech signal. Most of the present day ASR systems are developed using parameters that are derived based on spectral analysis, and the speaker variability is captured in terms of the distribution of these feature vectors. But, it is a fact that the spectrum of a signal is prone to channel characteristics and noise. Channel characteristics and noise play a prominent role in the performance of spectral feature-based systems [23] [24] [25]. Another drawback with the existing techniques is the way in which speaker-specific information is being captured. Mostly, they are statistical techniques, capturing the variability in terms of distribution of the feature vectors and hence large amount of data is required for a better estimate.

Since all the real world services have to deal with speech coming over telephone channel, the ASR systems have to be robust to environmental variations. Also, the requirement of large amount of data has to be overcome, as in the real world applications we may not have large amount of data to recognize a person. Hence, in order to make the ASR work in noisy conditions, and with less amount of data, features other than those derived based on spectral analysis also need to be explored.

The motivation to our work is the fact that the residual of a signal contains information mostly about the source, which might be more robust to environmental conditions. Since speech signal is produced as a result of an interactive process between the glottal excitation (source) and the vocal tract (system) [26] [27], the source information present in a given speech signal should also be speaker-specific. Hence, speaker recognition should be possible from source information present in the given speech signal.

The next section gives a brief overview of the organization of this thesis in terms of the chapters and their contents.

1.7 ORGANIZATION OF THE THESIS

This thesis is organized in 6 chapters. A brief overview about the organization of this thesis in terms of the chapters and their contents is as follows:

Chapter 2 gives a review of the research work done in the field of speaker recognition. It also introduces the feature space. The chapter starts with the classification of speaker recognition tasks and proceeds further with a discussion on the concept of speaker-specific Probability Density Function (PDF). Some traditional approaches to capture the speaker-specific distributions are also reviewed.

Chapter 3 introduces the concept of source and system from signal processing point of view. To facilitate the extension of this source-system concept for speech signals, a description of speech production mechanism is given. The concept of residual for ASR task is introduced. Work done on the utilization of the residual for ASR task is reviewed, and towards the end of the chapter, the need for a new approach is discussed.

Chapter 4 deals with the development of a source feature-based speaker recognition system using five layer AANN models. Analysis of AANN models is discussed. The procedure used to extract the source characteristics from the speech signal is explained. Performance of the speaker recognition system is evaluated on a database of 80 male speakers. Towards the end of the chapter, a performance comparison between spectral information based speaker recognition system and source information based SR system is given.

Chapter 5 focuses on the studies made for refining the performance of the system. The dependence of the system's performance on various parameters is examined. Robustness of the source features to amount of data available is demonstrated. The importance of training phase is also explained.

Chapter 6 concludes the thesis by summarizing the work. The scope for future work is also given.

CHAPTER 2

SPEAKER RECOGNITION: A REVIEW

Speaker Recognition is the process of automatically recognizing the speaker using the information obtained from speech data. In other words, it is authentication by voice. As mentioned earlier, it has several advantages over other authentication techniques. This chapter starts with a discussion on principles of speaker recognition and later presents an overview on recent advances in this field. The concept of feature space is introduced and various techniques for capturing the speaker-specific distribution of feature vectors in the feature space are discussed in detail. This chapter concludes with a discussion on limitations of some traditional approaches for speaker recognition, followed by a discussion on the new approach proposed to overcome these limitations.

2.1 TYPES OF SPEAKER RECOGNITION

Based on the application, speaker recognition can be divided into two categories [1] [2] [4] as: speaker identification and speaker verification.

2.1.1 Speaker Identification

Speaker identification [15] is the process of determining the speaker from a given test utterance from a set of registered speakers. Identification task can further be divided into two categories: closed-set identification and open-set identification. In the closed-set identification, the test speaker is always one of the N registered speakers. It need not be so in the open-set identification task. The block diagram of a typical closed-set speaker identification system is shown in the Fig 2.1. In the closed-set identification,

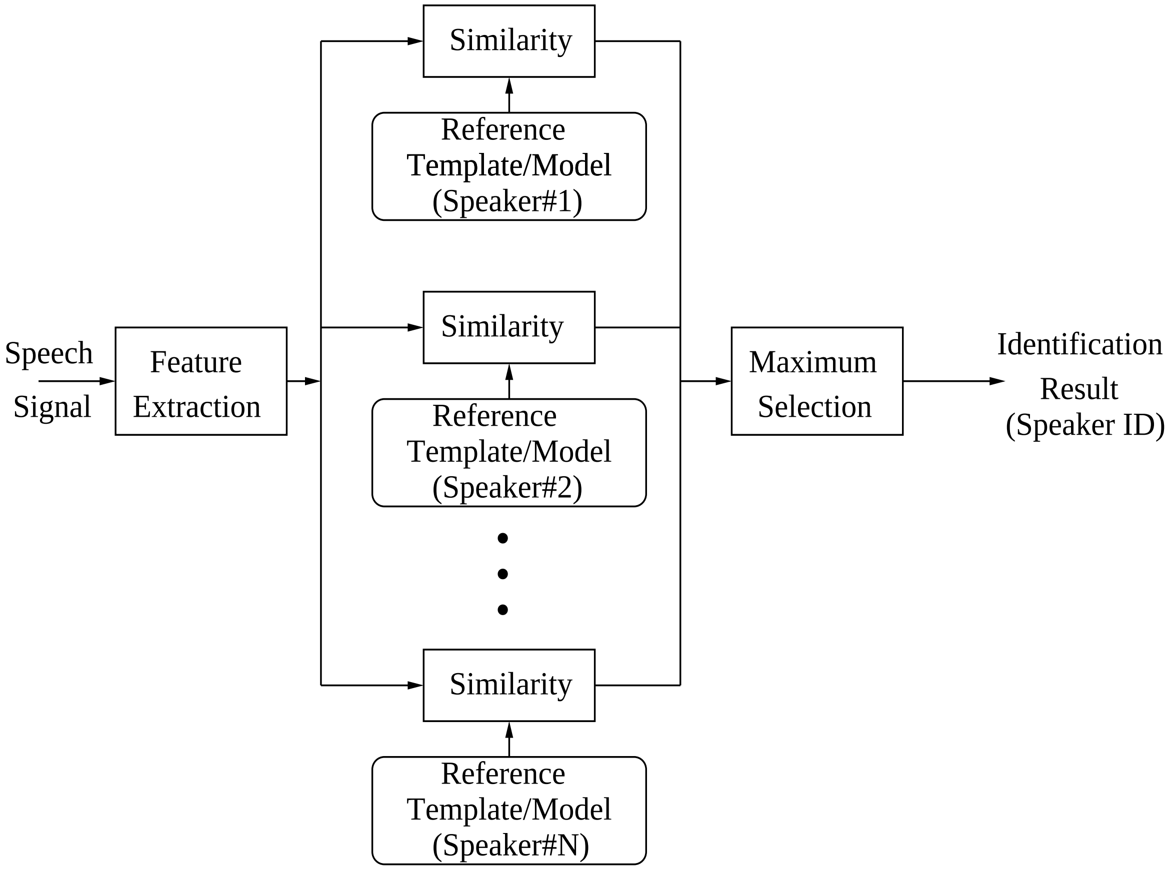


Fig. 2.1: Basic structure of a close-set identification system.

the speaker whose model best matches with the test utterance is declared as the identified speaker. The output of the system is the identity of the test speaker.

In the open-set identification, initially the speaker closest to the test speaker is found. Next, the distance of this speaker is compared with a threshold to come up with a decision. If the reference model for the unknown speaker is not present, an additional decision alternative, *the unknown does not match any of the models*, is required. Finding the optimum threshold is a difficult task in itself, and is an important issue to be addressed in the open-set identification.

2.1.2 Speaker Verification

Speaker verification [5] is the process of accepting or rejecting the identity claim of a speaker. Fig. 2.2 shows the block diagram of a typical speaker verification system.

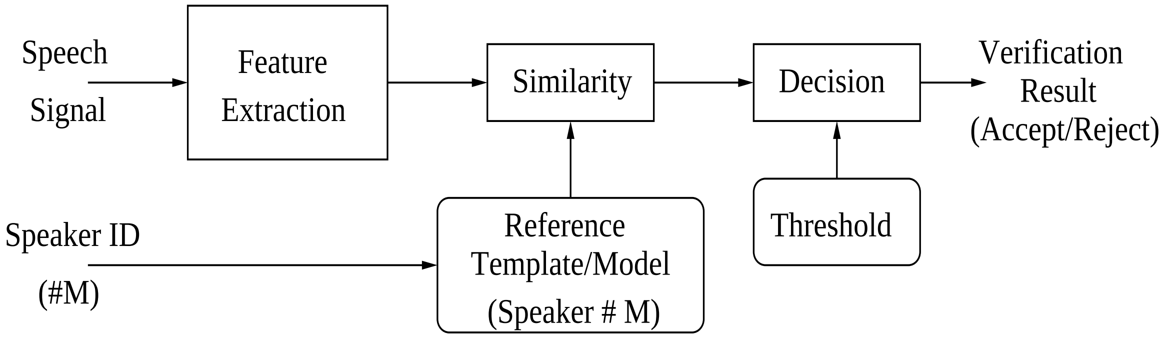


Fig. 2.2: Basic structure of a speaker verification system.

In speaker verification, when an identity claim is made by a speaker, the utterance is evaluated with respect to the model of the speaker whose identity is claimed. As in the open-set identification task, the concept of threshold is used in verification also. If the distance of the test utterance to the target model is below the threshold [28], then the claim of the speaker is accepted. The confidence with which the claim is accepted is presented as confidence score.

Unlike in identification, where number of decision alternatives are equal to the size of the population, in verification there are only two alternatives, accept or reject, regardless of the population size. Hence, the performance of the verification system does not depend on the size of the population.

2.2 TRADITIONAL METHODS FOR SPEAKER RECOGNITION

Traditional approaches for speaker recognition can be divided into three sub-categories based on the text employed for performing the task [3]. They are: Text-dependent methods, Text-independent methods, and Text-prompted methods.

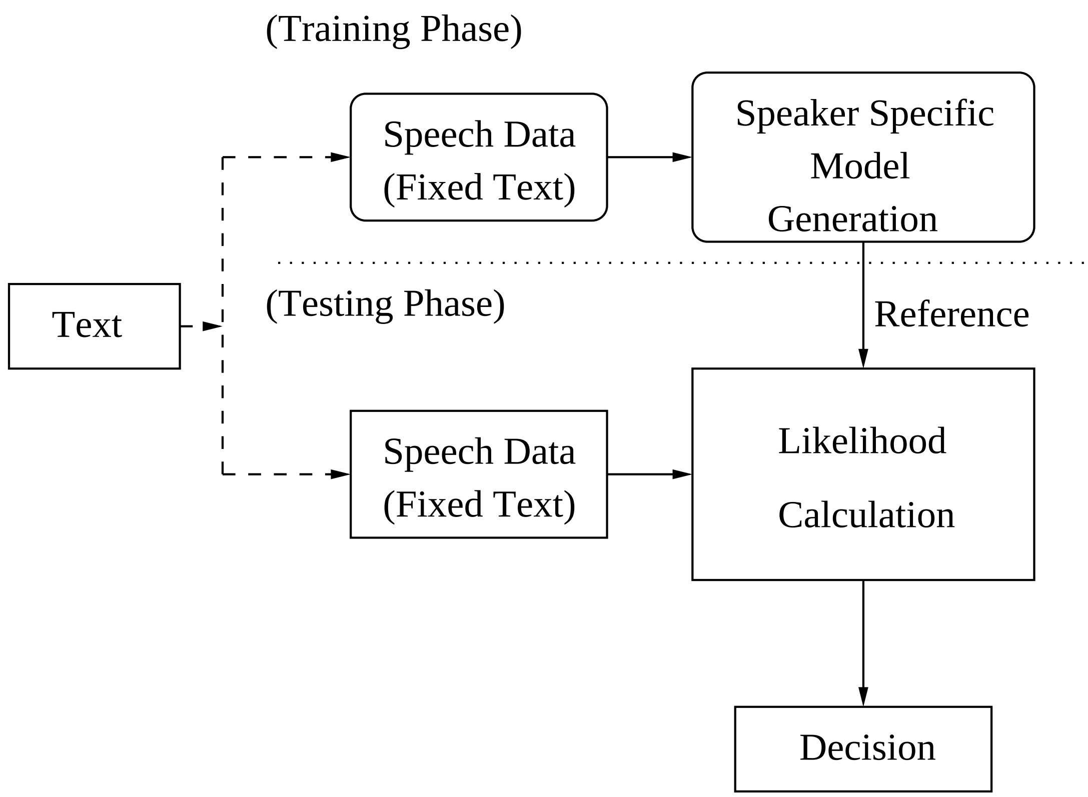


Fig. 2.3: Basic structure of a text-dependent system.

2.2.1 Text-dependent Speaker Recognition

Text-dependent methods require the speaker to provide utterances of the key words or sentences having the same text in both training and testing phases. Hence, text dependent systems are more convenient from the users point of view [3]. Fig. 2.3 shows the block diagram of a typical text-dependent system. Usually matching is done at the acoustic level, and therefore the performance is generally better. One common matching method employed for matching test and reference utterances is the Dynamic Time Warping (DTW) [29] technique. The performance of the system is related to the vocabulary that is chosen. For good performance, a phonetically balanced text is needed so that the system can capture the voice characteristics of the speaker in several phonetic contexts.

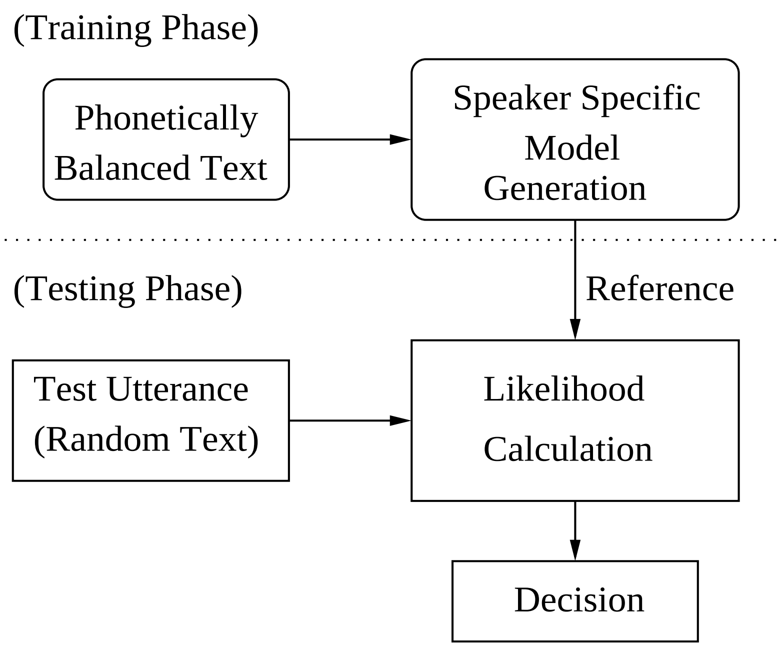


Fig. 2.4: Basic structure of a text-independent system.

2.2.2 Text-independent Speaker Recognition

In a text-independent speaker recognition system, the text of the utterances in training and testing need not be the same. Fig. 2.4 shows the block diagram of a typical text-independent system. In these systems, the statistical behavior of the speaker's feature vectors is captured during training. The same statistical characteristics are estimated at the time of testing, and they are compared with the characteristics captured in training phase. Since these methods use statistical characteristics, the performance is not as good as in the case of text-dependent systems. Various approaches to text-independent speaker recognition are dealt with in section 2.3.

2.2.3 Text-prompted Speaker Recognition

From security point of view there is a flaw in the earlier two methods. By storing recorded voice of a speaker and playing it back at the time of testing, one can cheat

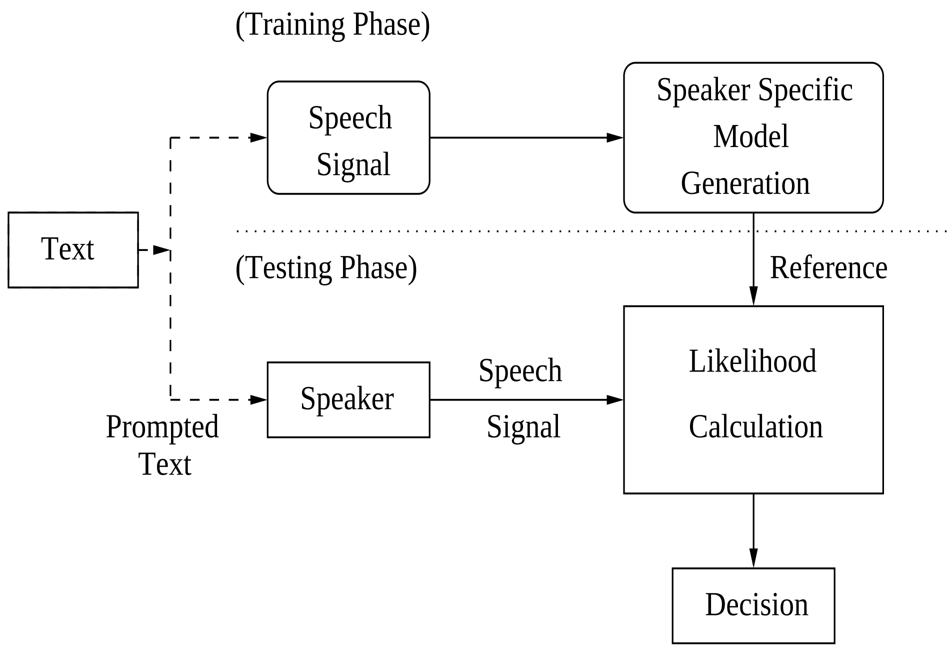


Fig. 2.5: Basic structure of a text-prompted system.

the system. A text-prompted system can overcome this problem to some extent. Fig. 2.5 shows the block diagram of a typical text-prompted system. In this system, the system prompts the user to speak some randomly selected words or sentences. The system first checks if the uttered words are same as prompted by the system. If the sequences matches, the claim is validated the way it is done in a text-dependent system.

In the present work, we are interested in developing a text-independent automatic speaker recognition system. It consists of two phases: Feature extraction phase and Distribution capturing phase.

2.3 FEATURE EXTRACTION

The purpose of feature extraction is to describe the acoustic properties of the speaker in the input speech data. Any information which is not useful in capturing the speaker-specific variability should be ideally eliminated or suppressed. In other words, it is the process of reducing data while retaining the classification information. The performance of an ASR system depends primarily upon the effectiveness of the feature vectors used.

2.3.1 Desirable Characteristics for Features

A set of desirable characteristics of feature vectors for speaker recognition are as given below [19].

- Efficient in representing the speaker-specific information.
- Easy to measure.
- Stable over time.
- Occur frequently in speech.
- Change little from one environment to another.
- Not susceptible to mimicry.
- High inter-speaker variations.
- Low intra-speaker variations.

For evaluating the features for their performance, a good measure of effectiveness is the F-ratio. It is defined as given below [30] [31] [32].

$$\mathrm{F} = \frac{\text{inter-speaker variation}}{\text{intra-speaker variation}} \tag{2.1}$$

F-ratio should be high so as to have high discrimination efficiency for a given feature vector. Sometimes, by weighting the feature parameters according to their effectiveness, the performance of the system can be improved [33] [34] [35].
After extracting a set of effective feature vectors, speaker characteristics are captured by estimating the distribution of these feature vectors in the feature space. Methods used for capturing the distribution are explained in the next section.

2.3.2 Feature Space

As with any pattern recognition task, the speech signal is first reduced to a sequence of feature vectors. Speaker recognition is often based on the premise that the speaker-specific information derived from speech utterance of an individual is characterized by a unique distribution of feature vectors. That is, the feature vectors extracted from the voice of an individual occupy a region in the feature space in a manner unique to him.

For better discrimination, the distributions of the feature vectors should have high inter-speaker variability and low intra-speaker variations as explained in the previous section. With increase in the size of the population, *crowding* [24] of the feature space occurs and this accounts for the degradation in the performance of the system.

Text-independent systems are based on modeling the speaker's acoustic feature space. The distribution for each speaker is determined from the feature vectors obtained from his speech. The speaker-specific distribution capturing is discussed in the next section.

2.4 DISTRIBUTION CAPTURING BY PROBABILISTIC MODELING

Probabilistic modeling refers to modeling of speakers by probability distributions, and deriving classification decision based on the probability or likelihood [15]. Text-independent speaker recognition is based on the premise that feature vectors, derived from the speech utterance of an individual, are characterized by a unique distribution. In the training phase, the speaker-specific feature distribution is captured. During testing, the probability that the test utterance is from a particular speaker's distribution is computed to arrive at a decision. Assuming that the distribution of the feature vectors is known for every speaker and has a continuous density p_i, the likelihood that a feature vector $\mathbf{x}$ is generated by i^{th} speaker is $p_i(\mathbf{x})$. Using Bayes rule, the probability that the speaker is i^{th} speaker is

$$P(speaker = i/\mathbf{x}) = \frac{p_i(\mathbf{x})P_i}{p(\mathbf{x})} \tag{2.2}$$

where, P_i is the prior probability of the i^{th} speaker, and $p(\mathbf{x})$ is the probability of feature vector $\mathbf{x}$ occurring from any of the speakers. Typically, priori probability for each of the speakers are assumed to be equal. The term $p(\mathbf{x})$ is the average of the speaker densities given by

$$p(\mathbf{x}) = \sum_{i=1}^{N} p_i(\mathbf{x}) P_i \tag{2.3}$$

where, N is the number of speakers.
Noting that $p(\mathbf{x})$ is same for all the speakers. If the prior probabilities are equal, the speaker to choose is simply the speaker having the highest likelihood.
Based upon the likelihood estimation, probabilistic modeling methods can be classified as [15]: Parametric methods and Nonparametric methods. Methods used for capturing the distribution of feature vectors are explained in the next section.

2.5 METHODS TO CAPTURE THE DISTRIBUTION OF FEATURE VECTORS

Approaches for text-independent speaker recognition can be broadly classified as:

- Parametric Approaches
  - Gaussian Mixture Model (GMM)
  - Hidden Markov Model (HMM)
- Nonparametric Approaches
  - Vector Quantization (VQ)
  - Artificial Neural Network (ANN)

Approaches that assume a model for the distribution are termed as parametric. The parameters of the model are estimated from the training set of feature vectors. In contrast, no such assumption is made in the nonparametric approaches.

2.5.1 Parametric Approaches

Parametric approaches are also model-based approaches. The parameters of the model are estimated using the training feature vectors. It is assumed that the model is adequate to represent the distribution. The most widely used parametric approaches are GMM and HMM based.

2.5.1.1 Gaussian Mixture Models

In GMM based approach, the underlying probability density function of the feature vectors of each speaker is captured using Gaussian mixtures [36] [37]. The complete Gaussian mixture density is parameterized by the mean vectors, covariance matrices, and mixture weights for all components. These parameters are represented as:

$$\lambda = \{p_i, \mu_i, \Sigma_i\}, i = 1, 2, 3, ...M \tag{2.4}$$

A Gaussian mixture density is a weighted sum of component densities given by the equation,

$$p(X/\lambda) = \sum_{i=1}^{M} w_i p_i(X) \tag{2.5}$$

where,

X is a D dimensional feature vector,

$p_i(X), i = 1, 2, 3...M$ are the component densities, and

$w_i, i = 1, 2, 3...M$ are the mixture weights.

Each component density is a D-variate Gaussian function of the form

$$p_i(X) = \frac{1}{|2\pi\Sigma|^{\frac{1}{2}}} exp\{-\frac{1}{2}(X-\mu_i)^T \Sigma_i^{-1}(X-\mu_i)\} \tag{2.6}$$

where μ_i is mean vector, and Σ_i is covariance matrix. The mixture weights satisfy the constraint that $\sum_{i=1}^{M} w_i = 1$.

During training, the goal is to estimate the parameters of GMM λ using the feature vectors collected from the training utterances. Most popular methods for estimating the parameters are Maximum Likelihood (ML) estimation and Expectation Maximization (EM) algorithm [38] [21]. The parameters are estimated to build a speaker-specific

GMM. One such GMM is developed for every speaker.

For speaker identification, a group of speakers are represented by GMMs λ_1, λ_2, $\lambda_3 ... \lambda_M$. The objective is to find the speaker model which has the maximum a posteriori probability for a given observation sequence. That is,

$$S = \underset{1 \leq k \leq S}{argmax} P_r(\lambda_k|X) \tag{2.7}$$

Assuming that all speakers are equally likely, using Bayes rule, the classification rule reduces to

$$S = \underset{1 \leq k \leq S}{argmax} P(X|\lambda_k) \tag{2.8}$$

For speaker verification, a threshold is fixed for the a posteriori probability and a speaker is accepted only if the probability of the claimed model is greater than the threshold, else he is rejected.

Usually GMM approach is not used directly. Instead, a background (BG) normalization technique [39] [40] is used to improve the performance. Here the fundamental assumption is that the BG model contains information about all the speakers and it is smooth. When the BG model is further trained with speaker's feature vectors, deviations are produced on the distribution surface of the BG model.

Discussion

- If the feature vectors do not follow a distribution that can be modeled by GMMs, the performance of the system will be poor.
- The performance of GMM-based system depends upon number of mixture components and database used [38] [41].
- It is a statistical approach, and hence requires a large amount of data to have a good estimate of the model.
- When training and testing environments are different, performance of the system decreases.

2.5.1.2 Hidden Markov Models

The HMM is a doubly embedded stochastic process where the underlying stochastic process is not directly observable. They have the capability of effectively modeling statistical variations in spectral features. In a variety of ways, HMMs can be used as probabilistic speaker models for both text-dependent and text-independent speaker recognition [42] [43] [44]. HMM not only models the underlying speech sounds but also the temporal sequencing among 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 DTW-based methods [45] [46]. But this temporal modeling does not aid in the text-independent case, because the sequence of sounds in the test utterance need not be the same as that in the training utterance.

In the training phase, an HMM for each speaker is obtained (i.e., parameters of model are estimated) using training feature vectors. The parameters of HMM are [27]:

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

  $$a_{ij} = P(q_{t+1} = j | q_t = i) \qquad 1 \le i, j \le N \tag{2.9}$$

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

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

  $$b_i(k) = P(o_t = v_k | q_t = j) \qquad 1 \le k \le M \tag{2.10}$$

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

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

  $$\pi_i = P(q_1 = i) \qquad 1 \le i \le N \tag{2.11}$$

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 testing phase, $P(O|\lambda)$ for each model is calculated, where $O = (o_1 o_2 o_3 .... o_T)$. Here the goal is to find out the probability, given the model, that the test utterance belongs to that particular model. The speaker whose model gives the highest score is declared as the identified speaker. GMM corresponds to the single-state continuous ergodic HMM [4].

Discussion

- The output of the HMM model is in terms of multiplication of several probabilities and therefore the scores are very small. Efficient scaling methods have to be derived to overcome this computational problem.
- As in the case of GMM, HMM is also a statistical approach, and hence requires large amount of training data for effective estimate of the model parameters.
- The system performance degrades when training and testing environments differ.

2.5.2 Nonparametric Approaches

As mentioned earlier, nonparametric approaches do not assume any model of the distribution. Since no model is assumed, the model parameters need not be estimated and hence these methods can capture arbitrary distributions. Two popular approaches in this category are VQ based and ANN based methods [47] [48] [49].

2.5.2.1 Vector Quantization

In the VQ-based method, codebooks are generated for each speaker. Codebooks consists of a small number of representative feature vectors to characterize a speaker [50] [51] [52] [53]. 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 [54] 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 [21] [55].

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 [56]. VQ can be considered as a degenerate case of single-state HMM with observation probability being replaced by the distance measure [4].

Discussion

- VQ based methods suffer from the problem of outliers, i.e., the vectors lying at the boundaries of the clusters create problems.
- A large amount of training data is required to have a good estimate of cluster representatives.
- Degradation of performance occurs if the training and testing environments are different.

2.5.2.2 Artificial Neural Networks

Artificial neural networks, especially Multilayer Feed Forward Neural Networks (MLFFNN) are well known for their mapping [57] [58] and generalization capabilities [59] [60] [61]. The potential of MLFFNN lies in its structure [60], as the processing units (neurons) operate in parallel. The nonlinearity of the processing units can effectively yield non-linear classification [62].

A MLFFNN consists of one input layer, with linear processing units, one or more hidden layers with either linear or nonlinear units, and one output layer with either linear or nonlinear processing units. The units in each layer are connected fully to the units in the next layer with suitable weights. The connection weights are adjustable, and the process of adjustment of weights is called as learning [59].

During training, the feature vectors are fed to the input layer of the network, and the corresponding output is calculated. This actual output is compared with the desired output and the difference between these two outputs, know as error, is taken as criteria to adjust the weights. The weights are adjusted such that the mean-square of this error is minimized. The network is trained till the training error converges to reach certain minimum value.

In the testing phase, each feature vector of the test utterance is fed as input to the trained network (models) and the error for each feature vector is computed. This error is accumulated over the entire test utterance. The accumulated error is considered for speaker recognition.

In [63] [64], one network for each speaker was used with each network having one node in the output layer. The desired output of the networks are kept binary: 1 if the frame belongs to the speaker, else 0. Here, the clustering property of MLFFNN was utilized. In [65], only one network was considered, with the number of output units being same as the number of speakers. In [66] AutoAssociative Neural Networks (AANN) are proposed for speaker recognition. The distribution capturing capabilities of AANNs is exploited for text-independent speaker recognition [67] [68].

There are many advantages in using ANN models. First of all, no assumption is made about the underlying distribution of the feature vectors. The distributions are learnt from the feature vectors themselves. ANN has good generalization capability. They are robust, for even if some of the units are corrupted, performance doesn't fall drastically.

Discussion

- There is no proof of convergence.
- Training is time consuming.
- Performance depends upon size of the network and training [4].

- Because of nonlinearity, analysis of network is difficult.

2.6 NEED FOR A DIFFERENT APPROACH

Most of the present day text-independent speaker recognition systems are GMM-based. But the problem with these systems is the assumption that the feature vectors of a speaker form a gaussian distribution. Feature vectors extracted from speech signal may have any arbitrary distribution which might not be effectively modeled using GMMs. A suitable nonparametric technique may overcome this disadvantage. Though one of the nonparametric techniques discussed in the previous section, namely Vector Quantization, do not impose any restriction on the distribution, it has other disadvantages. It requires a priori knowledge to have better modeling. Another major disadvantage with this is that it creates rigid boundaries in the feature space and hence will not allow any overlap in the feature vectors generated by two different acoustic classes.

To overcome these problems other nonparametric methods like ANN based methods are explored. The effectiveness of nonlinear models such as AANN models for speaker recognition is shown in [66] [69]. Efforts to use AANN models for speaker verification can also be found in [70] [71]. But even in these cases, the performance degrades if the training and testing environments are different. Also, large amount of speaker's data is required for developing model specific to the speaker.

The degradation of performance of the system with change in the environment can be attributed to the type of feature vectors being used. Automatic speaker recognition systems invariably use spectral features. But these spectral features are not robust to environment. In order to overcome this problem we should explore different type of features which can effectively represent the speaker-specific information and are robust to environment. We can solve the problem of requirement of large amount of speech data by using the features such that instead of capturing the distribution of the speaker's feature vectors, speaker-specific information is captured.

In the present work, we develop an approach to overcome some of the above problems.

Instead of spectral features, excitation source characteristics of speakers are used to capture the speaker variability. Source characteristics are likely to be more robust to noise, and hence the effects of environment on the system's performance can be reduced. In our approach, instead of capturing the distribution of feature vectors of the speakers, speaker-specific information is captured directly in terms of the source characteristics using Linear Prediction (LP) residual signal of the given speech signal. AANN models are used to capture the speaker-specific information. The next chapter is devoted to the concept of source and system for automatic speaker recognition. The concept of LP residual to represent the source information is also discussed in detail.

2.7 SUMMARY

In this chapter, a review of the research work done in the field of text-independent speaker recognition is presented. Apart from the techniques to capture the speaker-specific distributions, the limitations of these approaches are also discussed. The need for a new approach to overcome some of the existing limitations is highlighted.

CHAPTER 3

CONCEPT OF SOURCE AND SYSTEM FOR SPEAKER RECOGNITION

In the previous chapter, spectral-based methods for capturing the speaker variability in terms of distribution of the speaker's feature vectors were reviewed. In this chapter, a new method for capturing speaker variability in terms of speaker-specific information, using source characteristics of speakers is introduced.

Section 3.1 is an introduction to the concept of source and system. To understand the concept of source and system, Section 3.2 describes the speech production mechanism. Speech production mechanism is described in terms of source and system characteristics in Section 3.3. Section 3.4 addresses the speaker recognition task using source features. In Section 3.5, the concept of residual is discussed. Section 3.6 reviews the work done on utilization of the LP residual for speaker recognition task. The proposed approach is discussed towards the end in Section 3.7.

3.1 INTRODUCTION

A signal is defined as any physical quantity that varies with time, space or any other independent variable or variables [72] [73]. Speech sound is also a kind of signal for which the functional relationship between the signal and the independent variable is highly complicated [72]. To a high degree of accuracy, a segment of speech may be represented as a sum of several sinusoids of different amplitudes and frequencies.

Whenever any kind of physical work has to be done, some supply of energy is required and the work actually consists in converting this energy from one form into another. The generation of sounds is also a physical work and hence is governed very much

by the same laws as any other kind of phenomenon to be found in the universe [74]. They exemplify the effects of forces acting upon physical bodies to produce movements of various kinds and hence depends on a suitable source of energy. We can think of the energy supply as being the driving force and the system to which it is applied as the driven system. This driving force sets the systems into vibrations, which in turn produces sound. The amplitudes of these vibrations depend in part on how tightly the two elements are coupled together.
To have a better understanding of how this principle applies to speech signals, it is necessary to have a knowledge about the speech production mechanism, which is discussed next.

3.2 SPEECH PRODUCTION MECHANISM

Speech signal is produced as a result of time varying excitation of the time varying vocal tract system [27]. A schematic diagram of speech production mechanism is as shown in the Fig. 3.1

Speech production mechanism consists essentially of a vibrating source of sound coupled to a resonating system. For a great deal of the time in speech, the larynx is the source and the air column from larynx to the lips, that is the vocal tract is the system. But to produce some special sounds called nasal sounds, the vocal tract is replaced by nasal tract as the system. 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. A simplified representation of the complete physiological mechanism for creating speech is as shown in Fig. 3.2.

By the use of laryngeal muscles the vocal chords can be brought together so as to form as it were a shelf across the air way which leads from the lungs through trachea to

Fig. 3.1: Speech production mechanism.

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 approximately, as shown in Fig. 3.3. The ligaments and muscle fibers that make up these structures 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 showed that the dimensions of the air column involved were all 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. Hence the state of the vocal chords, shape and

Fig. 3.2: Representation of speech production mechanism.

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. 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 speaker recognition, is to effectively capture the variability due to the anatomical structure of the vocal tract which will help in automatic speaker recognition.

With this knowledge of speech production mechanism, we can extend the concept of source and system to the speech signals as discussed in the next section.

Fig. 3.3: Diagram of vocal fold motion.

3.3 CONCEPT OF SOURCE AND SYSTEM FOR SPEECH PRODUCTION MECHANISM

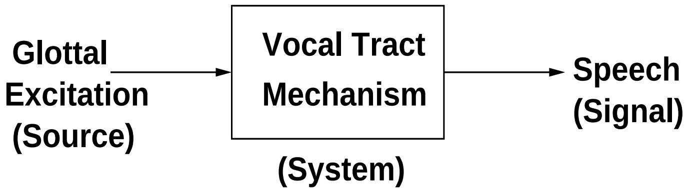


Fig. 3.4: Source and system representation of speech production mechanism.

Speech signals, as any other real world signals, are produced by exciting a system with a source [26]. A simple block diagram representation of the speech production mechanism is as shown in the Fig. 3.4. 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 and of which the amplitude, fundamental frequency, and 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. 3.4, 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. From signal processing point of view, this block diagram can be replaced with excitation and filter representation as shown in the Fig. 3.5. Here, 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.

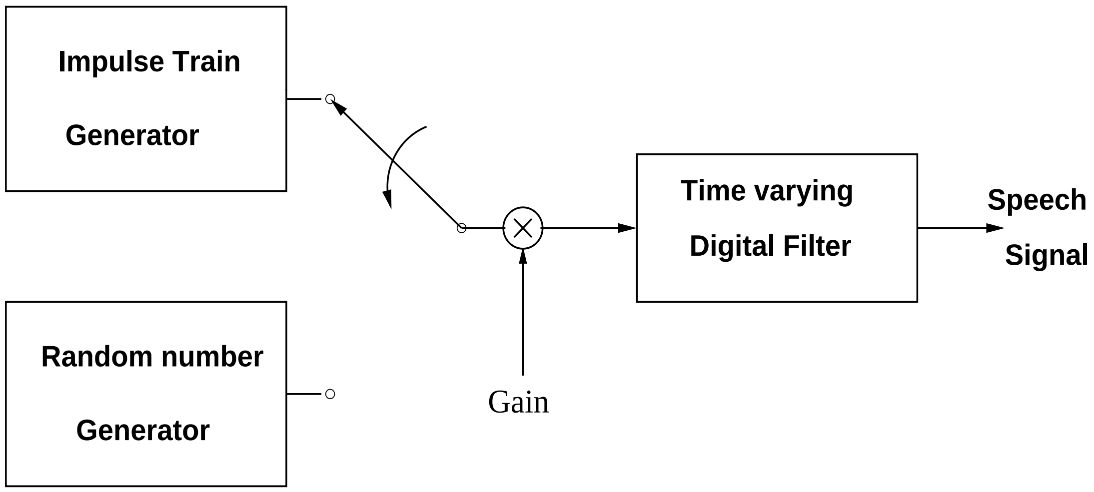


Fig. 3.5: Excitation and filter representation of speech production mechanism.

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 we denote the Fourier Transform of the source as $U(f)$ and if we consider the vocal tract as a time-invariant linear system, represented as $H(f)$, then the Fourier Transform of the output signal produced is given by

$$S(f)=U(f)H(f)$$

where, $S(f)$ is the Fourier Transform of the output signal.

In general the spectral envelope of the source function $U(f)$ is smooth. The transfer 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 spectral envelopes of the source, system, and the output signal are as shown in the Fig. 3.6.

Fig. 3.6: Spectral envelopes of source, system and output signal.

In brief, it is convenient to consider human speech production to be the result of the generation of one or more sources of sound and filtering of these sources by the vocal tract. To a large extent, the mechanism of source generation is independent of the filtering process, i.e., the properties of source tend not to be strongly influenced by the acoustic properties of the filters. Thus, it is appropriate to discuss source mechanism separately from the behavior of the vocal tract in response to the sources. To an approximation, source properties can be considered to be independent of the shape of the vocal tract mechanism. In other words, speech signal consists of two kinds of information, source information and system information. It is by the interplay of these two kinds of information, that the speech signal is generated.

Having known this information about the speech signal, in the next section we shall discuss the usefulness of either of the two kinds of information for speaker recognition task.

3.4 SOURCE CHARACTERISTICS FOR SPEAKER RECOGNITION

Speaker recognition systems have been developed mostly using spectral features for capturing speaker-specific information. Some of the spectral features used for speaker recognition task are short-time spectrum [75] [76], predictor coefficients [26] [27] [77] [78], cepstral coefficients [29] [79] [80], formant frequencies and bandwidths [81] [82], etc. In spectral analysis, the signal is considered as a band of sinusoidal signals, and the sum of the responses of the system to these sinusoidal signals is computed at the output. Hence, spectral analysis yields information about the system.

The speaker recognition systems capture the speaker variability in terms of system characteristics only, thereby neglecting the source characteristics. But we have seen in the previous section that speech signal carries information of both system as well as source, and hence the source information can also be used to capture the speaker-specific information present in the given speech signal.

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, which are specific to a speaker. Moreover, human listeners have proved themselves as robust speaker recognizers when presented with degraded speech and session variability [83]. This shows that source features might be more robust to degradations and session variability. It is this fact which has motivated us to address the speaker recognition task using source characteristics.

The significance of the source features in terms of speaker variability is shown in the Fig. 3.7. The speech utterances sampled at 8kHz were collected from three male speakers over a microphone. All the speakers spoke the sound unit /ka/. The significant instants of the glottal excitation are computed for the three speakers. The instants corresponding to the steady section of the utterances were displayed in the Fig. 3.7. It is clearly seen from the figure that the periodicity of the instants of glottal excitation for each of the three speakers is different from that of the other's.

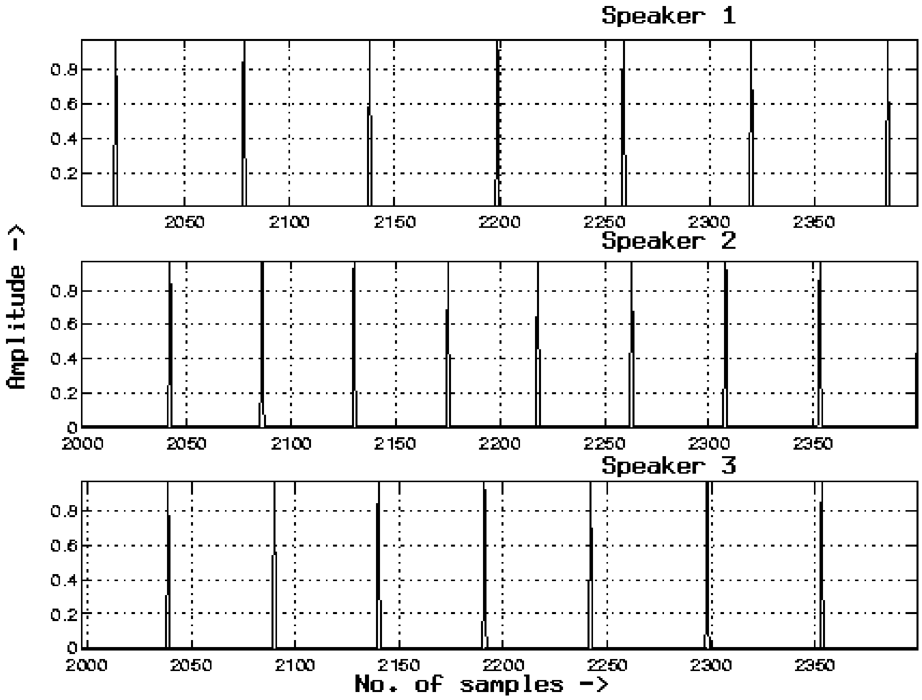


Fig. 3.7: Instants of significant excitations for three male speakers.

Though it is a known fact that this high level source information is certainly valuable in terms of speaker-specific information, not much effort has gone in developing speaker recognition system using this information due to practical difficulties involved in reliably extracting and using this information. In the present work we have derived a speaker-specific model using predominantly the source characteristics of the speech signal, and used this model for speaker recognition. AANN models are proposed to capture the speaker characteristics [84], and Linear Prediction (LP) residual signal is used as a representation of the source characteristics. The concept of residual for speaker recognition is explained in the next section.

3.5 LINEAR PREDICTION RESIDUAL

One of the most powerful speech analysis technique is the method of linear predictive analysis [78]. The philosophy of linear prediction is intimately related to the basic speech production model. The Linear Predictive Coding (LPC) analysis approach performs spectral analysis on short segments of speech with an all-pole modeling constraint [78]. Since speech can be modeled as the output of a linear, time-varying system excited by a source, LPC analysis captures the vocal tract system information in terms of coefficients of the filter representing the vocal tract mechanism. Hence, analysis of speech signal by LP results in two components, namely the synthesis filter on one hand and the residual on the other hand. In brief, the LP residual signal is generated as a by-product of the LPC analysis, and the computation of the residual signal is given below.

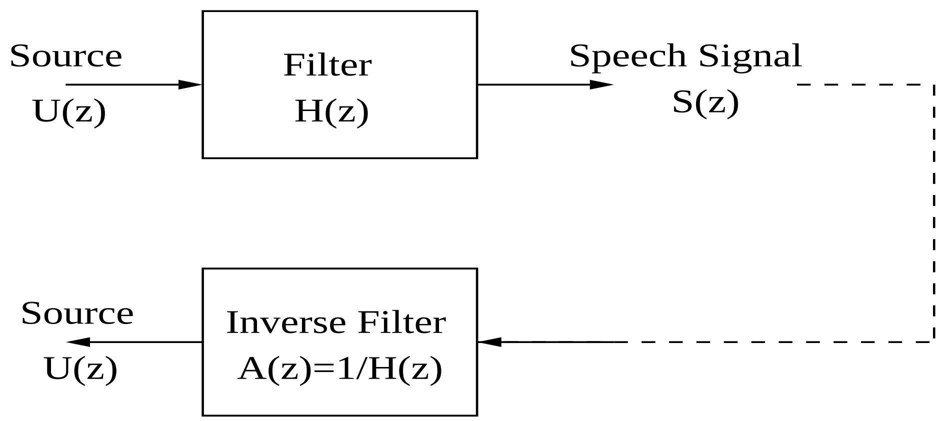


Fig. 3.8: Filter and inverse filter representation of the speech production mechanism.

In the previous sections we have seen the source and system representation of speech production mechanism. The discrete speech production representation of the same is as shown in the Fig. 3.8. If the input signal is represented by u_n and the output signal by s_n, then the transfer function of the system can be expressed as,

$$H(z) = \frac{S(z)}{U(z)} \tag{3.1}$$

where $S(z)$ and $U(z)$ are z-transforms of s_n and u_n respectively.
Consider the case where we have the output signal and the system and have to compute the input signal. The above equation can be expressed as

$$S(z) = H(z)U(z) \tag{3.2}$$

$$U(z) = \frac{S(z)}{H(z)} \tag{3.3}$$

$$U(z) = \frac{1}{H(z)}S(z) \tag{3.4}$$

$$U(z) = A(z)S(z) \tag{3.5}$$

where $A(z)$=$1/H(z)$ is the inverse filter representation of the vocal tract system.
Linear prediction models the output s_n as the linear function of past outputs and present and past inputs. Since prediction is done by a linear function, the name linear prediction. Assuming an all-pole model for the vocal tract, the signal s_n can be expressed as a linear combination of past values and some input u_n as shown below:

$$s_n = -\sum_{k=1}^{p} a_k s_{n-k} + Gu_n \tag{3.6}$$

where G is a gain factor.
Now assuming that the input u_n is unknown, the signal s_n can be predicted only approximately from a linear weighted sum of past samples. Let this approximation of s_n be $\tilde{s}_n$, where

$$\tilde{s} = -\sum_{k=1}^{p} a_k s_{n-k} \tag{3.7}$$

Then the error between the actual value s_n and the predicted value $\tilde{s}_n$ is given by

$$e_n = s_n - \tilde{s}_n = Gu_n \tag{3.8}$$

This error e_n is nothing but the LP residual of the signal. For a detailed description of LP analysis, see Appendix A.

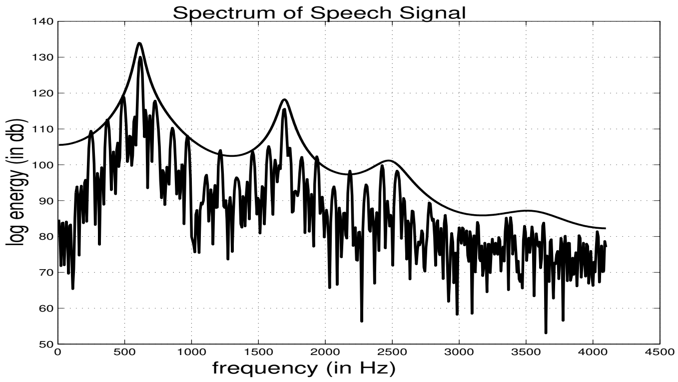
This LP residual, which is generated by LP analysis is usually ignored in all the major applications of speech analysis like speaker recognition. Only LPC coefficients are used to compute the feature vectors. But the residual signal is rich with source characteristics, which are also speaker-specific. Hence, the information present in the residual signal can be used for speaker recognition task. In the next section we shall review the work done in the direction of using the LP residual signal for speaker recognition task.

3.6 WORK DONE ON LP RESIDUAL FOR SPEAKER RECOGNITION

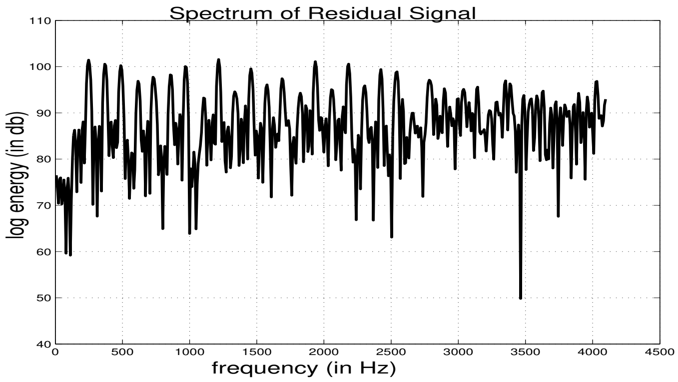
Though it is a known fact that residual contains information regarding the source, which is speaker-specific, not much work has been done in utilizing this information for speaker recognition task [2].

A few attempts [85] [86] have been made to utilize the speaker information present in the residual signal. In [86] the residual signal is converted into an one-sided autocorrelation sequence and the reflection coefficients are computed by performing FFT based cepstrum analysis. In [85] long-term average of the LP-residual real cepstrum is used to extract the information present in the residual signal. Some authors often summarize the whole residual in just one number represented by F_0, the fundamental frequency. But the residual on the whole carries richer information than the fundamental frequency.

It is quoted in the literature that information extracted from the residual signal using the above mentioned techniques yielded improvement in the performance when combined with the existing information [86]. From signal processing concepts, it is a known fact that the spectrum based features (like LPCCs), capture the gross level information. But the residual signal has much flatter spectrum (like white noise) representing the source characteristics rather than those of the vocal tract. The spectrum



(a)



(b)

Fig. 3.9: (a) Spectrum of speech signal. (b) Spectrum of residual signal.

of the signal and the residual for a speech segment are as shown in the Fig. 3.9. Spectral representation of the residual signal might not yield complete information present in the residual. The challenge is to extract this information effectively. For this, we propose a technique using AutoAssociative Neural Network (AANN) models, which is explained in the next chapter.

3.7 PROPOSED APPROACH USING RESIDUAL

In the proposed technique, AANN models are used for capturing the speaker specific information contained in the residual signal. Neural networks have the inherent ability

to model the nonlinearities contained in the underlying physical mechanism responsible for generating the input data. AANN trained with speech signal, captures the correlations present among the samples of the speech signal [87]. This concept motivated us to use AANN models for capturing information present in the residual.
Here, the hypothesis is that the source characteristics of the speaker may be present in the higher (> 2) order correlations among the samples of the speech signal, and AANN might capture these higher order correlations present among the samples of the speech data. The next chapter is devoted to the idea of effectiveness of source characteristics for speaker recognition task. A discussion on AANNs is also presented.

3.8 SUMMARY

In this chapter, we introduced the general idea of source and system, and extended this idea to the speech production mechanism. The concept of residual is introduced and a review of the work done on utilization of speaker-specific information present in residual for speaker recognition task is presented. Towards the end, the idea of source characteristics for speaker recognition using AANN models is proposed.

CHAPTER 4

SOURCE FEATURES FOR SPEAKER RECOGNITION USING AANN MODELS

In the previous chapter, we have discussed the concept of source and system for speaker recognition task. In this chapter, the effectiveness of AANN models to capture the source features for speaker recognition is studied. Comparison of performance with vocal tract system feature-based speaker recognition system is also presented in this chapter. To start with, a brief discussion on AANNs is given.

This chapter is organized as follows: The analysis of AANNs is presented in Section 4.1. Section 4.2 introduces the idea of AANNs for speaker recognition task. Section 4.3 deals with the details of source feature-based speaker recognition system. The data and the parameters used in our studies are also discussed here. Performance evaluation of the speaker recognition system is done in Section 4.4. In Section 4.5 a summary of this chapter is presented.

4.1 ANALYSIS OF AUTOASSOCIATIVE NEURAL NETWORKS

Autoassociative mapping neural network is basically a FeedForward Neural Network (FFNN) with network structure satisfying requirements for performing restricted autoassociation. FFNNs can be used for classification and mapping tasks [59] [88] [89] [90]. AANNs tries to map an input vector onto itself, and hence the name autoassociative or identity mapping. To understand the behavior of AANNs, the structure and mapping concepts of AANNs are discussed next.

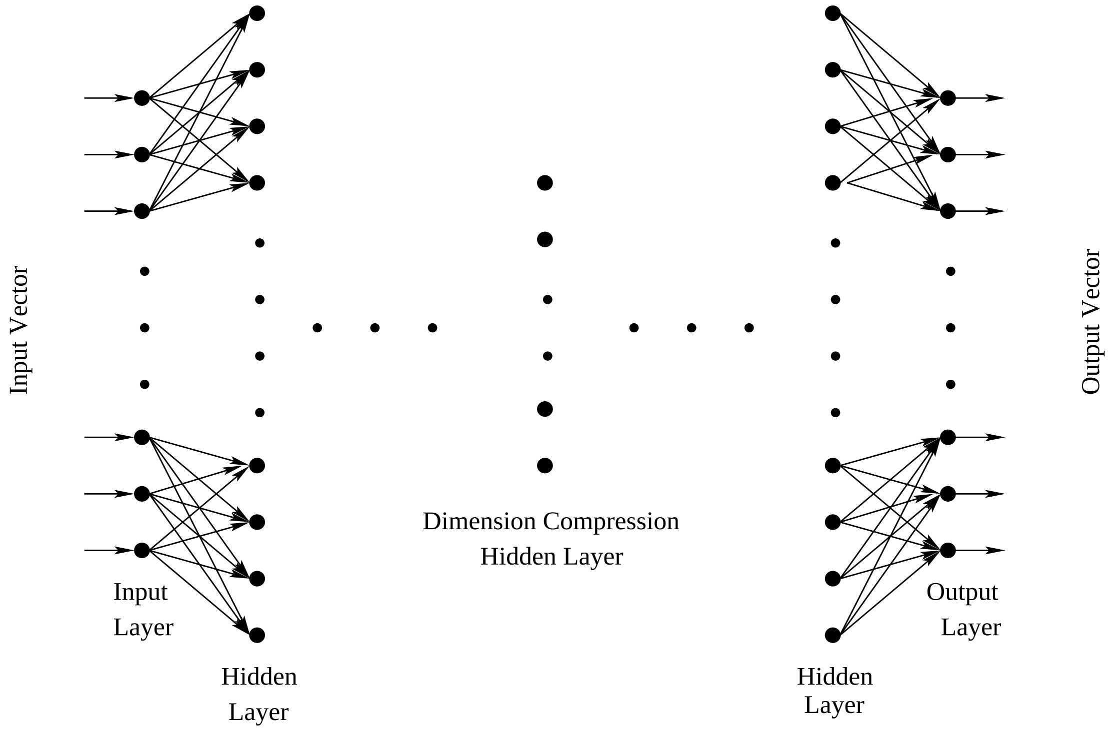


Fig. 4.1: General structure of AANN

4.1.1 Structure of AANN

AANNs are basically FFNNs with a general structure as shown in the Fig. 4.1. It consists of an input layer, an output layer, and one or more hidden layers. The number of units (also called as neurons) in the input and output layers are equal to the dimension of feature 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*. Output function of the units in the input layer is linear, whereas for units in hidden and output layers, it can be either linear or nonlinear. Structure of the AANN is generally specified in the form $x_iL, x_2N, ...x_dN, ...x_oN$, where L denotes the linear units and N the nonlinear units. x_i, x_d, and x_o denotes the number of units in input layer, hidden layer, and output layer, respectively. With this idea on the structure of the network, the concept of mapping is discussed next.

4.1.2 Concept of Mapping in AANN

Autoassociative mapping should reproduce an input vector at the output with least error. Ideally, the output should be the same as the input. AANN learns autoassociative mapping by training it in autoassociative mode with the training patterns.

From the network structure shown in the Fig. 4.1, the autoassociative mapping function (F) can be separated into two parts, F_1 and F_2, where

- F_1 is the transformation in the part of the network from input layer upto the dimension compression layer and
- F_2 is the transformation in the part of the network from dimension compression hidden layer to the output layer.

Since the number of units in the dimension compression hidden layer are less than the number of units in input and output layers, F_1 is basically a dimension reduction process and F_2 a dimension expansion process.

Dimension reduction is achieved in the network by projecting vectors in the input higher dimension space onto a nonlinear subspace of dimensionality equal to the number of units in the compression layer. Dimension expansion is achieved by mapping the lower dimension vectors onto a hyper-surface in the higher dimension output space. Dimensionality of hyper-surface is equal to number of units in the compression hidden layer.

Distribution modeling using AANN can be achieved by training AANN in autoassociative mode with the training feature vectors, using BackPropagation (BP) learning algorithm [91]. During testing phase, error obtained by testing the network with the test feature vector will give the likelihood of feature vector coming from the distribution learnt by the network. Attempts have been made for using AANNs for speaker recognition utilizing this distribution capturing ability [92] [93], and it was shown that the performance of AANNs based speaker recognition system is comparable with that of a GMM based system. AANNs for speaker recognition using source features is discussed next.

4.2 AANN MODELS FOR SPEAKER RECOGNITION

AANNs capture the distribution of the input feature vectors effectively. But, when an AANN is presented with raw data such as samples of speech or LP residual signal (which is of our interest here), the explanation of the behavior of AANN in terms of capturing the distribution is not appropriate. Adjacent frames in the residual signal may be widely separated in the input space, but are highly close in their behavior because of the higher order correlations present among the samples of the residual signal.

AANNs trained with speech signal captures the correlations present among the samples of the speech signal [87]. Since our hypothesis is that the source characteristics of speaker may be present in the higher (> 2) order correlations among the samples of given speech signal, AANN models have been used to capture this information. If speech signal is fed directly to the network, the 2^{nd} order correlations present among the samples (which is nothing but the spectral information) dictates the training of the network, suppressing the higher order correlations present if any. Since this 2^{nd} order correlation information is absent in the LP residual signal, the network is expected to capture the higher order correlations, which is the desired the source characteristics. The next section gives a detailed description of source feature-based speaker recognition system developed using AANN models.

4.3 SOURCE FEATURE-BASED SPEAKER RECOGNITION SYSTEM

For effectively capturing the speaker-specific source characteristics present in the LP residual signal, a five layer [66] AANN model with the structure as shown in the Fig. 4.2 is used. The structure of the network used in our studies is 40L 48N 12N 48N 40L. Speaker recognition being a pattern recognition task, the explanation of the system is divided into three phases as given below:

- Feature extraction phase

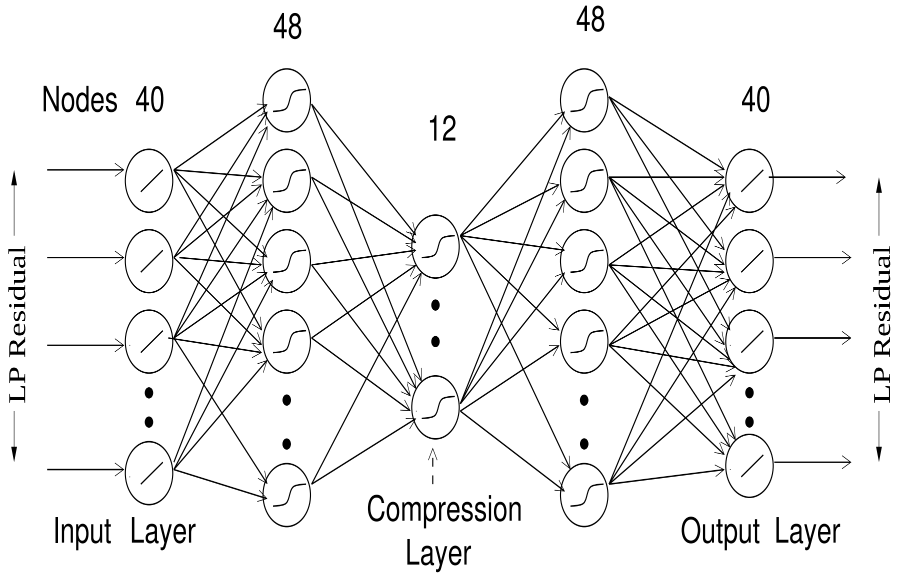


Fig. 4.2: AANN model for capturing source characteristics.

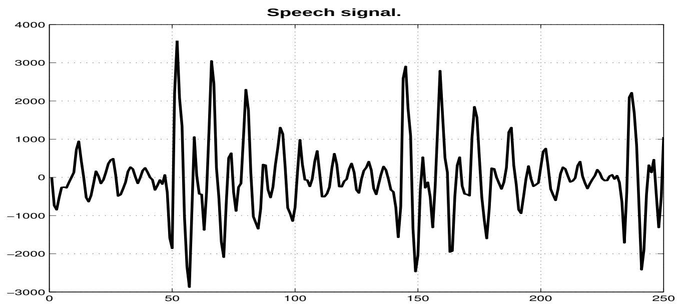
- Training phase
- Testing phase

Before proceeding to the description of the system, the various parameters used in this study are as listed below.

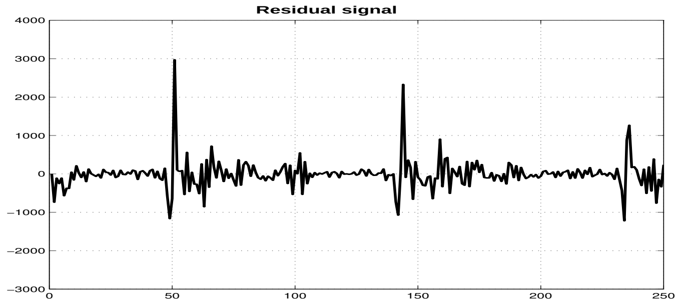
- Size of input (output) : 40 samples
- Frame shift : 1 sample
- LP order used to compute the residual : 8
- Number of epochs used in training : 60

4.3.1 Feature Extraction

LP order of 8 is used to compute the LP residual of the given speech signal. The residual signal of a typical speech segment obtained using LP order of 8 is as shown in the Fig. 4.3. To capture the source characteristics, a block size of 40 samples of the



(a)



(b)

Fig. 4.3: (a) Speech segment. (b) Residual of the speech segment obtained from 8^{th} order linear prediction.

LP residual signal is used as input vector to the AANN. But each of this 40 samples block is normalized to unit magnitude (norm=1) before feeding it to the network.

4.3.2 Training Phase

In the training phase, the models are generated by training the AANNs in autoassociative mode. The input to the network is the normalized blocks of the residual signal, each taken with one sample shift. The number of blocks used for training are nearly equal to the number of samples in the speech data of that speaker, after removing the silence and low energy frames. Each model is trained for 60 epochs using BP [91] learning algorithm. One such model is created for each speaker. After generating the models for all the speakers, the testing begins.

4.3.3 Testing Phase

In the testing phase, a test utterance of 60 seconds duration of each speaker is used. Input vectors (feature vector) are extracted as explained above. A block of normalized 40 samples of the LP residual signal is used as input to the AANN. For a model, the output for each block is compared with its input, to compute the squared error of that block. The error (E_i) for i^{th} block is transformed into a confidence value by using $c_i = exp(-\lambda E_i)$, where the constant $\lambda = 1$ throughout this study. This confidence value will be large for smaller values of error, i.e., blocks matching with the corresponding models. The c_i value will be low for large error value, thus giving less emphasis to nonmatching blocks. A given test utterance is compared with each of the claimant models to obtain the confidence value $c = \frac{1}{N}\sum_i c_i$ for each model, where N=number of blocks in the test utterance. This average confidence value per frame is used to compute the performance of the speaker recognition system. For a given test utterance, the score of the genuine speaker should be high, whereas that of an imposter should be low. The database used and the performance evaluation is discussed in the next section.

4.4 PERFORMANCE EVALUATION OF SPEAKER RECOGNITION SYSTEM

Before proceeding to the evaluation of the system, the database used for this study is discussed.

4.4.1 Speech Data

NIST 99 [25] development data is used for this study. Speech data of 80 male speakers is used. Each speaker's data consists of two 1 minute utterances, both collected over the same channel but in different sessions. One set of utterances is used for generating

Table 4.1: Performance of Speaker Recognition using source and system features. The table shows the rank of the speaker obtained by matching with 20 speakers.

	Speaker No.→	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
set I	**Rank of Model 1 (system features)**	**1**	**1**	**1**	**1**	**1**	**2**	**1**	**1**	**1**	**8**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**
	Rank of Model 2 (source features)	**2**	**1**	**1**	**1**	**1**	**1**	**4**	**1**	**1**	**1**	**1**	**2**	**1**	**1**	**1**	**17**	**1**	**1**	**1**	**1**
	Speaker No.→	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40
Set II	**Rank of Model 1 (system features)**	**1**	**1**	**1**	**4**	**1**	**1**	**1**	**1**	**1**	**1**	**2**	**1**	**1**	**1**	**1**	**1**	**5**	**1**	**1**	**1**
	Rank of Model 2 (source features)	**1**	**1**	**1**	**1**	**1**	**10**	**1**	**1**	**1**	**1**	**10**	**1**	**1**	**1**	**1**	**3**	**2**	**2**	**1**	**1**

speaker models and the other set is used as the test utterance in the testing phase. Details of the data are given in Appendix B.

4.4.2 Performance of the System

The 80 speakers are divided randomly into four independent sets of 20 speakers each. 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 speaker models of their respective sets. The average confidence value of all the 20 models for each of the test utterance is computed and this confidence value is used to rank the speakers. Ideally, the genuine speaker should have highest confidence value and hence rank one.

For comparison, a separate speaker recognition system based on vocal tract system characteristics is developed. The corresponding AANN model is as shown in the Fig. 4.4. The model using 19-dimensional weighted LP cepstral feature vectors captures the

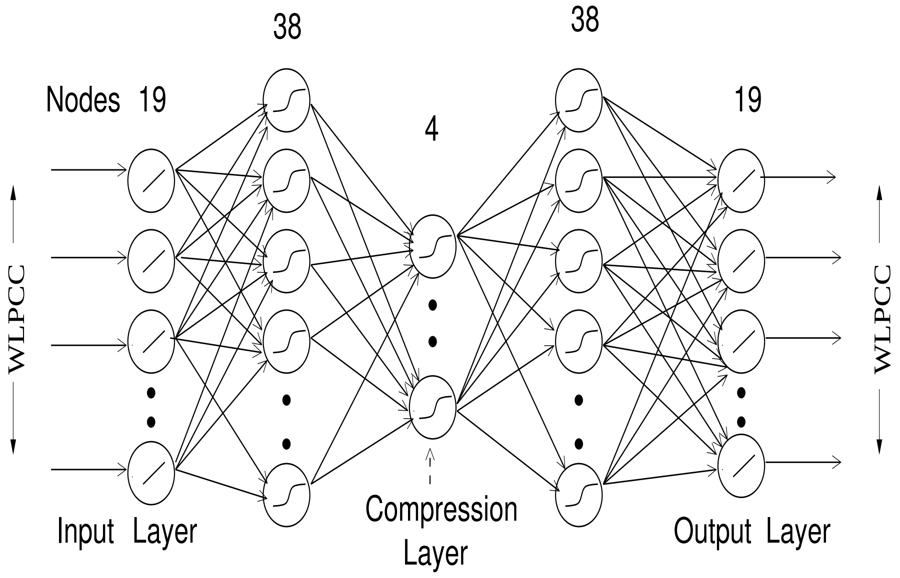


Fig. 4.4: AANN model for capturing vocal tract system characteristics.

distribution of the vocal tract system feature vectors of a given speaker [92]. The distributions are usually different for different speakers. Thus, an AANN model trained with one speaker's data captures the distribution of that speaker. Each model is trained with feature vectors derived from one minute of speaker data. The feature vectors are computed for every 27.5 msec frames separated by 13.75 msec. The model is trained using BP learning algorithm for 60 epochs [59]. Each feature vector is normalized to unit magnitude before giving as input to the network. One such model is generated for each speaker.

The behavior of the individual speakers for two sets of 20 speakers each, for both source feature-based and vocal tract system feature-based systems is as shown in the Table 4.1. The numbers in the table indicates the rank obtained by the genuine speaker for each of the 20 test utterances belonging to 20 speakers of the respective sets, taking the confidence value as criteria. The performance of both the systems is summarized in Table 4.2. Here, Model 1 is the vocal tract system features-based speaker recognition system and Model 2 is the source feature-based speaker recognition system.

Table 4.2: **Performance of the system feature-based and source feature-based speaker recognition systems for a set of 80 speakers.**

Type of Speaker recognition system ↓	No. of Models with		
	$Rank = 1$	$Rank = 2$	$Rank > 2$
Model 1 (system features)	70	3	7
Model 2 (source features)	64	7	9

From Table 4.2, it is evident that both features seem to give good performance. It is interesting to note that the source features are derived from the LP residual signal, which does not have any spectral information. Still it is giving recognition performance nearly as well as the system features. Another interesting observation that can be made from the Table 4.1 is the complementary nature of the source and the system components for speaker recognition. This complementary behavior can be used to develop a more robust and efficient system as explained next.

4.4.3 Performance of the Combined Model

Since the above result illustrate the complementary nature of the two types of information, we can improve the performance of either of the individual systems by combining these two kinds of information. A simple way of combining is to add the scores obtained by both the models, and then rank the test speaker according to new scores. In Table 4.3, these results are shown as rank of the combined model. The performance of the combined model is given in Table 4.4.

Table 4.3: **Performance of the Combined Model.**

Set I	Speaker No.→	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
	Rank of Combined Model	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**4**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**
Set II	Speaker No.→	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40
	Rank of Combined Model	**1**	**1**	**1**	**2**	**1**	**2**	**1**	**1**	**1**	**1**	**2**	**1**	**1**	**1**	**1**	**1**	**2**	**1**	**1**	**1**

The results of the Table 4.4 shows the improvement in the performance, in terms of overall ranking of the speakers. Yet, this is a simple way of combining the information. If better combining techniques [94] [95] are used for combining both kinds of information, the performance of the combined system may be improved over either of the individual systems.

Table 4.4: **Performance of the Combined model in comparison to source feature-based system, for a set of 80 speakers.**

Type of Speaker recognition system ↓	No. of Models with $Rank = 1$	$Rank = 2$	$Rank > 2$
Model 2 (source features)	64	7	9
Combined Model (system+source)	72	6	2

Though the results quoted above show the effectiveness of the proposed source feature-based speaker recognition system, there is definitely a need for further studies to be made as the various parameters chosen in this study are selected based on some preliminary experimentation. Moreover, there is a need for further analysis of the cases where the system performed poorly. The various studies made in this direction are discussed in the next chapter.

4.5 SUMMARY

In this chapter, an analysis of AANNs has been presented. The proposed source feature-based automatic speaker recognition system was explained and the performance of the system was studied. The effectiveness of the source features in comparison with the spectral features for speaker recognition has been shown. The complementary nature of source and vocal tract system information was discussed, and the improvement in the speaker recognition performance by combining these two evidences was illustrated.

CHAPTER 5

SPEAKER RECOGNITION STUDIES

5.1 INTRODUCTION

Performance of the source feature-based automatic speaker recognition system was discussed in the previous chapter. Various parameters used in this study are not optimal and are selected based on some initial experimentation. Hence, a further study on the dependence of the performance of the speaker recognition system on various parameter values is required. Also, to understand the various issues involved, detailed analysis of poorly performing speaker models is necessary. A detailed study of the behavior of source feature-based speaker recognition system is presented in this chapter.

This chapter is organized as follows: Section 5.2 deals with the effect of training on the performance of the system. The role played by the structure of neural network on the performance of the system is presented in Section 5.3. Time optimization studies through data reduction are discussed in Section 5.4. The performance of the optimized source feature-based speaker recognition system in comparison to that of the spectral feature-based system is discussed in Section 5.5.

5.2 SIGNIFICANCE OF TRAINING

Since the performance of a speaker recognition system depends on the generated speaker models, a detailed study is made to analyze the effect of training on the performance of the system. In the studies so far, the number of epochs used in the training phase for generating speaker models are 60. By a detailed study on variation of training error behavior, it is observed that in a few cases (mostly for poorly performing speakers), the training error is not converging. Fig. 5.1 shows the variation of training error with respect to number of epochs, for normally trained and poorly trained cases. This indicates that the learning is not complete and hence the generated model is not trained well.

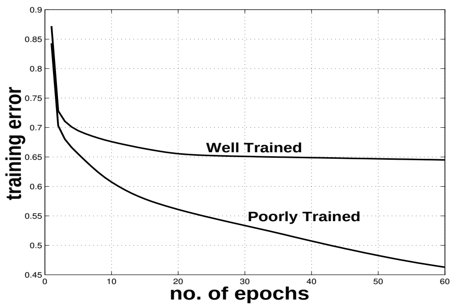


Fig. 5.1: Training error for well trained and poorly trained models.

To analyze the effects of training on the performance of the system, the poorly performing speaker models are retrained for an additional 60 epochs. Newly generated speaker models are tested to evaluate the performance. The change in the behavior of the individual speakers of Set-I with respect to training is as shown in the Table 5.1. The performance variation of the speaker recognition system with respect to training is as shown in the Table 5.2.

From the Table 5.2, it is clear that on extending the training beyond 60 epochs, the performance of the system improved significantly. A similar study was made on the spectral feature-based speaker recognition system, and a similar trend in the performance was observed. This study indicates that the performance of the system depends primarily on the speaker models generated in the training phase. It also indicates that training differs from speaker to speaker. For optimal performance, speaker-specific training has to be used for generating speaker models.

Table 5.1: Variation in the performance of the speaker's of Set I, before and after retraining the defective models.

	Speaker No.→	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Set I	Rank before Retraining	**2**	**1**	**1**	**1**	**1**	**1**	**4**	**1**	**1**	**1**	**1**	**2**	**1**	**1**	**1**	**17**	**1**	**1**	**1**	**1**
	Rank after Retraining	**1**	**1**	**1**	**1**	**1**	**1**	**2**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**16**	**1**	**1**	**1**	**1**

Table 5.2: Performance of the source feature-based speaker recognition system before and after retraining, for 80 speakers.

Number of epochs used↓	No. of Models with		
	$Rank = 1$	$Rank = 2$	$Rank > 2$
Before retraining (60 epochs)	64	7	9
After retraining (120 epochs)	70	5	5

5.3 EFFECT OF NETWORK STRUCTURE ON THE PERFORMANCE

5.3.1 Role of Dimension Compression Hidden Layer

The structure of the AANN model plays a significant role in capturing the distribution of the given data. The units in the dimension compression layer reduce the dimension of the data. The number of units in the compression layer determines the number of components captured by the network. But depending upon the data, the use of fewer components of the data may lead to loss of information.

Table 5.3: Effect of Number of Nodes in Compression Layer on System's Performance. The table shows the ranks obtained by 20 speakers of set I.

No. of Nodes in Compression layer ↓	No. of Models with		
	$Rank = 1$	$Rank = 2$	$Rank > 2$
12	16	2	2
8	15	3	2
6	14	2	4

In the present approach, AANN is used for capturing the higher order correlations that are present among the samples of the given speech signal. Since the proposed approach is not based on the distribution capturing ability of the AANN, the role played by the number of units in the compression layer on the performance of the system is expected to be insignificant. To analyze the behavior of the system, the performance is obtained with varying number of nodes in the compression hidden layer. The results are as shown in the Table 5.3.

From the Table 5.3, it is interesting to note that the performance of the system is nearly retained even after reducing the number of nodes in the compression layer from 12 to 6. This behavior is not totally unexpected for the fact that it is the higher order correlations present among the samples of the speech signal that the network is expected to capturing and not the distribution of the feature vectors in higher dimensional feature space. This study demonstrates the robustness of the source features to variations in the network structure.

Table 5.4: **Effect of Input Vector Size (Network Structure) on System's Performance. The table shows the ranks obtained by 80 speakers.**

Network Structure ↓ (L=linear, N=non-linear)	No. of Models with $Rank = 1$	$Rank = 2$	$Rank > 2$
30L 37N 12N 37N 30L	56	9	15
40L 48N 12N 48N 40L	64	7	9
50L 64N 12N 64N 50L	69	5	6
80L 100N 12N 100N 80N	74	1	5

5.3.2 Effect of Input Vector Size

In all the previous studies, the size of the input vector used is 40 (block of 40 samples of residual signal). To study the effect of input vector size on the performance of the system, four cases are considered, each consisting of models trained with a different input vector size. The sizes of the input vector considered in this study are 30, 40, 50, and 80. Number of units in the compression hidden layer are kept constant (12) across all the cases. Each of them are evaluated independently, and the results are given in the Table 5.4.

From the Table 5.4, it is clear that higher the input vector size, the better is the performance of the system. Since it is the higher order correlation information that the network is expected to capture, the performance will be better if the block size of the residual signal used as input to the network is large. Another reason to this behavior is the nature of the LP residual signal. A typical segment of the LP residual signal is as shown in the Fig. 4.3. It can be observed that around the instants of significant excitation, the amplitude of the samples of residual signal is high and as we move away, the amplitude falls drastically. Around the instants of excitation, the correlation information among the samples is high. Typically, the time interval between two instants of excitation will be in the range of 70-80 samples. As the size of the input vector increases, more and more frames span over the instants of excitation and the amount of correlation information captured will be high, and hence is the improvement in the performance. But the price that has to be paid for this improvement in the performance is the increase in the network structure and hence the computating time for training the speaker models. A compromise has to be reached between the size of input vector (network size) and the computating time to optimize the performance of the system.

Though the studies show the effectiveness of source features in comparison with spectral features for ASR task, source feature-based system is computationally very expensive. While generating a speaker model using spectral features take 4 minutes, source feature-based system takes 7 hours for generating the same model on a Pentium III system. To make the source feature-based system work in real-time, the training time has to be drastically reduced. The two major factors contributing to high training time are network structure and amount of data used. But our previous studies showed that more the size of the input vector and hence the network structure, the better is the performance of the source feature-based system. The other alternative to reduce the computating time is to reduce the amount of data. The next section focuses on the studies made for optimizing the computating time through data reduction.

5.4 TIME OPTIMIZATION STUDIES

5.4.1 Size of Data

All the traditional speaker recognition systems explained in Chapter 2 are statistical methods. Since statistical methods capture speaker variability in terms of the distribution of the speaker's feature vectors in the feature space, the performance of

Table 5.5: **Performance variation of the speaker recognition system with the amount of training data used (for 80 speakers).**

Amount of training data used (in seconds)↓	No. of Models with		
	$Rank = 1$	$Rank = 2$	$Rank > 2$
60	64	7	9
12	63	8	9
6	59	11	10
3	43	18	19

these systems depends on the amount of data available for training and testing. If the amount of available data is small, the captured distribution of the feature vectors in the higher dimensional feature space will not be accurate. This results in a poor performance. The studies discussed below emphasizes the role played by the amount of data used in training and testing phases on the performance of the source feature-based system.

Training Data

In the evaluation studies quoted in the previous chapter, one minute of speech data was used for generating each of the 40 speaker models. Using different amounts of training data for generating the speaker models, four different source feature-based speaker recognition systems are developed to study the influence of the size of training data on the system's performance. The four systems use 60 seconds, 12 seconds, 6 seconds, and 3 seconds of speech data for generating the speaker models. The four systems are evaluated independently, as explained in the previous chapter. The result is as shown in the Table 5.5. The behavior of one set of (set-II) models with varying

Table 5.6: **Variation in the performance of the speaker's of Set II, with change in amount of training data used.**

	Speaker No.→	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Set II	Rank obtained for 60 sec.	**1**	**1**	**1**	**1**	**1**	**10**	**1**	**1**	**1**	**1**	**10**	**1**	**1**	**1**	**1**	**3**	**2**	**2**	**1**	**1**
	Rank obtained for 6 sec.	**1**	**1**	**1**	**1**	**1**	**4**	**5**	**1**	**1**	**1**	**5**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**3**	**14**

training data is as shown in the Table 5.6.

From the Table 5.5, it is evident that 6 seconds of data seems to be enough for capturing the speaker-specific information. Since the speaker variability is captured in terms of higher order correlations present among the speech samples, reduction in the amount of speech data for generating speaker models does not effect the performance of the system significantly. But as can be seen from the table, when the training data is reduced to 3 seconds, the performance dropped indicating that certain minimum amount of speech data is required for capturing the higher order correlation information present among the samples. From the Table 5.6, it is interesting to note the varying behavior of the speakers with varying amounts of training data used. This shows that the performance of the system is predominantly determined by the models generated. The data used during the training phase determines the performance of the each model. This feature may be used for improving the performance of the system by generating more than one model for each speaker and combining the evidence from each of the speaker's models for taking a decision.

Testing Data

To examine the effect of amount of testing data on the performance of the system, we considered two different cases, each using different amount of testing data. The two cases considered in this study use 60 sec and 6 sec of data in the testing phases. Models trained with 60 sec of speaker's data are used in this study. The performance

Table 5.7: Performance variation of the speaker recognition system with the amount of testing data used (for 80 speakers).

Amount of testing data used (in seconds)↓	No. of Models with		
	$Rank = 1$	$Rank = 2$	$Rank > 2$
60	64	7	9
6	63	8	9

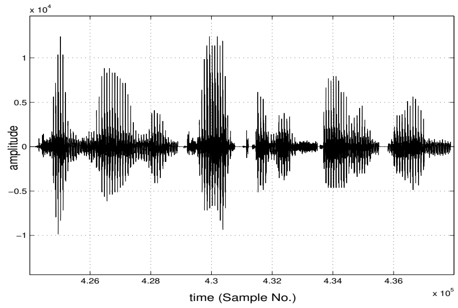
variation with respect to amount of testing data used is shown in the Table 5.7.

It can be noted from the Table 5.7 that the amount of data used during testing does not have significant influence on the performance of the system. This result once again emphasizes the fact that the performance of the system primarily depends upon the models generated in the training phase, and any deficiency in the training will result in degradation of the performance.

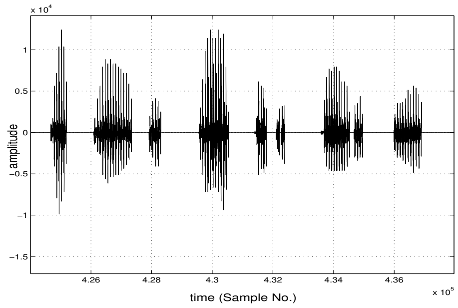
5.4.2 Nature of Data

The performance of any pattern recognition task depends primarily on the quality of the input data. To a large extent, the quality of speech data might also influence the performance of the speaker recognition system. The studies discussed below illustrates the effect of quality of speech data on the system's performance.

Before processing the speech signal for extracting appropriate feature vectors, the silence and non-speech regions have to be eliminated. Apart from not carrying any speaker-specific information, silence and non-speech frames add confusibility during testing. The algorithm used in this study to detect the speech frames is based on the



(a)



(b)

Fig. 5.2: Silence detected speech segment with (a) 0.1 threshold (b) 0.9 threshold.

amplitude of the speech signal. The speech signal is blocked into frames using the specified frame size and frame shift. The maximum positive amplitude in each frame is determined. The sum of mean and a fraction (1/10) of the standard deviation of these positive amplitudes is considered as the maximum amplitude value in the signal. A fraction (in percentage, determined by the selected constant value) of the maximum amplitude is taken as the threshold for a frame to be considered as a speech frame. The frames above the threshold are marked as speech frames, whereas the remaining frames are marked as silence/non-speech frames.

In the previous studies the silence and non-speech regions were detected and eliminated by setting the constant value to 10%. But analysis showed that the value is too small a threshold to effectively remove all the silence and non-speech regions. A study to evaluate the behavior of the system for higher thresholds is made and the results are

Table 5.8: **Performance variation of the speaker recognition system with the nature of speech data used (for 80 speakers).**

Type of data used ↓	Performance (No. of models) for training data of					
	6 seconds			3 seconds		
	$Rank = 1$	$Rank = 2$	$Rank > 2$	$Rank = 1$	$Rank = 2$	$Rank > 2$
non-contiguous speech data	59	11	10	43	18	19
contiguous speech data	63	9	8	55	11	14

as explained below. Increase in the threshold value removes all the non-speech regions and takes into consideration only those regions where the amplitude of the signal is very high.

With increase in the threshold value, there is a general trend in the increase in the scores for all the speakers. The change in the score values for one set of 20 speakers for two different threshold values is as shown in the Table 5.10. It can be observed from the table that increase in the confidence value for a genuine speaker is more when compared to that of imposter speakers.

Increase in the threshold value leads to breakup of the speech segments into smaller, non-contiguous segments. Fig. 5.2 shows the silence detected speech segment with threshold values of 0.1 and 0.9 respectively. This problem can be overcome to certain extent by considering only those regions in the speech signal where contiguous high energy speech segments are present for generating the models and for estimating the confidence scores. Considering only contiguous 100 msec speech regions for generating models, the performance of the system is obtained. The result is as shown in the

Table 5.9: **Performance comparison of the optimized and unoptimized source models with that of the system model (for 40 speakers).**

Type of system ↓	No. of Models with			Computational time required
	$Rank = 1$	$Rank = 2$	$Rank > 2$	
source model (unoptimized)	64	7	9	7 hours
source model (optimized)	71	3	6	30 minutes
system model	70	3	7	4 minutes

Table 5.8. For comparison, the result obtained using non-contiguous speech data is also given in the table.

Table 5.8 clearly demonstrates the significance of the nature of speech data used. It can be observed that the change in the performance of the system developed using 3 sec of data is more when compared to that developed using 6 sec of data, indicating the fact that when the amount of data and hence the available speaker-specific information is reduced, the nature of the speech data has to be good. Since we are interested in higher order correlations, the performance of the system can be improved by using only high energy, *contiguous* speech data. This study also shows that the quality of the speech data is as important a factor as the amount of speech data used.

5.5 OPTIMIZED SPEAKER RECOGNITION SYSTEM

The various studies mentioned above demonstrates the dependence of the performance of the source feature-based speaker recognition system on various parameters used. Based on the above studies, the performance of the system can be improved by properly choosing the parameter values.

In evaluating the overall performance of the source feature-based speaker recognition system, the various parameter values choosed are as given below: For each speaker, 6 seconds of data is used for training as well as testing. High energy and 100 msec contiguous speech segments are considered. A normalized block of 80 samples of the LP residual, each taken with one sample shift is used as input vector to the network. Number of epochs used for training are fixed to 60. The performance and the computational time of the optimized system is evaluated and the result is as shown in the Table 5.9. For comparison, the performance of the spectral-based system is also given in the table.

From the Table 5.9, it is clear that the performance of the optimized source model is as good as that of the system model. Though the optimized source model takes 30 minutes for generating each speaker model when compared to 4 minutes taken by system model, it has to be noted that the performance of the source model is obtained by using only 6 sec of data, whereas that of the system model is obtained by using 60 sec of data.

5.6 SUMMARY

In this chapter, experiments conducted to study the effects of various parameter values on the performance of the source feature-based speaker recognition system were presented. Analysis of the poorly performing speakers showed that for optimal performance, the models have to be trained well and the training has to be speaker-specific. The study made on the effect of number of units in dimension compression hidden layer on the performance showed the robustness of the source feature-based system to variations in network structure. It is shown that higher the size of the input vector (block of residual samples), the better is the performance. The main issue discussed in this chapter is the time optimization study, where it was shown that even 6 seconds of data was enough to capture the speaker variability in terms of source characteristics.

Table 5.10: **Change in the confidence values of genuine and imposter speakers for silence thresholds of 0.25 and 0.5, for 20 speakers of Set-II.**

Speaker No.	Change in the scores of	
(set-II)↓	Genuine speaker	Imposter speaker (avg.)
1	0.000443	0.000228
2	0.001139	0.000493
3	0.001193	0.000634
4	0.000856	0.000223
5	0.00055	0.000472
6	0.000402	0.000215
7	0.00047	0.000068
8	0.001168	0.000692
9	0.000184	-0.000333
10	0.00086	0.000385
11	-0.000131	0.000039
12	0.001549	0.000380
13	0.001058	0.000455
14	0.000137	-0.000171
15	0.001175	0.000386
16	0.000485	0.000142
17	0.002676	0.001649
18	0.000155	0.000343
19	0.001112	0.000739
20	-0.000156	0.000142

CHAPTER 6

SUMMARY AND CONCLUSIONS

6.1 SUMMARY OF THE WORK

The objective of this work is to illustrate the effectiveness of source features for text-independent speaker recognition task. The motivation behind this work is the fact that, any signal is produced as a result of exciting a resonating system by a force which sets the resonating system into vibrations. This force is nothing but the source of the generated signal, and hence the generated signal should carry information pertaining to both source as well as the system.

Speaker recognition is a pattern recognition problem and its performance depends upon the effectiveness of the features used, and how well the speaker characteristics have been captured in the training phase. All the present day automatic speaker recognition systems capture speaker variability mostly in terms of spectral features alone. But spectral features capture speaker characteristics in terms of the vocal tract system information. The source characteristics, which might also carry significant speaker information has been neglected. In this work, we proposed source feature-based text-independent speaker recognition. Here the speaker variability is captured in terms of source characteristics like glottal excitation. AANNs have been used for capturing the speaker-specific source characteristics.

LP residual signal has been used for capturing the source characteristics. Here our hypothesis was that source characteristics might be present in higher (> 2) order correlations present among the speech samples. Since the second order correlation information, which is nothing but the spectral information has been removed in spectral analysis, the residual signal should contain the higher order correlation information.

Nonlinear techniques may be needed to capture this information. AANN, which is a FFNN, is known for its ability to capture the correlations present among the samples of the signal, when trained with samples of speech signal. This feature of AANNs has been explored for capturing the speaker-specific source information.

It has been shown that AANNs trained with samples of LP residual signal capture the correlation information present among the samples of the signal. The performance of the source feature-based speaker recognition system has been evaluated on a database consisting of 80 male speaker conversational telephone speech. By comparing the performance of the source feature-based system with vocal tract system (spectral feature based) information-based system, it was shown that the performance for both the features is good. The complementary nature of the source and system information has been demonstrated through this study. By adding the scores from the source and system models, it has been shown that overall improvement can be obtained by combining two kinds of information.

Through a careful analysis of the poorly performing speakers, the significance of the training phase was illustrated. It was shown that the performance of the system depends primarily on the training phase of the system. For better performance, speaker-specific training has to be used.

Since the parameters used in the initial study were not optimal, further studies have been made to refine the performance of the system with respect to various parameters. The study made on the effect of number of nodes in the compression hidden layer indicated the robustness of the source features towards the network structure. The study on the effect of input vector (block of residual samples) size on the performance showed that the performance is better when the input vector size is large (80 to 100 samples). But increasing the input vector size increase the network structure, which in turn increases the time for training and testing. To reduce the computation time, various time optimization studies have been made. Studies made on the amount of data showed that data as little as 6 sec is enough to capture speaker-specific information in the training phase. This emphasizes the fact that speaker variability is being captured in terms of correlations present among the samples of the LP residual signal and not in terms of distribution of some feature vectors. Through the study made on the effects of nature of speech data on the performance of the system, it was shown that high amplitude contiguous speech segments carry better correlation information and hence yield better performance.

By properly tuning the various parameter values, the source feature-based speaker recognition system can be made as effective as the vocal tract system (spectral) information-based speaker recognition system. But, a compromise has to be reached between the computation time and the performance of the system.

6.2 MAJOR CONTRIBUTIONS OF THE WORK

- AANN based text-independent speaker recognition using source features has been proposed, and the effectiveness of the source features in comparison with the spectral features has been illustrated.
- The complementary nature of the source information and vocal tract system information has been demonstrated. The scope for improvement in the performance of the speaker recognition system by suitably combining this two kinds of information was discussed.
- The role played by training phase on the performance of the speaker recognition system was discussed. It was showed that speaker-specific training has to be used for improving the performance.
- The significant contribution of this work is the study made on the amount of data required to capture the speaker variability. It has been shown that 6 sec of data is enough to capture the speaker variability in terms of source characteristics.

6.3 SCOPE FOR FUTURE WORK

- Computation time and the performance optimization can be achieved by using more sophisticated learning algorithms (like conjugate gradient descent algorithm) in place of BP algorithm.
- To understand the issues clearly, a through theoretical analysis is required.
- Suitable methods have to be derived for combining the source and vocal tract system (spectral) information more effectively.
- The same concept can be extended for speaker segmentation and speaker tracking tasks.
- Since only 4-6 sec of data is enough for generating the speaker models, multiple models can be generated for each speaker and multiple evidences can be combined for improving the performance of the system.
- The effectiveness of the source features against spectral features in noisy conditions has to be studied.

APPENDIX A

LINEAR PREDICTION ANALYSIS

In LP analysis of speech, an all-pole model is assumed for the system producing speech signal $s(n)$. A p^{th} order all-pole model assumes that sample value at time n can be approximated by linear combination of past p samples. i.e.,

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

If $\hat{s}(n)$ denotes the prediction made by the all-pole model then, the prediction error is given by,

$$e(n) = s(n) - \hat{s}(n) = s(n) - \sum_{k=1}^{p} a_k s(n-k) \tag{A.2}$$

This error is nothing but the LP residual of the given speech signal.

For a speech frame of size m samples, the mean square of prediction error over the whole frame is given by,

$$E = \sum_{m} e^2(m) = \sum_{m}[s(m) - \sum_{k=1}^{p} a_k s(m-k)]^2 \tag{A.3}$$

Optimal predictor coefficients will minimize this mean square error. At minimum value of E,

$$\frac{\partial E}{\partial \mathbf{a_k}} = 0 \ , \qquad k = 1, 2, ...p. \tag{A.4}$$

Differentiating Eqn A.3 and equating to zero we get,

$$\mathbf{R}\ \mathbf{a} = \mathbf{r} \tag{A.5}$$

where, $\mathbf{a} = [a_1\ a_2 \cdots a_p]^T$, $\mathbf{r} = [r(1)\ r(2) \cdots r(p)]^T$, and $\mathbf{R}$ is a Toeplitz symmetric autocorrelation matrix given by,

$$\mathbf{R} = \begin{bmatrix} r(0) & r(1) & \cdots & r(p-1) \\ r(1) & r(0) & \cdots & r(p-2) \\ \vdots & & \ddots & \vdots \\ r(p-1) & & \cdots & r(0) \end{bmatrix} \tag{A.6}$$

Eqns A.5 can be solved for prediction coefficients using Durbin's algorithm as follows:

$$E^{(0)} = r[0] \tag{A.7}$$

$$k_i = \frac{r[i] - \sum_{j=1}^{L-1} \alpha_j^{(i-1)} \cdot r[|i-j|]}{E^{(i-1)}} \qquad 1 \leq i \leq p \tag{A.8}$$

$$\alpha_i^i = k_i \tag{A.9}$$

$$\alpha_j^i = \alpha_j^{(i-1)} - k_i \cdot \alpha_{i-j}^{(i-1)} \tag{A.10}$$

$$E^{(i)} = (1 - k_i^2) \cdot E^{(i-1)} \tag{A.11}$$

The above set of equations are solved recursively for $i = 1, 2, ..., p$. The final solution is given by

$$a_m = \alpha_m^{(p)} \qquad 1 \leq m \leq p \tag{A.12}$$

where, a_m's are linear predictive coefficients (LPCs).

Cepstral coefficients can be extracted from the predictor coefficients using recursive algorithm as follows.

$$c_0 = \ln \sigma^2 \tag{A.13}$$

$$c_m = a_m + \sum_{k=1}^{m-1} \frac{k}{m} \cdot c_k \cdot a_{m-k} \qquad 1 \leq m \leq p \tag{A.14}$$

$$= \sum_{k=1}^{m-1} \frac{k}{m} \cdot c_k \cdot a_{m-k} \qquad m > p \tag{A.15}$$

APPENDIX B

NIST DATABASE

NIST 99 database is a standard database used for checking the performance of speaker recognition systems. Salient features of NIST database are:

1. It consists of 539 speakers, 230 male and 309 female.
2. Recording of the utterances was done over telephone channel in noisy environment.
3. Data consist of conversational speech segments.
4. The speech signal is sampled at 8 KHz.
5. Each speaker has two 1 minute utterances, collected over same channel but in two different sessions.

In the present work, a subset of 80 male speakers is used. In all the experiments two 1 minute utterances collected in two different sessions is used. A block size of 40 (samples of LP residual signal) and a block shift of 1 sample are used for capturing the source component of the speech production system.

BIBLIOGRAPHY

[1] B. S. Atal, "Automatic recognition of speakers from their voices," *Proc. IEEE*, vol. 64, pp. 460–475, Apr. 1976.

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

[3] D. O'Shaughnessy, "Speaker recognition," *IEEE ASSP Magazine*, pp. 4–17, 1986.

[4] 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.

[5] A. E. Rosenberg, "Automatic speaker verification: A review," *Proc. IEEE*, vol. 64, pp. 475–487, Apr. 1976.

[6] J. M. Naik, "Speaker verification: A tutorial," *IEEE Communications Magazine*, pp. 42–48, Jan. 1990.

[7] J. G. Daugman, "High confidence visual recognition of persons by a test of statistical independence," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 15, pp. 1148–1161, Nov. 1993.

[8] Rama Chellappa, Charles L. Wilson and Saad Sirohey, "Human and machine recognition of faces: A survey," *Proc. IEEE*, vol. 83, no. 5, pp. 705–740, May 1995.

[9] Ashok Samal and Prasana A. Iyengar, "Automatic recognition and analysis of human faces and facial expressions: A survey," *Pattern Recognition*, vol. 25, no. 1, pp. 65–77, 1992.

[10] K. Karu and A. K. Jain, "Fingerprint classification," *Pattern Recognition*, vol. 29, no. 3, pp. 389–404, 1996.

[11] M. Kawagoe and A. Tojo, "Fingerprint pattern classification," *Pattern Recognition*, vol. 17, no. 3, pp. 295–303, 1984.

[12] O. Tosiand H. Oyer, "Experiments on voice recognition," *J. Acoust. Soc. Amer.*, vol. 51, no. 6, pp. 2030–2043, 1972.

[13] Frank Devaert, "Recognizing emotion in speech," *Proceedings of Int. Conf. Spoken Language Processing*, Oct 1996.

[14] A. Sutherland and M. Jack, "Speaker verification," *Aspects of Speech Technology*, pp. 184–215, 1988.

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

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

[17] S. Furui, A. Sutherland, and M. Jack, "Recent advances in speaker recognition," *Pattern Recognition Letters*, vol. 18, pp. 859–872, 1997.

[18] J. P. Campbell, "Speaker recognition: A tutorial," *Proc. IEEE*, vol. 85, pp. 1436–1462, Sept. 1997.

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

[20] P. Satyanarayan, *Short segment analysis of speech for enhancement*. Ph. D dissertation, Indian Institute of Technology, Department of Electrical Engg., Madras, Feb 1999.

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

[22] NIST, "Speaker recognition workshop notebook," *Proc. NIST 2000 Speaker Recognition Workshop, University of Maryland, USA*, Jun 26-27 2000.

[23] D. A. Reynolds, "The effects of handset variability on speaker recognition performance: Experiments on switch board corpus," in *Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*, pp. 113–116, 1996.

[24] D. A. Reynolds and *et al.*, "The effects of telephone transmission degradations on speaker recognition performance," in *Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*, pp. 329–332, 1995.

[25] NIST, "Speaker recognition workshop notebook," *Proc. NIST 1999 Speaker Recognition Workshop, University of Maryland, USA*, Jun 3-4 1999.

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

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

[28] T. N. T. Matsui and S. Furui, "Robust methods to update model and a-priori threshold in speaker verification," vol. 1, pp. 97–100, May 1996.

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

[30] M. Sambur, "Selection of acoustic features for speaker identification," *IEEE Trans. on ASSP*, vol. 23, no. 2, pp. 176–182, 1975.

[31] S. Pruzansky and M. V. Mathews, "Talker-recognition procedure based on analysis of variance," *J. Acoust. Soc. Amer.*, vol. 36, no. 11, pp. 2041–2047, 1964.

[32] W. S. M. Jr., "Two statistical feature evaluation techniques applied to speaker recognition," *IEEE Trans. Comput.*, vol. 20, pp. 979–987, 1971.

[33] N. Ney and R. Gierloff, "Speaker recognition using a feature weighting technique," in *Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*, pp. 1645–1648, 1982.

[34] Y. Tohkura, "A weighted cepstral distance measure for speech recognition," in *Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*, pp. 761–764, 1986.

[35] Y. Tohkura, "Weighted cpestral distance measure for speech recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 35, no. 10, pp. 1414–1422, 1987.

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

[37] L. L. J. He and G. Pahm, "A discriminative training algorithm for gaussian mixture speaker models," *Proceedings of EUROSPEECH'97*, vol. 2, pp. 959–962, 1997.

[38] 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.

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

[40] L. P. Heck and M. Weintraub, "Handset–dependent background models for robust text–independent speaker recognition," in *Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*, (Munich, Germany), Apr 1997.

[41] H. Gish, "Robust discrimination in automatic speaker identification," in *Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*, pp. 289–292, 1990.

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

[43] M. Forsyth, "Discriminating observation probability (dop) hmm for speaker verification," *Speech Comm.*, vol. 17, pp. 117–129, 1995.

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

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

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

[47] J. P. Eatock and J. S. Mason, "Automatically focusing on good discriminating speech segments in speaker recognition," in *Proceedings of Int. Conf. Spoken Language Processing*, (Kobe, Japan), pp. 133–136, Nov. 1990.

[48] S. Vela and H. A. Murthy, "Speaker identification- a new model based on statistical similarity," *Int. Conf. on computational linguistics, speech and document processing*, 1998.

[49] Y. Gong and J. Haton, "Text-independent speaker recognition by tragectory space comparision," in *Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*, vol. 1, (New Mexico), pp. 285–288, Apr. 1990.

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

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

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

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

[54] M. B. I. Booth and B. Watson, "Enhancement to dtw and vq decision algorithms for speaker recognition," *Speech Comm.*, vol. 13, pp. 427–433, 1993.

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

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

[57] K. I. Funahoshi, "On the approximate realization of contineous mapping by neural networks," *Neural Networks*, vol. 2, pp. 183–192, 1989.

[58] K. Hornik, "Approximate capability of multilayer feedforward neural networks," *Neural Networks*, vol. 4, pp. 251–257, 1991.

[59] B. Yegnanarayana, *Artificial Neural Networks.* New Delhi: Prentice-Hall of India, 1999.

[60] J. Sietsma and R. J. F. Dow, "Creating neural networks that generalize," *Neural Networks*, vol. 4, pp. 67–79, 1991.

[61] R. P. Lippmann, "An introduction to computing with neural nets," *IEEE ASSP Magazine*, vol. 4, pp. 4–22, Apr. 1989.

[62] F. B. P. Gallinari, S. Thiria and F. F. soulie, "On the relation between discriminant analysis and multilayer perceptron," *Neural Networks*, vol. 4, pp. 349–360, 1991.

[63] M. T. L. C. Y. T. S. S. Y. I. C. Jou, S. L. Lee and Y. O. Tsay, "A neural network based speaker verification system," in *Proceedings of Int. Conf. Spoken Language Processing*, pp. 1273–1276, 1990.

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

[65] T. Z. H. Yin, "Speaker recognition using static and dynamic cepstral fetures by a learning neural network," in *Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*, vol. 2, pp. 1277–1280, 1990.

[66] M. Shajith Ikbal, *Autoassociative Neural Network Models for Speaker Verification.* MS dissertation, Indian Institute of Technology, Department of Computer Science and Engg., Madras, May 1999.

[67] M. Smith, *Neural networks for statistical modeling.* New York: Van Nastrand Reinhold, 1993.

[68] H. Bourlard and Y. Kamp, "Auto-association by multilayer perceptrons and singular value decomposition," *Biol. Cybernet.*, vol. 59, pp. 291–294, 1988.

[69] Hemant Misra, *Development of a Mapping Feature for Speaker Recognition.* MS dissertation, Indian Institute of Technology, Department of Electrical Engg., Madras, May 1999.

[70] 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.

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

[72] J. G. Prokis and D. G. Manolakis, *Digital signal processing: principles, algorithms, and applications.* New Delhi: Prentice-Hall of India, 1997.

[73] A. V. Oppenheim and A. S. Willsky, *Signals and systems.* New Delhi: Prentice-Hall of India, 1998.

[74] D. B. Fry, *The physics of speech.* New york: Cambridge University Press, 1979.

[75] A. V. Oppenheim and R. W. Schafer, *Digital signal processing.* Englewood Cliffs, New Jersey: Prentice-Hall of India, 1975.

[76] S. M. Kay, *Modern spectral Estimation: Theory and Applications.* Englewood Cliffs, New Jersey: Prentice-Hall of India, 1988.

[77] 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.

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

[79] J. G. P. J. R. Deller and J. N. L. Hansen, *Discrete-time processing of speech signals.* New York: Macmillan, 1993.

[80] K. T. Assaleh and R. J. Mammone, “New LP–derived features for speaker identification,” *IEEE Trans. Speech, Audio Processing*, vol. 2, pp. 630–638, Oct. 1994.

[81] L. Shinan and A. Almeida, “The effects of voice disguise upon formant transition,” in *Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*, pp. 885–888, 1986.

[82] T. Takagi and H. Kuwabara, “Contributions of pitch, formant frequency, and bandwidth to the perception of voice-personality,” in *Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*, (Tokyo), pp. 889–892, 1986.

[83] J. H. Li Lui and G. Palm, “A comparision of human and machine in speaker recognition,” *Proceedings of EUROSPEECH'97*, 1997.

[84] S. P. Kishore and B. Yegnanarayana, “Speaker verification: Minimizing the channel effects using autoassociative neural network models,” in *Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*, (Istanbul), pp. 1101–1104, 2000.

[85] P. Thevenaz and H. Hugli, “Usefulness of lpc-residue in text-independent speaker verification,” *Speech Comm.*, vol. 17, pp. 145–157, 1995.

[86] J. H. Li Lui and G. Palm, “On the use of features from prediction residual signals in speaker recognition,” *Proceedings of EUROSPEECH'97*, pp. 313–316, 1997.

[87] A. V. N. S. Anjani, *Analysis of autoassociative neural network models for processing degraded speech.* MS dissertation, Indian Institute of Technology, Department of Computer Science and Engg., Madras, June 2000.

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

[89] B. Yegnanarayana, *Artificial neural networks for pattern recognition*, vol. 19. Sadhana, Apr. 1994.

[90] J. M. Zurada, *Introduction to artificial neural network systems.* Singapore: Information access and distribution, 1992.

[91] D. E. Rumelhart, P. Smolensky, J. L. McClelland, and G. E. Hinton, “Schemata and sequential thought processes in PDP models,” in *Parallel Distributed Processing: Explorations in the Microstructure of cognition* (J. L. McClelland, D. E. Rumelhart and PDP Research Group, ed.), vol. 2, ch. 14, Cambridge: MIT Press, 1986.

[92] S. P. Kishore and B. Yegnanarayana, "Speaker verification using autoassociative neural network models," *IEEE Trans. Acoust., Speech, Signal Processing (communicated).*

[93] Hemant Misra, M. Shajith Ikbal, and B. Yegnanarayana, "Spectral mapping as a feature for speaker recognition," in *National Conference on Communications(NCC)*, (IIT, Kharagpur), pp. 151–156, Jan 29-31 1999.

[94] L. W. Ke Chen and H. Chi, "Methods of combining multiple classifiers with different features and their applications to text-independent speaker identification," *Int. J. Pattern Recognition and Artificial Intelligence*, vol. 11, no. 3, pp. 1–18, 1997.

[95] J. J. Hull and S. N. Srihari, "Decision combination in multiple classifier systems," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 16, pp. 66–75, Jan. 1994.

LIST OF PUBLICATIONS

PRESENTATION IN CONFERENCES

1. B. Yegnanarayana, K. Sharat Reddy and S. P. Kishore, "Source and System Features for Speaker Recognition Using AANN Models", Proc. IEEE int. conf. Acoust., Speech, Signal Process., Vol 1, 2001.