TEXT-DEPENDENT SPEAKER VERIFICATION USING SEGMENTAL, SUPRASEGMENTAL AND SOURCE FEATURES

A THESIS

submitted by

Jinu Mariam Zachariah

for the award of the degree

of

MASTER OF SCIENCE

(by Research)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY MADRAS

MARCH 2002

THESIS CERTIFICATE

This is to certify that the thesis entitled **Text-Dependent Speaker Verification Using Segmental, Suprasegmental And Source Features** submitted by **Jinu Mariam Zachariah** 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 her under my 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-600 036

Date:

Prof. B. Yegnanarayana

Dept. of Computer Science and Engg.

ACKNOWLEDGEMENTS

I express my deepest gratitude to Prof. B. Yegnanarayana for the valuable guidance provided to me throughout this research work. I am extremely grateful to him for his constant motivation and encouragement. Sir, my heartfelt thanks to you for the attention, support and patience shown to me. The discussions with you and your constructive criticisms have helped me learn the way to carry out a disciplined and organised research and improve my writing abilities and presentation skills.

I am grateful to Dr. C. Chandrasekhar for the guidance and support given to me in this work. I thank Prof. S. Raman, the Head of Department of Computer Science and Engineering, for providing good facilities and fostering an excellent research environment. I also thank Prof. C. Pandu Rangan, the former Head of the Department of Computer Science and Engineering.

I thank Dr. Hema A. Murthy and Dr. Sukhendu Das for the several fruitful discussions we had, especially during the short presentations during Monday meetings. I acknowledge the help received from Mr. S. Rajendran during the course of this work.

I express my heartfelt thanks to Prof. R. Kalyana Krishnan for the support and advice given to me on several occasions. Sir, I have learnt a lot from you.

I am privileged to have a wonderful research group around me. I thank my seniors Mathew, Jyotsna, Siva, Anjani, Kishore, Sharat, Prasada Reddy, Kiran[P], and Devaraj for the help given to me on several occasions. My special thanks to Jyotsna and Sharat for the several discussions and fun we had during the course of my work. I specially thank all my labmates, Prasanna, Surya, Nayeem, Sreenivasa Rao, Palanivel, Kiran, Ramesh, Vinod, Venkatesh, Anil, Gupta, Anvita, Dhananjay, Guruprasad, Gaurav, Shivnaga, Leena, Ram, Siva, Vishwam, and Meera.

I thank Prasanna and Nirmala for the care shown and the help given to me during my stay in IIT. You were supportive and caring at times that were extremely difficult for me. Your help, described in words, can never fully convey the comfort that it provided. Thanks to Prasanna for the patience with which he has gone through my thesis. Thanks to Surya for going through the thesis and suggesting improvements and also for the discussions we had, that have helped me lot.

I am thankful to the good circle of friends I had in IIT, which has made my life here a memorable one. My heartfelt thanks to Anitha, Rajani, Jaison, Renny, Vipin and Manoj. I also thank Sudar, Uma, Valli, Rajashree, Vijayalakshmi, Priya, Aparna, Chandana and Sabia, who were my friends at Sarayu hostel.

I thank Anitha Joseph for introducing me to IIT and this research programme.

I shall always remain indebted to my parents, P. V. Zachariah and Jessy Zachariah, my brothers, Jibu and Jeen and my sister Jinsu for giving me the moral support that I needed, without which this would not have been possible. Thanks to Ajith, though miles apart, he has poured in lots of words of care and love, which has relaxed me during the final stages of my work.

Above all, I thank the Lord Almighty whose unwavering grace has been my companion through the thick and thin.

Jinu Mariam Zachariah

ABSTRACT

KEYWORDS: *automatic speaker verification; dynamic time warping; intonation; duration; source characteristics*

Automatic speaker recognition is the task of recognizing a person's voice based on the information obtained from the speech signal. The uniqueness in the voice of a speaker is due to several factors such as the shape and size of the vocal tract system, the dynamics of the articulators, the characteristics of the excitation source, the rate of vibration of the vocal folds, the accent imposed by the speaker, and the speaking rate. These features are reflected in the speech signal, and therefore can be used for speaker recognition. Speaker recognition can be classified into two categories, based on the nature of the task, as speaker identification and speaker verification. Speaker identification is the process of determining to which of the registered speaker the test utterance belongs, and speaker verification is the process of accepting or rejecting the identity claim of the speaker. Depending on the text to be spoken, speaker recognition can be performed in a text-dependent or text-independent mode. The voice individuality of a speaker associated with the speech sound unit can be exploited directly in the case of text-dependent speaker verification. It requires less amount of training data when compared to a text-independent system. In this thesis we focus on text-dependent speaker verification.

The most important task in a text-dependent speaker verification system is the process of detecting the begin and end of the speech utterance. Present day systems mostly rely on amplitude of the speech signal. This technique fails when the speech data is noisy. Hence we have explored an alternate method which uses the events of the speech signal for marking the endpoints. The basic system exploits the uniqueness in the shape and size of the vocal tract system for recognizing the speaker. This information, which is reflected in the short-term spectral features, is extracted from the reference and test utterances and they are matched using the technique of dynamic time warping. It is interesting to note that human beings recognize speakers mostly from the prosodic features such as intonation and duration, and source characteristics such as the glottal vibrations. Since these features have more intra-speaker variability compared to spectral features, they are difficult to generalize. Hence not much work has gone in this direction to use these features for speaker verification. Though the prosodic and source features have large variability, they are more robust to channel/handset variations, when compared to the spectral features. In this thesis, we have explored methods to extract duration and pitch information, and studied their effectiveness for speaker verification using speech data collected through microphone as well as telephone. Duration information is extracted using the characteristics of the dynamic time warping path. Pitch information is obtained by matching the pitch frequency of similar frames of the reference and test utterances. The characteristics of the excitation source of the speech signal is present in the correlation among the samples of the linear prediction residual signal. Studies have shown that autoassociative neural network can be used to model this information. We have explored this proposed approach for text-dependent speaker verification. Though the prosodic and source features give reasonably good performance, it is not adequate for practical applications. Since these features form an independent source of information, they may be used to complement the decision based on spectral features in many cases. Hence by combining the evidence from all these features we can get better performance than a classifier which is based on any single feature alone. In this work we have attempted to combine the evidence from the spectral, prosodic (duration and intonation) and source features to improve the performance of a text-dependent speaker verification system.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

ANN	- Artificial Neural Network
AANN	- Autoassociative Neural Network
DFT	- Discrete Fourier Transform
DTW	- Dynamic Time Warping
EER	- Equal Error Rate
FA	- False Acceptance
FR	- False Rejection
GMM	- Gaussian Mixture Model
HMM	- Hidden Markov Model
IDFT	- Inverse Discrete Fourier Transform
LP	- Linear Prediction
LPC	- Linear Prediction Coefficients
VQ	- Vector Quantization

CHAPTER 1

INTRODUCTION

1.1 INTRODUCTION TO AUTOMATIC SPEAKER RECOGNITION

Automatic speaker recognition is the task of recognizing a person's voice by a machine from the information obtained from his/her speech signal. Speech signal contains information regarding the message conveyed, the speaker's identity, the language used for communication, the emotional state of the speaker, and the gender of the speaker. Speaker recognition task involves extraction of the unique speaker-specific features from the speech signal to recognize the speaker. The uniqueness in the voice of a speaker is due to several factors such as the shape and size of the vocal tract, the dynamics of the articulators, the rate of vibration of the vocal folds, the accent imposed by the speaker and the speaking rate [1]. These factors are reflected in the speech signal, and hence are useful for speaker recognition. Speaker recognition has become an important biometric technique. It has several applications such as control of confidential information, remote access to computers, forensic applications, voice dialing, banking transactions over the telephone, tele-shopping, voice mail and database access services.

Speaker recognition can be divided into three stages [2–7] :

1. Feature extraction
2. Pattern classification
3. Decision making

In the feature extraction stage, the desired set of features which are unique to a speaker are extracted from the speech signal. In the pattern classification stage, a score value which signifies the similarity between the reference and the test utterance is computed. Based on the score value obtained, a decision is made in the decision making stage.

Speaker recognition has two operational phases: Training phase and testing phase. A block diagram of the training phase is given in Fig.1.1. In the training phase, the speaker gives a reference speech data, from which the desired set of features are extracted. These features are then stored as the reference data for that speaker. In the testing phase, the speaker gives test speech data for which the same set of features are computed. The pattern comparison stage compares the reference pattern and the test pattern and generates a similarity score. A recognition decision is made by the decision making stage. The block diagram of the testing phase is given in Fig.1.2.

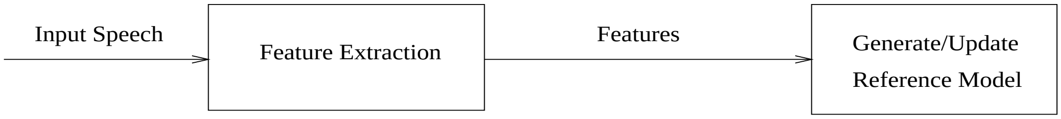


Fig. 1.1: Block diagram for training phase of speaker recognition

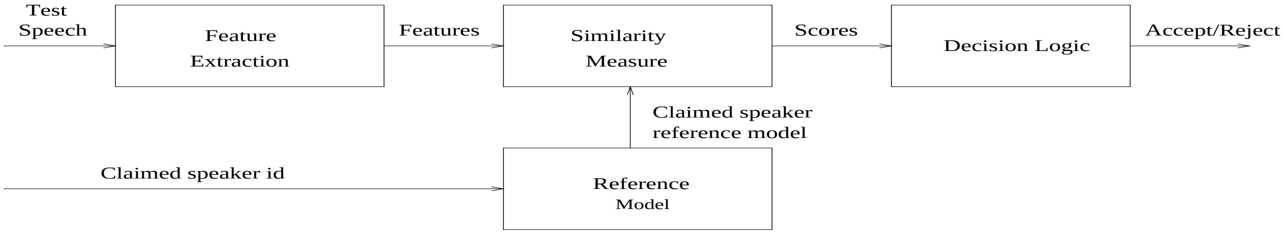


Fig. 1.2: Block diagram for testing phase of speaker recognition

1.2 CLASSIFICATION OF SPEAKER RECOGNITION SYSTEM

Speaker recognition can be divided into speaker identification and speaker verification [8].

- Speaker identification is the process of determining to which of the registered speaker, a given utterance belongs [9]. The block diagram of a speaker identification system is shown in Fig.1.3.
- Speaker verification is the process of accepting or rejecting the identity claim of a given speaker [10]. Fig.1.4 shows the schematic representation of a speaker verification system.

Speaker identification can be a closed set identification or an open set identification. In the closed set identification, it is assumed that the test utterance belongs to one of the N enrolled speakers (N decision levels). In the case of open set identification, there is an additional decision level to determine whether the test utterance was uttered by an enrolled speaker or not, that is, there are $N+1$ decision levels. Speaker verification is an open set problem. The fundamental difference between identification and verification is in the number of decision alternatives. In identification, the number of decision alternatives is equal to the size of the population, whereas in verification there are only two decision alternatives, accept and reject. Hence, the performance of a speaker identification system decreases as the size of the population increases, whereas the performance of a speaker verification system is independent of the size of the population.

Based on the text to be spoken, speaker recognition methods can also be divided into text-dependent speaker recognition and text-independent speaker recognition. Text-dependent speaker recognition systems require the speaker to provide utterances of the same text for both training and testing, whereas text-independent speaker recognition

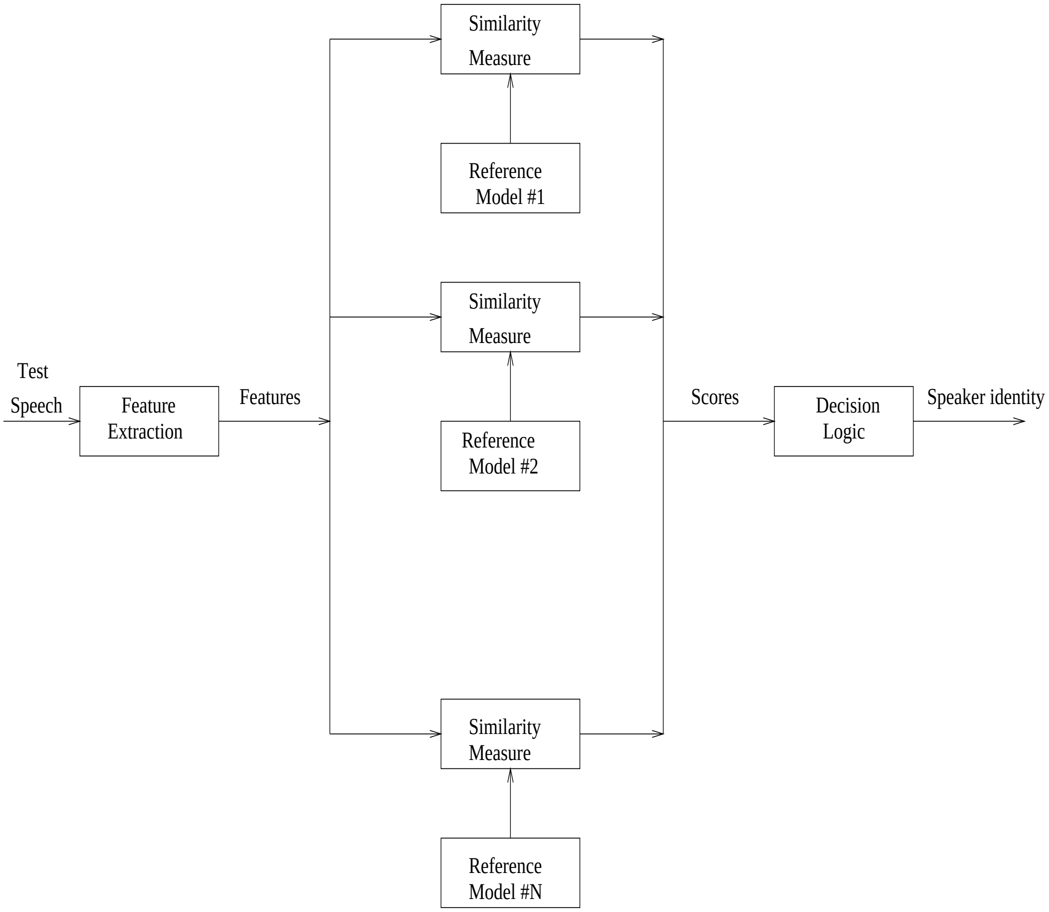


Fig. 1.3: Block diagram of a speaker identification system

does not depend on the text spoken. In this research work, we focus on text-dependent speaker verification systems.

1.3 COMPARISON OF TEXT-DEPENDENT AND TEXT-INDEPENDENT MODES OF SPEAKER VERIFICATION

Text-dependent methods are usually based on template matching techniques in which the time axes of the test pattern and the reference template of the target speaker are aligned, and the distance between them accumulated over the whole utterance is calculated. Since this method can exploit directly the individuality of speaker's voice

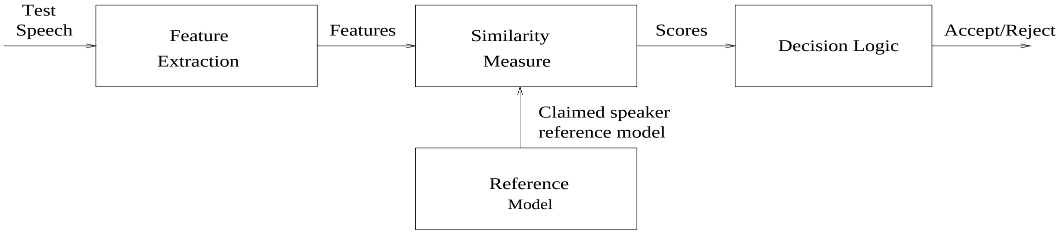


Fig. 1.4: Block diagram of a speaker verification system

associated with each phoneme or syllable, it generally achieves higher recognition performance than the text-independent methods [11]. Since the text-dependent speaker verification can be accomplished using template matching techniques, it requires less amount of data. In contrast, the text-independent systems requires large amount of data since they rely on statistical or neural network modeling techniques [12–14] . In the case of a text-dependent speaker verification task, it is easier to exploit the prosodic features or features which reflect the speaking style of an individual. This is because the prosodic features show less intra-speaker variability for a given fixed text. Both the text-dependent and the text-independent systems have a serious problem. These systems can be easily defeated by an imposter by playing back the recorded voice of a registered speaker. In this regard the text-dependent system has advantage over the text-independent system, since in the case of a text-dependent speaker verification system, the user can be prompted to speak a sentence selected at random, which cannot be predicted before hand.

1.4 ISSUES IN TEXT-DEPENDENT SPEAKER VERIFICATION SYSTEM

A challenging task in any biometric application is the development of a reliable practical system. The issues involved in the development of a reliable system for text-dependent speaker verification can be viewed from three different angles. The text-dependent speaker verification, like other speaker verification systems, involves three entities in the process: man, machine, and the environment. Speech signal produced by man in a given environment is the input to the machine (which includes the signal capturing device and the speaker recognition algorithm), as shown in Fig.1.5. The issues in a practical system are those issues that deal with each of these entities. These issues include mainly the intra-speaker variations which are reflected in the speech produced, the variation in the quality of speech due to difference in the signal capturing device, the issues at various stages of the speaker recognition algorithm, and the variation in the quality of speech signal due to environmental conditions.

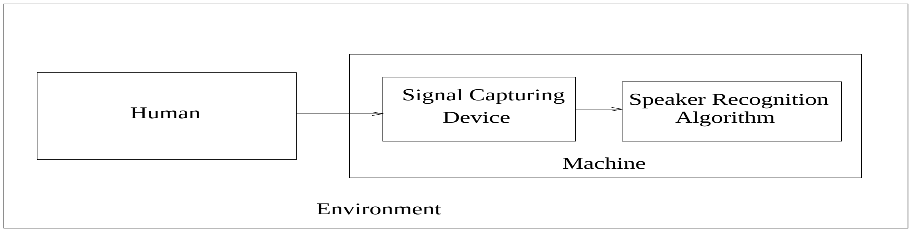


Fig. 1.5: The entities involved in a speaker verification task

A severe problem faced by almost all existing speaker verification systems is the variation introduced by the speaker himself. Intra-speaker variation of the *genuine* speaker such as those due to aging, stress, emotional state of the speaker, health, Lombard effect (effect of articulation variability as the speaker attempts to communicate in a noisy environment) [15], variation in the speaking style are all manifested in the speech data given by the speaker. An *imposter* speaker may try to mimic another registered speaker by giving the recorded speech data of the target speaker. Hence, systems for real-world applications must be robust to the type of intra-speaker variation introduced by a genuine speaker, and should reject attempts by an imposter speaker.

The recording devices used for capturing the signal, range from high quality fixed-mount microphones to low cost telephone handsets. Speaker phones and wireless handsets are also becoming popular devices. Different handsets produce speech signal with different spectral properties. Even for the same handset, the channel may be different. Channels with different characteristics introduce varied nature of degradation in the spectral and noise characteristics of the speech signal [6]. In real world application, there is also a possibility of mismatch between the type of handset or channel used for training and testing (known as mismatched condition). The direction of the incoming speech signal, and also the distance between the speaker and the microphone, as in hands-free signal capturing devices, play a very crucial role in the quality of the speech signal captured.

A speaker recognition system consists of preprocessing, feature extraction, pattern classification, and decision making stages. An important task in the preprocessing stage is the process of location of the begin and end of the speech signal. Improper marking of begin or end may result in wrong alignment of the reference and the test patterns. In the present systems, mostly amplitude of the speech signal is used for locating the begin and end of an utterance. But this technique fails in cases where the speech data is noisy, or where high amplitude regions such as glitches or cough or breathing sounds are present. Hence, it is necessary to come up with a technique other than a simple amplitude threshold method for detecting the end points.

The selection of the speaker-specific features to be extracted is an important factor. It should have low intra-speaker variability, and very high inter-speaker variability. The features should be consistent irrespective of the quality of the speech data. Proper representation of the features is a key issue in the feature extraction. It should be such that the speaker-specific information is retained properly.

The classification algorithm to be used plays a crucial role in the performance of the system. Many pattern recognition methods do exist such as the template matching technique, statistical approach, and neural network based techniques. Suitable techniques should be used depending on the type of the selected feature.

The environment conditions in which the recording is done may have background noise or may have other people talking in the background. In the case of cellular devices, which are mostly used outdoor, the environment is highly noisy. Speech data recorded in such conditions have a low Signal-to-Noise Ratio (SNR). In such conditions, it is important to use features which are robust to degradations due to environment.

1.5 ISSUES ADDRESSED IN THIS THESIS

In the previous section, a brief description of the issues that arise in a text-dependent speaker verification system was given. In this section, the issues addressed in this thesis are listed.

The basic system for text-dependent speaker verification performs a template matching between short-term spectral features of the reference and test utterances. The first and the foremost task in a text-dependent speaker verification system is the process of location of the begin and end of the speech utterance. Currently, systems use techniques based on the amplitude of the speech signal for locating the end points of an utterance. These techniques perform poorly when the speech data is degraded by channel or environment noise or when there are high amplitude non-speech glitches or cough or breathing sounds. In this work, we have explored an alternate way for locating the end points of the speech utterance.

The speaker-specific characteristics in the speech signal are present at different levels, namely, the segmental level and the suprasegmental level. The basic system for text-dependent speaker verification makes use of the segmental features. Segmental features or short-term spectral features are those which are extracted from 10-30 msec of the speech signal, and they reflect the shape and size of the speech production system and their dynamics. The short-term spectral features are sensitive to frequency characteristics of channel, and also to the variation of the noise characteristics of the environment or the channel. Hence, we need to explore other features which remain relatively consistent, and which can be reliably extracted even from noisy or degraded speech data. It has been observed that the basic system for text-dependent speaker verification ignores the speaker-specific features which are present at the suprasegmental level, such as pitch and duration. In fact the features which have been ignored in the basic system are those features which turn out to be robust against channel and handset variations. These features can be used as additional evidence to the information obtained from the spectral feature for making a decision. In this work, we have attempted to incorporate pitch and duration information from the speech signal for the task of text-dependent speaker verification.

In addition to the suprasegmental features, the source characteristics of the speech signal are also generally ignored in the basic system. By source characteristics we mean the characteristic of the excitation source of the vocal tract system for speech production. Studies have shown that the source characteristics contains speaker-specific information [16]. We study the effectiveness of the methods proposed in [16] for text-dependent speaker verification task.

It is known that human beings implicitly integrate several features for speaker verification. Since the segmental features, suprasegmental features and the source features of the speech signal are having different physical significance, the evidence produced by classification based on these features are likely to produce complementary results. Literature says that it is indeed possible to combine the evidence produced by complementary classifiers [17]. Hence we attempt to combine the evidence present at various levels for improving the robustness of a text-dependent speaker verification system.

1.6 ORGANIZATION OF THE THESIS

A brief introduction to automatic speaker recognition has been given in this chapter. The issues that arise in a text-dependent speaker verification system and the issues addressed in this thesis have been discussed. The rest of the thesis is organized as follows:

In **chapter 2** techniques used for speaker verification are reviewed. The chapter deals mainly with different types of features and classification schemes.

In **chapter 3** a description of the existing basic system is given. It uses segmental features. An alternative method for locating the end-points of an utterance is described here. Performance evaluation of the basic system is carried out on a speech database collected through microphone and telephone. Details of the speech data used in this study is given in this chapter.

In **chapter 4** methods for capturing the suprasegmental features (pitch and duration) are given. The effectiveness of these features for text-dependent speaker verification discussed here.

In **chapter 5** the effectiveness of source information for text-dependent speaker verification is discussed. Analysis of the source features is carried out for microphone and telephone speech data.

In **chapter 6** a method for combining the evidence from the short-term spectral features, suprasegmental features, and the source features is described. A hierarchical approach is used to combine the evidence present in the features.

Chapter 7 gives a summary of the thesis work with conclusions from the studies made in this thesis.

CHAPTER 2

REVIEW OF APPROACHES USED FOR AUTOMATIC SPEAKER RECOGNITION

2.1 INTRODUCTION

Speaker recognition by humans is remarkable, but it is appreciated better when one attempts to make the machines perform a similar task. The basic principles of automatic speaker recognition have been discussed in the previous chapter. In this chapter, we present an overview of the techniques used in the literature for automatic speaker recognition. A speaker recognition system consists of two parts: Measurement and classification [18]. In the first part, a number of parameters are abstracted from the test speech data. These parameters characterize the pattern. The resulting set of parameters in turn act as input to a classification scheme, which compares the parameters with stored information or reference patterns, to arrive at a decision for the class membership of the test pattern.

This chapter is organized as follows: Section 2.2 gives a description of the ideal characteristics required for the features. The speaker-specific features that have been used for speaker recognition are described here. The pattern classification techniques which have been used in the literature are discussed in Section 2.3. Section 2.4 discusses attempts to combine the evidence from classifiers that use different features, or different modeling techniques. Section 2.5 gives a summary of the chapter.

2.2 FEATURES FOR SPEAKER RECOGNITION

The individuality in a speaker's voice is due to anatomical and learned differences [3]. Anatomical differences are the result of the variations in the sizes and shapes of the components of the vocal tract, namely, larynx, pharynx, vocal folds, tongue, teeth and the oral and nasal cavities. These anatomical differences lead to differences in the resonant or formant frequencies, bandwidth and the fundamental frequency. Learned differences are the result of differences in the patterns of coordinated neural commands to the individual articulators. Such differences gives rise to variations in the dynamics of the vocal tract, such as the rate of formant transitions and co articulation effects. The learned characteristics are also reflected in stress, accent, and speaking rate. The features and their representations are discussed briefly here.

2.2.1 Desirable characteristics of the features

One of the important steps in achieving successful speaker recognition is the selection of features capable of efficiently representing the speaker-dependent information in speech. A desirable set of features for speaker recognition are [18]:

- The features should vary as much as possible among speakers but be as consistent as possible for each speaker.
- The features should be stable over time, and should not be affected by the speaker's health.
- They should be easily measurable.
- They should occur naturally and frequently in normal speech.
- The features should not be affected by reasonable background noise, and should not be dependent on specific transmission characteristics.
- They should not be susceptible to mimicry

In respect of evaluating the features, a good measure of effectiveness is the $F - ratio$, which is defined as the ratio of the variance of the speaker means to the average inter-speaker variance [18–21]. For speaker recognition, a good measure is to have a high $F - ratio$. In such a case, the distribution of the features of an individual will be narrower, and the distributions of the speakers are widely separated from each other.

2.2.2 Short-term spectral features

Short-term spectral features are known to contain the information regarding the shape and size of the vocal tract. Several representations of the speech spectra such as Linear Prediction Coefficients (LPC) [22,23], orthogonal LP coefficients [24], autocorrelation coefficients, LPC-derived cepstral coefficients [25], Mel-warped cepstral coefficients [2], etc have been examined. We discuss two commonly used representations of the spectrum: LP cepstrum and mel-warped cepstrum.

2.2.2.1 LP cepstrum

The cepstrum, which is the Fourier transform of the log spectrum, is commonly used in speaker recognition applications because of its ability to capture the formant structure and the spectral tilt [25]. Atal compared the effectiveness of LP coefficients, autocorrelation coefficients and LP cepstral coefficient for speaker recognition [3]. It was reported that the LP cepstrum feature set was the best feature when used with Mahalanobis distance measure. Furui found LP cepstra to be better than log area ratio representation [10]. The reason behind this is that the spectral envelope reconstructed from the truncated set of cepstral coefficients is smoother than that reconstructed from the LP coefficients, and therefore provides a stable representation from one repetition to another of a particular speaker's utterance. It was suggested in [26] that the process of cepstral mean subtraction (subtracting the cepstral mean from each cepstral vector) helps in reducing the effects of the transmission channel.

2.2.2.2 Mel-warped cepstrum

The mel warping transforms the frequency scale to place less emphasis on high frequencies. It is based on the nonlinear human perception of the frequency of sounds. For higher frequencies, resolving power of the ear is less. The mel scale reflects this by using nonlinear warping of the frequency scale, i.e., by reducing the frequency resolution as frequency increases. Since the mel scale emphasizes perceptually important aspects of the speech signal, Mel Frequency Cepstral Coefficients (MFCC) are commonly used in speaker recognition [27, 28].

2.2.2.3 Other forms of representation

Line Spectral Pairs (LSP) is another feature that has been suggested recently [29]. Spectral mapping was found to be an important feature for text-independent speaker recognition [30–32]. The lower order cepstral coefficients contain the gross spectral information (mostly speech), whereas the higher order LP cepstral coefficients contain finer spectral information (speech and speaker information). A feedforward neural network is used to map a LP cepstral vector derived from a low order LP analysis to an LP cepstral vector derived from a higher order LP analysis. The neural network learns the speaker-specific mapping during training. RASTA (RelAtive SpecTra) processing with linear prediction has been used to arrive at a new feature called RASTA-PLP (RelAtive SpecTra Perceptual Linear Prediction), which has been used successfully for speaker recognition [33, 34]. The use of delta/differenced cepstra as a feature set is found in automatic speech recognition and automatic speaker verification [35, 36]. The differenced cepstra is a first order approximation to the first order differentiation of the cepstra. The motivation behind using differenced cepstra is to capture the transitional nature of the spectrum [10], which is useful to represent the dynamics of the articulators.

2.2.3 Fundamental frequency and intonation

The fundamental frequency (F_0) of the speech signal is the average rate of the vibration of the vocal folds, whereas intonation is the variation of the fundamental frequency as a function of time [37]. In early speaker recognition systems, pitch and gain were used [1, 5, 38–42]. Statistics of the frame-level pitch has been recently used in speaker recognition system with good results [43]. It has been shown by Atal that the temporal variations of pitch is indeed a speaker-specific characteristic [44]. In order to capture the local dynamics of the intonation pattern that characterize an individual's speaking style, the speaker's F_0 movements were modeled by fitting a piecewise linear model [45]. Parameters of this model were then used as statistical features for speaker verification. Prosodic features were used for speaker recognition in Hindi in a recent study [46–48]. The raw pitch frequencies at certain critical regions in the pitch contour were used as a feature. Feedforward and Adaptive Resonance Theory (ART) network models were used in this study. It was found that the performance with the feedforward neural network is better than with ART network system. In [49], the pitch frequency at the syllable nuclei and the duration of the words were used as components of feature for text-dependent speaker verification. Changes in emotion and health of the speaker cause large intra-speaker variation in the pitch. Moreover, the difficulty in explicit measurements of the speaker-specific features of the pitch contour has severely limited the use of pitch. The advantage of suprasegmental features, such as intonation, over short-term spectral features is that these features are robust against channel variations, and it is difficult to mimic the entire intonation pattern characteristics of a person [44].

2.2.4 Speaking rate

The rate at which a person speaks is an important speaker-specific feature. The duration of the entire utterance in normalized form was used in the study made by Atal [44]. It has been shown in [48, 49] that the duration of words can be used for text-dependent speaker verification. Word Boundary Hypothesis for Hindi language has been used for hypothesizing the boundaries of the words in an utterance [50].

2.3 PATTERN MATCHING

Pattern matching in speaker verification involves computing a match score, which is a measure of the similarity of the input feature vectors to some model. This section discusses briefly three different categories of pattern matching techniques:

- Template models
- Stochastic or probabilistic models
- Artificial neural network models

2.3.1 Template models

In template models, the pattern matching is deterministic. The given test pattern is assumed to be an imperfect replica of the template. A distance (D) between the test pattern and the template is used to measure the similarity between them [2]. The most common techniques are based on dynamic time warping and vector quantization.

2.3.1.1 Dynamic time warping

Dynamic Time Warping is a method to compensate for the variability in speaking rate in template-based system. It was originally developed for isolated word recognition application [51], and later adapted by Furui for text-dependent speaker verification [10]. The template model is a sequence of feature vectors $(Y_1, Y_2, ..., Y_N)$. The DTW is used to find a match between the template model and the input sequence $(X_1, X_2, ..., X_M)$. In general, N is not equal to M due to variations in human speech. The matching score, (D), is given as

$$D = \sum_{i=1}^{M} d(X_i, Y_{j(i)}) \tag{2.1}$$

where the template indices $j(i)$ is given by the DTW algorithm, and $d(.)$ represents the distance between the feature vectors. $j(i)$ corresponds to the index of the feature vector of the template sequence Y, which matches with the i^{th} vector of the input sequence X.

Given the reference and test input data, the DTW algorithm does a constrained, nonlinear mapping of one time axis onto the other to align the two, by minimizing D. The accumulated distance is used as the matching score. Chapter 3 gives a detailed description of the DTW technique and its application for speaker verification. The disadvantages of DTW based speaker verification system are the following [52, 53]:

1. The performance of the system critically depends on how accurately the end-points of the speech signal are located.
2. Computation time needed to perform the DTW match could be high
3. Large storage requirement

2.3.1.2 Vector quantization

Another form of template model which uses multiple templates to represent the frames of speech is known as Vector Quantization(VQ) [54, 55]. A VQ codebook is designed by standard clustering procedures for each enrolled speaker using his training data. Consider a set of N training feature vectors $\{\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_N}\}$ which constitute the feature space S characterizing the variability of a speaker. The feature space (S) of the speaker is partitioned into M components $\{S_1, S_2, ..., S_M\}$. Each partition S_i forms a non-overlapping region, and it is represented by the corresponding centroid $\mathbf{b_i}$ of S_i. Partitioning is done such that the average distance D is minimized over the whole training set. The match score D for N frames of speech is

$$D = \frac{1}{N} \sum_{i=1}^{N} \min_{1 \leq j \leq M} \mathbf{d}(\mathbf{x_i}, \mathbf{b_j}) \tag{2.2}$$

where, $d(\mathbf{x_i}, \mathbf{b_j})$ denotes the distance between $\mathbf{x_i}$ and $\mathbf{b_j}$. Each mean can be considered as a spectral template representing the phonetic sound cluster of speaker's speech. In speaker identification, for a set of L test feature vectors $\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_L}$, the quantization errors with respect to each code book are individually accumulated across the whole test data. Average distance with respect to the k^{th} code book is,

$$D_k = \frac{1}{L} \sum_{i=1}^{L} \min_{1 \leq j \leq M} d(\mathbf{x_i}, \mathbf{b_j}) \tag{2.3}$$

The speaker recognition decision for a system of K enrolled speakers is given by

$$\mathbf{speaker} = \mathbf{arg}\,(\min_{1 \leq k \leq K} (D_k)) \tag{2.4}$$

In the case of speaker verification, average distance of the code book of the claimed speaker is found. This is compared with a preset threshold to accept or reject the identity claim made by the speaker. The disadvantage of using this method is that it needs large amount of training data to generalize the model.

2.3.2 Stochastic models

In stochastic models, the pattern matching problem can be formulated as measuring the likelihood of an observation, given the speaker model. The feature vectors extracted from different speakers have different distribution in the feature space. Probabilistic modeling involves capturing such speaker-specific distribution during training. During testing, the probability that the test feature has come from a particular speaker's distribution is obtained to arrive at a decision. Based on the likelihood estimation, probabilistic modeling can be either parametric or non-parametric. If the model is parametric, then it assumes a structure characterized by parameters of the distribution.

Non-parametric models on the other hand make minimal assumption regarding the distribution. Hidden Markov Models (HMM) and Gaussian Mixture Models (GMM) are parametric models.

2.3.2.1 Hidden Markov models

HMM is a popular stochastic model for modeling sequences [25]. This temporal structure modeling is advantageous for text-dependent speaker verification task. In conventional Markov models, each state corresponds to a deterministically observable event. In HMM, the observations are a probabilistic function of the state, i.e., the model is a doubly embedded stochastic process, where the underlying stochastic process is not directly observable (it is hidden). HMM captures the underlying speech sounds as well as the temporal sequencing of the sounds. This temporal structure modeling is advantageous for text-dependent tasks. In the case of text-independent systems the sequence of sounds units in testing and training case will differ. HMMs have the following disadvantages:

1. To estimate the speaker model reliably, a large amount of training data is required
2. The performance degrades if the conditions for training and testing are different

2.3.2.2 Gaussian mixture models

Gaussian mixture model assumes a Gaussian mixture density to distribution of the feature vectors in the feature space [12]. The complete Gaussian mixture density is parametrized by the mean vectors (μ), covariance matrices ($\sum$) and mixture weights (p). These parameters are completely represented by the notation

$$\lambda = \{p_i, \mu_i, \textstyle\sum_i\},\ \ i = 1, 2, ..., M$$

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

$$p(\mathbf{x}|\lambda) = \sum_{i=1}^{M} p_i b_i(\mathbf{x}) \tag{2.5}$$

where,

$\mathbf{x}$ is a D dimensional feature vector,

$b_i(\mathbf{x}),\ i = 1, ..., M$ are the component densities, and

$p_i,\ i = 1, ..., M$ are the mixture weights

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

$$b_i(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{D}{2}} |\sum|^{\frac{1}{2}}} exp\{-\frac{1}{2}(\mathbf{x} - \mu_i)^t \sum_i^{-1} (\mathbf{x} - \mu_i)\} \tag{2.6}$$

where, μ_i is the mean vector, and $\sum_i$ is the covariance matrix.

The vertical bar indicate the determinant, i.e, $|\sum|$ is determinant of $\sum$. The mixture weights satisfy the constraint $\sum_{i=1}^{M} p_i = 1$.

During training, the goal is to estimate the parameters of the GMM (λ) from the feature vectors collected from the training utterances of the speaker. For speaker identification, a group of speaker is represented by $\lambda_1, \lambda_2, ..., \lambda_M$. The objective is to find out the speaker model which has the maximum likelihood for a given observation sequence. For speaker verification, a simple method is to fix a threshold for the likelihood. Speaker will be accepted if the probability for the target model is greater than the threshold, else he will be rejected.

2.3.3 Artificial neural networks

Artificial neural networks have been used for text-independent speaker recognition applications [30, 31, 56, 57]. Artificial neural networks (ANN) have a powerful and flexible architecture for solving classification problems. ANNs are easy to implement and are suitable for discriminative training. The results of ANN based approach are comparable to VQ based approaches. Time delay neural networks (TDNN) and recurrent neural networks (RNN) have been investigated to use the contextual information [58]. Application of auto-associative neural networks (AANN) for source and system features have been proposed for text-independent speaker verification [16, 59]. The distribution estimation capability of AANN is exploited for text-independent speaker verification.

2.4 COMBINING MULTIPLE CLASSIFIERS FOR SPEAKER RECOGNITION

We have seen that there are several different features and classification algorithms for speaker recognition task. But individually they are not ideal for a practical real-world application [60]. Efforts have been made to lump several features together. It has been suggested that a combination of static and dynamic spectral information improves the performance of the speaker recognition system [35]. A 24-dimensional feature vector consisting of 12 cepstral and 12 delta-cepstral coefficients was used in this study. However, no significant improvement in the performance of the system was observed. It is also seen that lumping of features lead to *curse of dimensionality* problem [38]. Studies have been made in [61] by several combinations of MFCC, RASTA-PLP, delta MFCC, delta RASTA-PLP. Combining features from the vocal tract characteristics and voice source characteristics were shown to be useful by Matsui and Furui [62]. Dimension of the concatenated features were compressed using the technique of Linear Discriminant Analysis (LDA) in [63]. Recently, it has been observed that classifiers and features of different types can complement one another in the classification performance [64]. In a recent study, the DTW, second order statistics and HMM methods were combined to improve the performance of the speaker verification system. Chen combined different features and different classifiers under three frameworks, namely, linear opinion pool [60], winner-take-all and evidential reasoning to improve the performance of the text-independent speaker verification system. In another experiment, the output of two neural networks were combined by training a new network on top of the two independent neural networks to improve the decision making [65]. It has been shown in [49] that combining evidence from multiple networks which differ in the architecture and which use the same input feature vector has helped in improving the performance of the text-dependent speaker verification system.

2.5 SUMMARY

In this chapter, we have discussed several speaker-specific features and pattern modeling techniques for speaker recognition. It has been noted that a single feature descriptor and a particular classification algorithm may not be suitable for a practical application. In this work, we have explored the possibility of incorporating suprasegmental and source characteristics, into the basic system, for speaker verification. The existing basic system uses only the short-term spectral features for speaker verification. The description of the basic system and the proposed refinements are described in the following chapters.

CHAPTER 3

SPEAKER VERIFICATION USING SEGMENTAL FEATURES

The uniqueness of a speaker's voice is associated with the physiological and the behavioural characteristics of the speaker. By physiological characteristics, we mean the shape and size of the speaker's speech production system, and the dynamics of the articulators. Variations in the physiological characteristics from speaker to speaker produce differences in the characteristic resonances in the spectrum of the speech signal. The basic system uses the traditional approach to capture the physiological characteristics by extracting the short-term spectral features from a small segment of 10-30 msec of the speech signal, called segmental features. The basic system for text-dependent speaker verification consists of training and testing phases. In the training phase, reference templates are generated for every speaker. During the testing phase, the test pattern generated from the test utterance is compared with the reference pattern, and a decision is made using a decision logic.

This chapter is organized as follows: In Section 3.1, the method of extracting segmental features, and the process of matching reference and test patterns using the Dynamic Time Warping (DTW) technique are explained. The basic system uses amplitude of the speech signal to locate the begin and end of the utterance. This method fails for a noisy speech data. An alternate method for locating the endpoints of the speech utterance based on vowel onset points is described in Section 3.2. Performance evaluation of the system is given in Section 3.3. Section 3.4 gives a summary of this chapter.

3.1 DESCRIPTION OF THE BASIC SYSTEM FOR TEXT-DEPENDENT SPEAKER VERIFICATION

Speaker verification consists of the following four stages:

1. Preprocessing
2. Feature extraction
3. Pattern classification
4. Decision making

3.1.1 Preprocessing

3.1.1.1 Endpoint detection based on amplitude

Locating the begin and end points of an utterance is important in a text-dependent speaker verification system. Correct detection of the endpoints of a speech utterance reduces the computation, and also increases the accuracy of aligning the reference and test utterances. Endpoint detection is not trivial if the Signal-to-Noise Ratio (SNR) is small. An algorithm based on the measurement of the amplitude of the speech signal in the time domain is used in the basic system. It assumes Gaussian distribution of the amplitudes. The speech signal is blocked into frames of size 20 msec with a frame shift of 6.4 msec. The maximum positive amplitude of each frame is determined ignoring the negative excursions for simplicity. The average of the maximum positive amplitudes is taken, and the mean and standard deviation are computed. Mean plus standard deviation is taken as the maximum amplitude, instead of taking the maximum amplitude of the entire utterance. The reason for doing this is that an inadvertent large amplitude found in the speech utterance should not be picked up as the maximum, even though most of the speech is of low amplitude. Ten percent of the maximum amplitude is taken as the threshold for a frame to be considered as a speech frame.

That is, if the maximum amplitude in a frame is less than the threshold, it is considered as a silence or non-speech frame. Otherwise it is considered as a speech frame.

3.1.1.2 Pre-emphasis

The speech samples in each frame are preprocessed using a difference operator to emphasize the high frequency components. This is done using a preemphasis filter of the following form:

$$H(z) = 1 - z^{-1} \tag{3.1}$$

3.1.1.3 Windowing

After preemphasizing, each frame is windowed using a window function. The windowing ensures that the signal discontinuities at the begin and end of each frame is minimized. The window function used is the Hamming window given below.

$$w(n) = 0.54 - 0.46cos((2\pi n)/(N-1)), \qquad 0 \leq n \leq N-1 \tag{3.2}$$

where N is the number of samples in a frame.

3.1.2 Feature extraction

The speech signal is represented as a sequence of feature vectors consisting of spectral components. The shape of the vocal tract contributes to the speaker's voice characteristics at the segmental level. That is, the shape information of the vocal tract is reflected indirectly in the envelope of the short-term spectrum. The spectral information for each frame of the speech signal is represented using cepstral coefficients and delta-cepstral coefficients [10].

Extraction of cepstral and delta-cepstral coefficients: Voiced speech production can be considered as air forced through the vocal folds producing periodic pulses, which are filtered by the shape of the vocal tract. Hence voiced speech can be modeled as quasi-periodic pulse followed by a linear time-invariant (LTI) filter. Let $s(t)$ denote the speech signal, $h(t)$ the impulse response of LTI filter and $e(t)$ the periodic pulse signal. Then,

$$s(t) = h(t) * e(t) \tag{3.3}$$

The first step is to compute the short-term spectrum of the signal. The magnitude spectrum $|S(f)|$ of the speech signal can be viewed as consisting of a rapidly varying part $(|E(f)|)$ and a slowly varying part $(|H(f)|)$.

$$|S(f)| = |H(f)|.|E(f)| \tag{3.4}$$

By taking logarithm on both sides of equation (3.4), a desirable separation of the vocal tract components and the periodic pulses is achieved.

$$\log|S(f)| = \log|H(f)| + \log|E(f)| \tag{3.5}$$

The cepstrum is computed by taking the inverse DFT of equation (3.5). The excitation part $E(f)$ and the vocal tract system part $H(f)$ manifest themselves at larger and smaller values of the quefrency (time), respectively. Since the coefficients at the smaller values of quefrency are mainly decided by the shape of the vocal tract, they are used as features for speech/speaker recognition. These coefficients are called cepstral coefficients. In addition to the cepstral parameters, the differenced cepstra are used as additional features in speech/speaker recognition. These features are also called as transitional features. It has been found that instantaneous and transitional representations are relatively uncorrelated, and provide complementary information for speaker recognition. Thus cepstral and differenced cepstral parameters together represent the vocal tract characteristics of a speaker effectively. The cepstral coefficients are weighted appropriately before using them for speech/speaker recognition. It is because the lower order cepstral coefficients are sensitive to overall spectral slope, and higher cepstral coefficients are susceptible to noise. A tapered window of the form given below is used to minimize these sensitivities [25].

$$w(i) = 1 + (N/2)sin(i\pi/N), \qquad i = 0, 1, 2, ..., N-1 \tag{3.6}$$

where N is the number of cepstral coefficients being used.

3.1.3 Pattern classification by template matching using DTW

The feature vector consists of 20 weighted cepstral coefficients and 5 delta cepstral coefficients for each frame of the speech signal. A template-based pattern comparison approach is used in this system. For training or enrollment, a reference template is created and is stored for every speaker. For testing , the test pattern generated from the test utterance is compared with the reference template by the DTW technique [51].

Dynamic Time Warping: DTW is a dynamic programming pattern matching algorithm with nonlinear time-normalization. This nonlinear time-normalization takes care of the timing difference between the two speech patterns by warping the time axis of one so that the best match is attained with the other. The process of time-normalization requires certain constraints on the warping path. A brief description of typical constraints along with the practical DTW algorithm is given in Appendix A. Speech can be represented as a sequence of feature vectors as shown below.

$$\begin{aligned} A &= \mathbf{a_1}, \mathbf{a_2}, ..., \mathbf{a_x}, ..., \mathbf{a_X} \\ B &= \mathbf{b_1}, \mathbf{b_2}, ..., \mathbf{b_y}, ..., \mathbf{b_Y} \end{aligned} \tag{3.7}$$

Consider the x-y plane shown in the Fig.3.1, where the patterns A and B are represented along the x and y axis, respectively. The match between them can be repre-

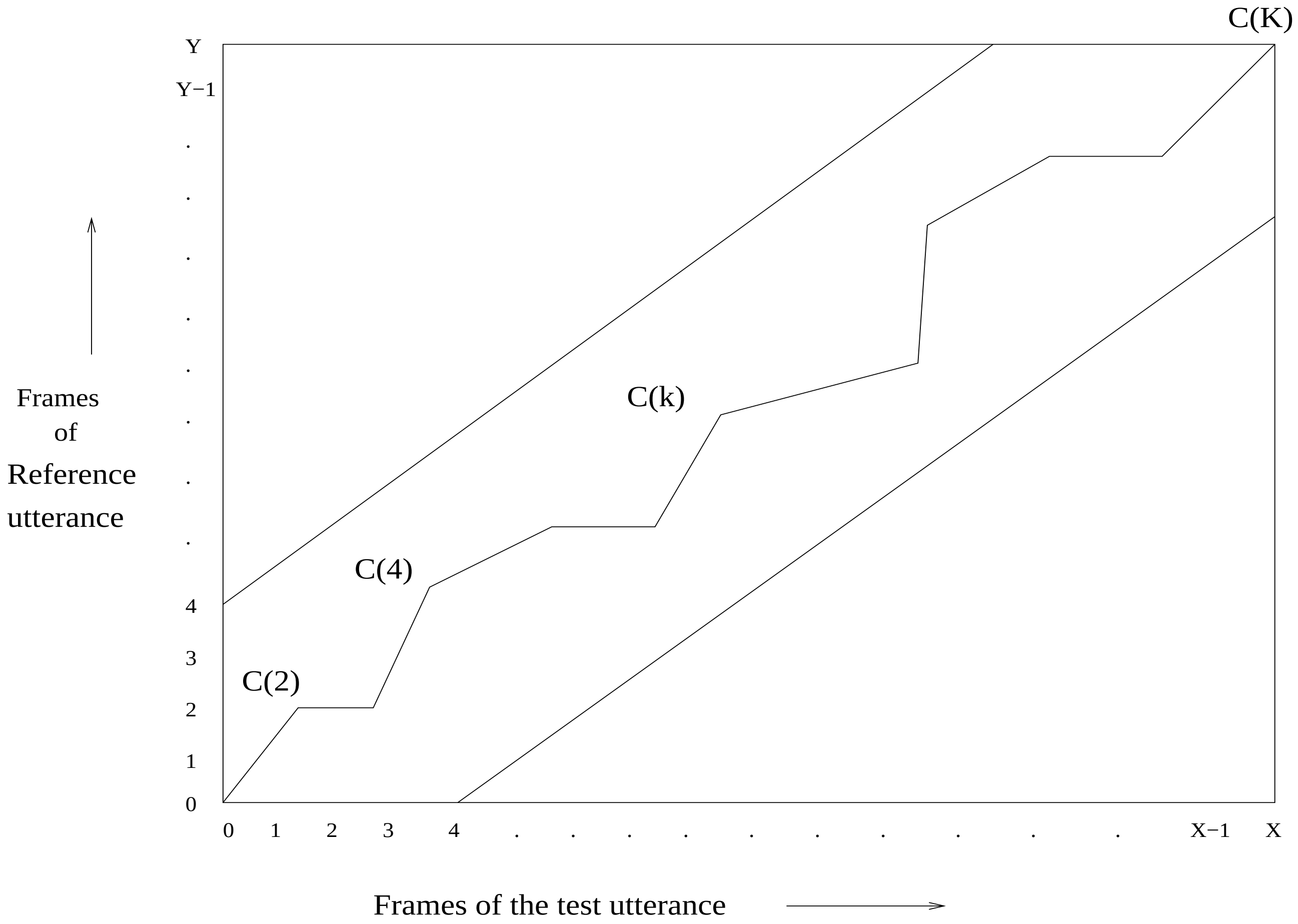

C(K)
Y
Y−1
Frames
of
Reference
utterance
C(k)
C(4)
C(2)
4
3
2
1
0
0 1 2 3 4 . . . . X−1 X
Frames of the test utterance

Fig. 3.1: A typical Dynamic Time Warping Path sented by a sequence of K points $C(1), C(2), ..., C(k)..., C(K)$, where each $C(k) = (x(k), y(k))$, and $x(k)$ is a frame of the test utterance and $y(k)$ is a frame of the reference utterance. Then the sequence,

$$F = C(1), C(2), ...C(k), ..., C(K), \qquad where \quad C(k) = (x(k), y(k)) \tag{3.8}$$

represents a a mapping from the time axis of the pattern A onto that of the pattern B. This mapping is called the warping function. The region between two lines marked in the x-y plane defines the allowable grid region or the region through which the warping path can traverse (see Appendix A). As a measure of the difference between the test and reference feature vectors $\mathbf{a_x}$ and $\mathbf{b_y}$, a distance $d(C)$ is computed between them as shown,

$$d(C(k)) = \|\mathbf{a_x} - \mathbf{b_y}\| \tag{3.9}$$

$$E(F) = \sum_{k=1}^{K} d(C(k)) \tag{3.10}$$

The distance $E(F)$ is the similarity score which is used in the decision making stage. The score attains minimum value when the time differences are adjusted optimally.

DTW for Speaker Verification: In the case of text-dependent speaker verification, the test and reference templates will have similar sequence of sound patterns, as the text is fixed. The DTW algorithm eliminates the differences in time between the two speech patterns by a nonlinear warping. The speech signal is normally affected by the transmission channel response. This could be interpreted as a convolution operation. In the cepstral domain each cepstral coefficient c_x is represented as

$$c_x = c_x^s + c_x^c \tag{3.11}$$

where c_x^s represents the linguistic information and the speaker information, and c_x^c represents the channel effect. When DTW is performed between the reference template and the test pattern, a similarity score $E(F)$ is obtained. The similarity score is the Euclidean distance between the feature vectors of the test and reference templates. The difference between the cepstral coefficients (c_x^d) is used in the computation of the Euclidean distance.

$$c_x^d = c_x^r - c_x^t \tag{3.12}$$

where the superscripts t and r refers to test and reference patterns, respectively. By writing each cepstral coefficient as

$$c_x^r = c_x^{rs} + c_x^{rc} \tag{3.13}$$

$$c_x^t = c_x^{ts} + c_x^{tc} \tag{3.14}$$

and, assuming that the channel characteristics being the same during the enrollment and verification phase, i.e., $c_x^{rc} = c_x^{tc}$, equation (3.12) can be written as

$$c_x^d = c_x^{rs} - c_x^{ts} \tag{3.15}$$

Since the text is fixed, the linguistic information being compared is the same. So c_x^d should carry speaker discriminatory information, which can be used for text-dependent speaker recognition.

3.1.4 Decision logic

The training phase involves generation of reference templates for each speaker. The templates are the cepstral feature vectors (20 weighted cepstral coefficient and 5 delta cepstral coefficients) for each 20 msec segment of the speech signal. Three reference templates are generated for every speaker and stored for future recall during the testing phase. The reason for generating three reference templates is to take care of the intra-speaker variability of the speakers. The template represents the vocal tract characteristics of a given speaker. Verification is performed by identification. For this purpose, for each speaker nine other speaker models are used as background speakers. During the testing phase, the test pattern is matched against the three reference templates of the target speaker, and one reference template for each of the nine background speakers, yielding 12 similarity scores. The decision is made based on the ranking of the 12 similarity scores. The 12 similarity scores are sorted in ascending order. If the first two ranks are obtained by the claimed speaker templates, then the claim is accepted, else the claim is rejected.

3.2 ALTERNATE METHOD FOR ENDPOINT DETECTION USING VOWEL ONSET POINTS

One of the severe limitations of using DTW technique for text-dependent speaker verification is that the alignment of the reference and the test pattern critically depends on the accuracy of the located endpoints. In the basic system, it was seen that the technique for endpoint detection based on amplitude fails in cases such as noisy speech data or telephone speech. Hence, it is necessary to use the knowledge of the events that occur in the speech signal to locate the endpoint. In this study, we have used the Vowel Onset Points (VOP) as anchor points to locate the begin and end of the speech utterance. VOP is the instant at which the onset of the vowel takes place.

The VOPs are obtained from the Hilbert envelope of the Linear Prediction (LP) residual signal [66, 67]. LP residual (see Appendix B) of the speech signal contains the information regarding the characteristics of the excitation source. The actual excitation function for speech is essentially either a quasi periodic pulse train (for voiced speech sounds) or a random noise source (for unvoiced sounds). The Hilbert envelope

$h(n)$ of the residual signal $r(n)$ is obtained as

$$h(n) = \sqrt{r^2(n) + r_h^2(n)} \tag{3.16}$$

where, $r_h(n)$ is Hilbert transform of $r(n)$. The Hilbert transform of a signal $r(n)$ is obtained by exchanging the real and imaginary parts of the DFT of $r(n)$, and then computing the IDFT. Table 3.1 gives the algorithm used for locating the VOPs [67].

Table 3.1: **Algorithm for automatic detection of the vowel onset points in continuous speech using source feature**

1. Preemphasise the input speech
2. Select only the high SNR (Signal to Noise Ratio) portions of the input speech (upto 2.5 kHz) by low pass filtering
3. Compute the LP residual using 8^{th} order LP analysis, with a frame size of 20 ms and frame shift of 10 ms
4. Compute the Hilbert envelope of the LP residual
5. Obtain the VOP evidence plot from the Hilbert envelope for every sample shift using Gabor filter
6. Identify the peaks in the evidence plot using peak picking algorithm
7. For each peak, if there is no negative region with reference to the next peak, then eliminate such spurious peaks
8. Hypothesize the remaining peaks as VOPs

The VOPs are processed to eliminate any spurious occurrences. This is done by checking the following conditions. The time gap between VOPs are not likely to exceed 500 msec in the case of a fixed text continuous speech utterance. Hence, if any VOP lies outside this range it is considered to be spurious. The VOPs should essentially be followed by a vowel region of minimum 50 msec. By this process, we eliminate any spurious VOPs, if present. The first and last VOPs are used to locate the end-points. After locating the first VOP, a continous silence region of 300 msec is located by searching back wards from this point. The start of this silence region is marked as the begin of the speech utterance. Similarly, the end of the speech utterance is located using the knowledge of the location of the last VOP.

3.3 PERFORMANCE EVALUATION OF THE SYSTEM BASED ON SEGMENTAL FEATURES

Performance evaluation of the text-dependent speaker verification system (basic system) based on short-term spectral features is discussed in this section. The system has been analyzed on both microphone speech data and telephone speech data. The quality of the speech signal obtained through these devices differ mainly due to differences in the transmission channel. A typical telephone channel has a passband of approximately 300-3300 Hz. Signal energy outside this range are attenuated. However, the nominal bandwidth can vary from one call to another and a mismatch of the passband characteristics between the training and testing sessions may occur. In addition to the bandwidth limitation, telephone channel introduces distortion to the spectral characteristics of the speech signal. It also introduces additive noise and glitches.

3.3.1 Speech data collection

The difference in the quality of the speech recorded through telephone as well as microphone has made it necessary to collect the speech data through both of these devices for evaluating the performance of the system. Speech data was collected from 30 cooperative speakers for one sentence in Hindi language. The sentence consists of five words and is approximately 2 seconds in duration. 22 utterances were collected for each speaker. Hence, in this process, we have collected 22 utterances through microphone and 22 utterances through telephone for each speaker. Recording was done in the laboratory environment. The speech data collected through both microphone as well as telephone was sampled at 8000 Hz, and the data was stored as 8 bit wave format.

3.3.2 Performance analysis

The performance of the system is specified by means of Equal Error Rate (EER), which is the average of the False Rejection (FR) rate and the False Acceptance (FA) rate. False rejection rate is the number of rejected cases of the genuine speaker tests expressed as a percentage of the total number of the genuine speaker tests conducted, whereas the false acceptance rate is the number of accepted cases of the imposter speaker tests expressed as a percentage of the total number of the imposter speaker tests conducted. Studies were made on one sentence for 30 speakers. For each speaker, three utterances were used for creating the three reference templates, and the test utterances were selected from the remaining set. Out of the remaining 19 utterances, 15 utterances of the speaker were used for conducting the genuine speaker tests. Hence the total number of genuine speaker tests for 30 speakers is 450 (30×15). Imposter tests for a speaker are conducted by giving the utterances of the remaining 29 speakers in the database. For each speaker, two utterances each of the same text of the remaining 29 speakers are taken for testing. Hence, the number of imposter speaker tests is 1740 $(30 \times 29 \times 2)$.

3.3.3 Results of the system

In order to illustrate the effectiveness of the endpoint detection based on VOP compared to the amplitude based approach, we have analyzed the system using both the methods. Table 3.2 shows the performance of the text-dependent speaker verification system using short-term spectral features for microphone speech and telephone speech of 30 speakers. False rejection rate indicates the percentage of the number of genuine speakers who have been rejected in the 450 genuine speaker tests conducted, and false acceptance rate indicates percentage of the number of imposter speakers who have been accepted in the 1740 imposter speaker tests.

The basic system which uses the amplitude based approach fails in cases where the speech data is noisy, especially in cases of telephone speech, where high amplitude spikes/glitches and channel noise are present. The errors in locating the end-points cause improper alignment of the reference and test utterances and hence result in arriving at a wrong decision. The performance of the system was evaluated using the VOP knowledge for end-point detection. It can be seen that the performance of the system for microphone speech as well as telephone speech data has significantly improved by exploiting the knowledge of vowel onset points for begin and end detection.

Table 3.2: **Performance of the text-dependent speaker verification system with microphone and telephone speech using spectral information.**

Endpoint detection	Speech Data	False Acceptance Rate (%)	False Rejection Rate (%)	Equal Error Rate (%)
Amplitude	Microphone	4.9	2.0	3.45
	Telephone	5.8	5.7	5.75
VOP	Microphone	3.9	0.2	2.05
	Telephone	4.19	0.6	2.4

Comparing the performance of the system for microphone and telephone speech data, it can be seen that the performance of the system degrades when telephone speech is used. This is mainly due to three reasons. Firstly, the bandwidth limitation of the transmission channel, which causes loss of high frequency information. These high frequency components may contain significant speaker-specific information, when compared to the low frequency components. Secondly, the spectral features get distorted due the frequency response characteristics of the pass-band of the transmission channel. Thirdly, the channel adds noise and glitches to the speech signal, which results in degradation of the extracted spectral features.

The results of the system indicates the necessity of exploring other robust features for speaker verification. Literature suggests that the suprasegmental features such as pitch and duration features are not significantly affected by channel or handset variations, and they can be reliably extracted from the speech degraded by noise [3]. Moreover, they act as additional independent evidence for speaker verification task. Hence, in this thesis we have explored the use of pitch and duration features for speaker verification.

3.4 SUMMARY

A brief description of the basic system which uses the segmental features is given in this chapter. The feature extraction process and the DTW algorithm has been described. Performance of the basic system was analyzed for microphone and telephone speech data. An alternate method for detecting the end-points of the utterance using the knowledge of the VOPs is proposed. It is shown that the performance of the system has improved by using this alternate method for begin and end detection of the speech utterance. The short-term spectral features are sensitive to the channel characteristics. This results in the degradation of the performance of the system for telephone speech data, when compared to the performance for microphone speech data. Hence we explore the use of robust suprasegmental features such as duration and pitch, for speaker verification. The system based on duration and pitch information is described in the next chapter.

CHAPTER 4

DURATION AND PITCH INFORMATION FOR SPEAKER VERIFICATION

In the previous chapter, it was suggested that, in addition to the spectral features, it might be possible to improve the performance of a text-dependent speaker verification system by incorporating other robust features. The basic system exploits only the physiological characteristics such as the shape and size of the speech production system, and it ignores the behavioural characteristics such as the speaking style, speaking rate, stress and accent which are also unique to a speaker. Behavioural characteristics such as intonation (variation in the fundamental frequency as a function of time), duration and stress are reflected at the suprasegmental level or in longer duration of the speech segment. Suprasegmental features are difficult to generalize due to the qualitative nature of these features. Moreover, suprasegmental features have large intra-speaker variability [68]. Though the suprasegmental features are more vulnerable to intra-speaker variability, they are less sensitive to channel or hand-set variation, when compared to short-term spectral features. They may provide additional evidence for speaker verification. With this motivation, we have attempted to incorporate pitch and duration information for speaker verification task.

This chapter is organized as follows: In Section 4.1, we describe a method to extract duration information, and present the results of system using duration information alone. Section 4.2 deals with the extraction of pitch information and performance of the system using the pitch information alone. We conclude our discussion on pitch and duration information in the last section.

4.1 DURATION INFORMATION

The speaking rate of an individual is an important evidence in speaker recognition task. Survey of the literature shows very few attempts in using duration information for speaker recognition task, although attempts do exist for using pitch information [44,49]. Interest for research in the use of duration information for speaker recognition appears to be less because of the difficulty in measuring the duration of the units such as syllable, word or phrase. This has become difficult due to the implicit need of locating the boundaries of the units. Hence, it is useful to extract duration information without explicitly locating the boundary of any type of unit. This is accomplished by using the nature of the dynamic time warping path. The nature of the dynamic time warping path indicates the extent of mismatch between the relative duration of the units of the reference and the test utterances.

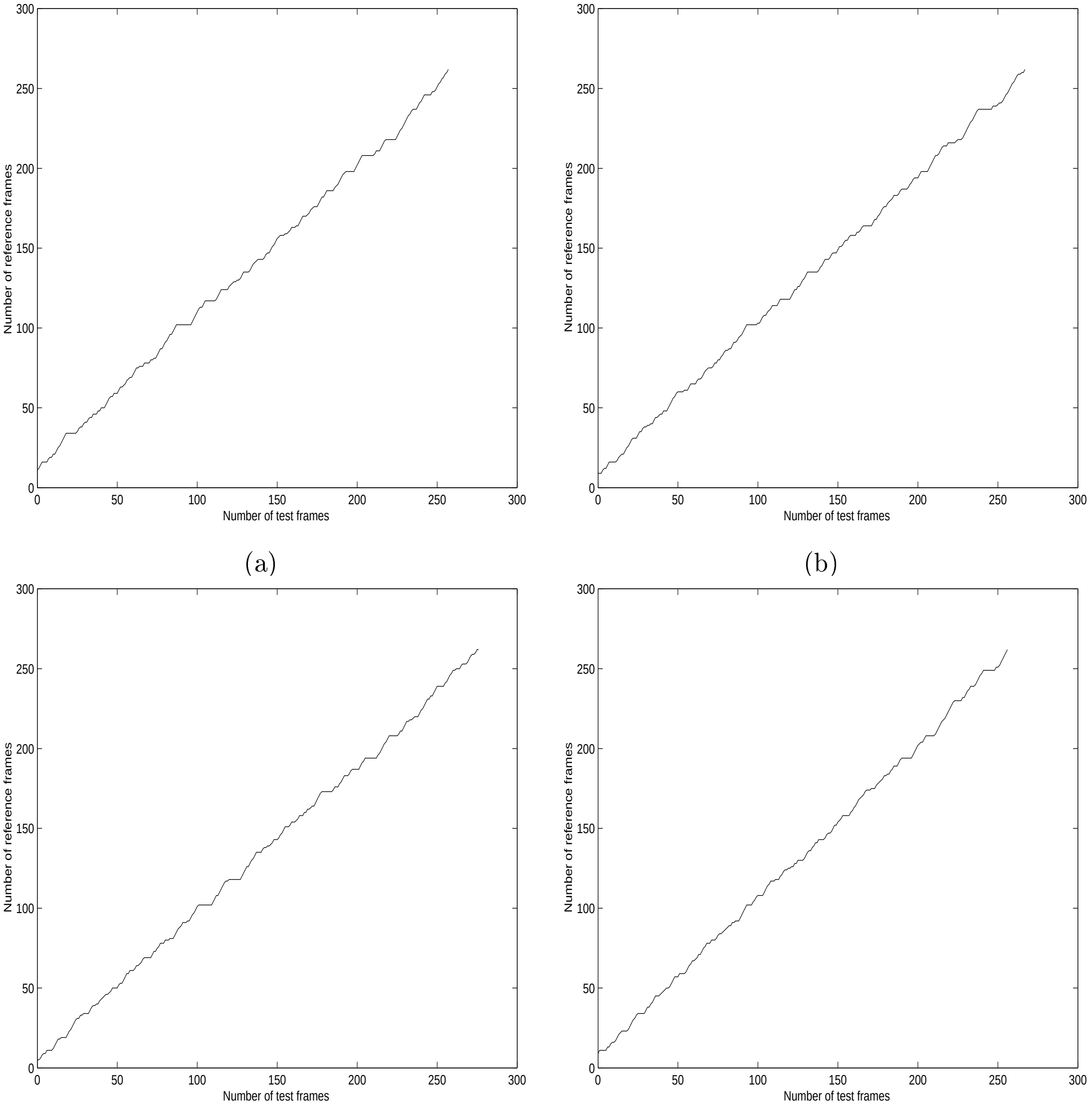
4.1.1 Nature of the dynamic time warping path

It should be noted that, although the text-dependent speaker verification system uses fixed text, it is often difficult for a genuine speaker to reproduce the test utterance in exactly in the same way as he/she has done for producing the reference utterance. Dynamic time warping is a non-linear time normalization technique used for aligning the frames of the reference and test utterance which are of varied durations.

The description of the DTW technique was given in section 3.1.3. The basic system uses this technique only for template matching, and it ignores information present in the nature of the resulting warping path. The DTW path is represented by a sequence of K points $C(1), C(2), ..., C(k)..., C(K)$, where, $C(k) = (x(k), y(k))$, $x(k)$ is a frame of the test utterance, and $y(k)$ is a frame of the reference utterance. An analysis has been carried out to study the nature of the warping path by matching the reference and test patterns of genuine and imposter speakers. It has been observed that the nature of the warping path that joins the points $C(k), k = 1, ..., K$, follows the diagonal line of the x-y plane closely for genuine speakers, whereas it deviates significantly from the diagonal line for imposter speakers. The significance of this behaviour of the warping path is discussed in the next section.

4.1.2 Duration information from the nature of the warping path

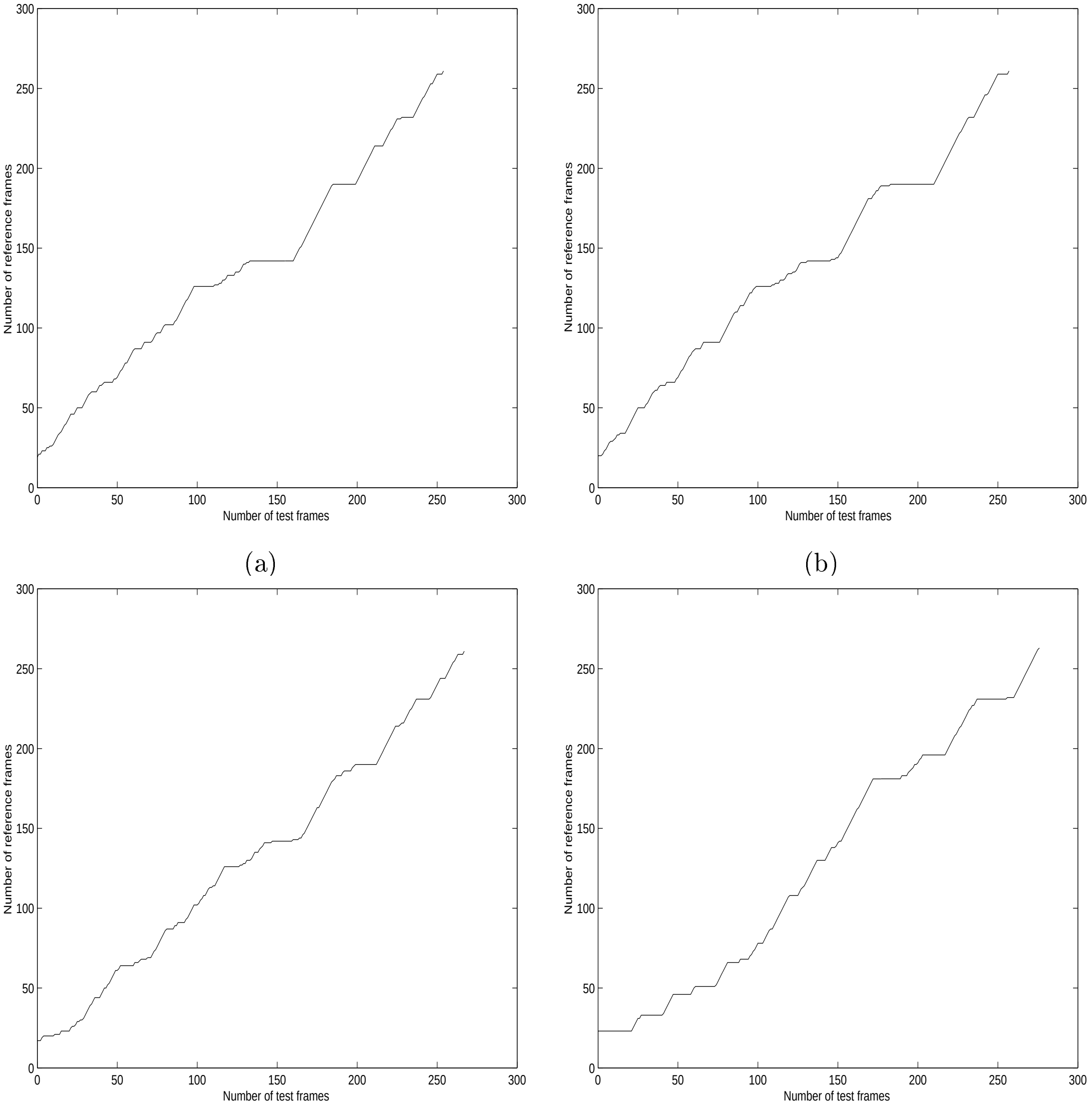
The duration of the test utterance by a genuine or an imposter speaker for a given text may or may not match with the duration of the reference utterance of the target speaker. As a result, it is not possible to arrive at a conclusion based on the absolute duration (amount of time taken) of the entire utterance, or even the absolute duration of each of the units (such as syllable or word or phrase) in the utterance. But it is interesting to note that, although the total duration of the utterance of the same text may vary from that of the reference utterance for the genuine speaker, the relative durations or the percentage durations of the units in the utterance (time taken for the unit expressed as a percentage of the total time of the utterance) is almost consistent. This consistency in the relative durations of the units of the reference and test utterance results in a warping path which is almost straight. It can be noted that if a mismatch occurs between the relative duration of the units of the reference and test utterances, then the nature of the warping path will be highly distorted. In other words, the extent of mismatch between the relative durations of the units of the reference and test utterances are related to the distortion of the warping path, irrespective of the absolute duration of the entire utterance. Fig.4.1 and Fig.4.2 shows some examples of the warping paths of genuine and imposter speaker test patterns respectively, matched against the reference template.



(a) (b)

(c) (d)

Fig. 4.1: Warping paths of genuine speaker utterances: (a), (b), (c) and (d) shows the warping path of four different genuine test utterances matched against the same reference template



(a)

(b)

(c)

(d)

Fig. 4.2: Warping paths of imposter speaker utterances: (a), (b), (c) and (d) shows the warping path of four different imposter test utterances matched against the same reference template

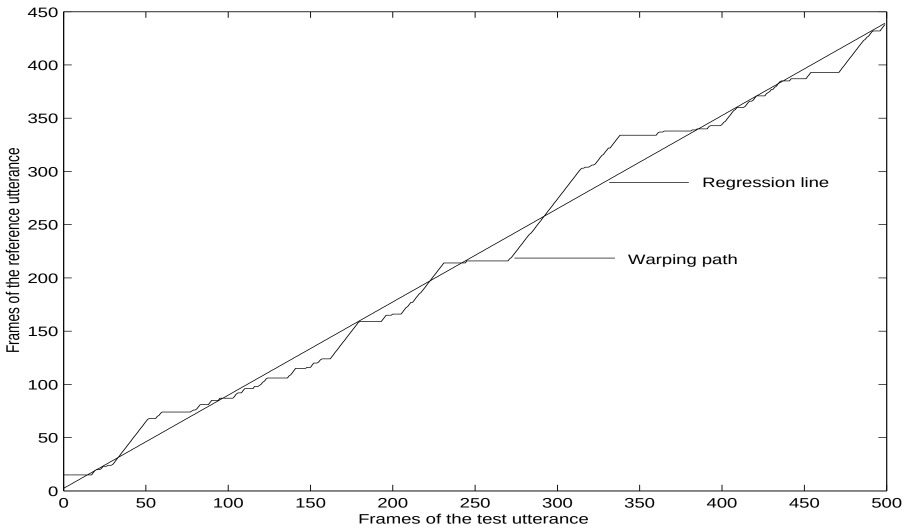


Fig. 4.3: Regression line for an imposter warping path

4.1.3 Extraction of distortion of the warping path

Fig.4.3 shows the warping path of an imposter speaker, and the straight line indicates the regression line or line of best fit of the warping path. The deviation of each point $y(k)$ of the warping path from its regression line is an indication of the mismatch in the relative durations between the units of the reference and test utterances. The regression line of the warping path can be indicated by $y'(k) = mx(k) + c$, where $y'(k)$ is the point on the regression line corresponding to the frame $x(k)$ on the x-axis, m is the slope of the regression line, and c is the intercept of the regression line. The slope and the intercept of the regression line is computed by means of the least squares method. The deviation of the actual warping path from the regression line is indicated by the normalized sum of squared error (E), which is given by,

$$E = \frac{\sum_{k=1}^{K}(y'(k) - y(k))^2}{K} \tag{4.1}$$

where K = Number of points in the warping path.

4.1.4 Results of the system using duration information

In order to have a comparative study of the effectiveness of duration information and spectral information for speaker verification, the same speech database described in section 3.3 is used. The test utterances are matched against the three reference models of the target speaker and one reference model for each of the nine background speakers using the DTW technique. The normalized error as shown in equation (4.1) is computed for each of the 12 tests. These error values are then sorted in ascending order. If the target model comes in the first and second position, then the claim is accepted as a genuine one, else it is rejected. Table 4.1 shows the results of the text-dependent speaker verification system tested on microphone and telephone speech data using the duration information alone.

Table 4.1: **Results of the text-dependent speaker verification based only on duration information for microphone and telephone speech data**

Speech Data	False Acceptance Rate (%)	False Rejection Rate (%)	Equal Error Rate (%)
Microphone	7.18	6.00	6.59
Telephone	6.49	6.00	6.24

The analysis shows that the distortion of the warping path is an important evidence for speaker recognition task. In fact it exploits the mismatch in the relative durations of the units in the speech utterance. Comparing the performance of the system for microphone and telephone speech data, it can be observed that the duration information does not critically depend on the channel characteristics. The poor performance of the system using duration information alone is due to the intra-speaker variability. Another reason for the poor performance of the system is due to the fact that the pause introduced in the utterance between the words is not taken care of.

4.2 PITCH INFORMATION

In addition to the duration information, the pitch information also contributes to the uniqueness of the speaker's voice at the suprasegmental level. Pitch frequency is the acoustic correlate of the rate of vibration of the vocal folds. The uniqueness of the rate of vibration of the vocal folds is due to the differences in the size of the vocal folds, and also due to the speaking style or the accent imposed by the speaker (which is a learnt attribute). The physiological constraints mentioned above determine the average pitch of the speaker over the entire utterance. The speaking style determines the pitch pattern of the utterance or the variation of the pitch frequency as a function of time, i.e., called intonation. The local variations of the pitch contour is more representative of the speaker than the average pitch of the utterance [69]. Hence, though a speaker's average pitch can be mimicked, it is indeed difficult for an imposter speaker to reproduce the pitch pattern of the utterance.

In this thesis, we have attempted to incorporate the pitch pattern information for a text-dependent speaker verification system. The basic system which is described in chapter 3 uses only the features of the shape and size of the vocal tract system of the speaker, but ignores the pitch information. The evidence obtained from the pitch information is likely to be complementary, at least for some speakers, to that obtained from the spectral information of the speech signal. It has also been mentioned in the literature, that the pitch features are insensitive to channel variations [44]. They can be reliably extracted from even remote speech signal (signal recorded when the speaker to microphone distance is large) and noisy speech data. Hence pitch information plays a significant role in adverse recording conditions. This fact motivated us to explore techniques for discriminating speakers using pitch information as an additional evidence in the basic system for text-dependent speaker verification system.

4.2.1 Need for a new approach

Several methods exist in the literature which shows attempts made to discriminate between speakers using the average pitch as well as the intonation pattern. Attempts have been made to use statistical methods which assume Gaussian distribution to the frame level pitch frequencies [43]. These methods requires large amount of training data for deriving the statistical parameters. Acquisition of large amount of data is not possible in the case of text-dependent speaker verification system. Other methods use the raw pitch frequency at certain anchor points of the utterance as input to a recognition model such as in a neural network model or a statistical model. These methods fail, as locating the anchor points in the utterance such as syllable nuclei, are prone to errors, depending on the noise level of the speech data or the type of text used. Though attempts have also been made to exploit the local variations of the F_0 contour, it has not yet been possible to exploit the characteristics of the pitch contour. In this research work, the similarities of the intonation pattern of the reference and test utterances are captured by using the DTW algorithm. The pitch frequency of the utterance has been extracted using the Simple Inverse Filter Tracking (SIFT) algorithm. The advantages of SIFT algorithm over other techniques based on autocorrelation functions, cepstrum, inverse filter tracking, parallel processing time domain methods, data reduction, average magnitude difference functions, group delay functions, are [70–73]:

1. Easy voiced/unvoiced decision
2. Implementation is easy
3. It is efficient, and accurate F_0 is obtained under clean speech conditions.

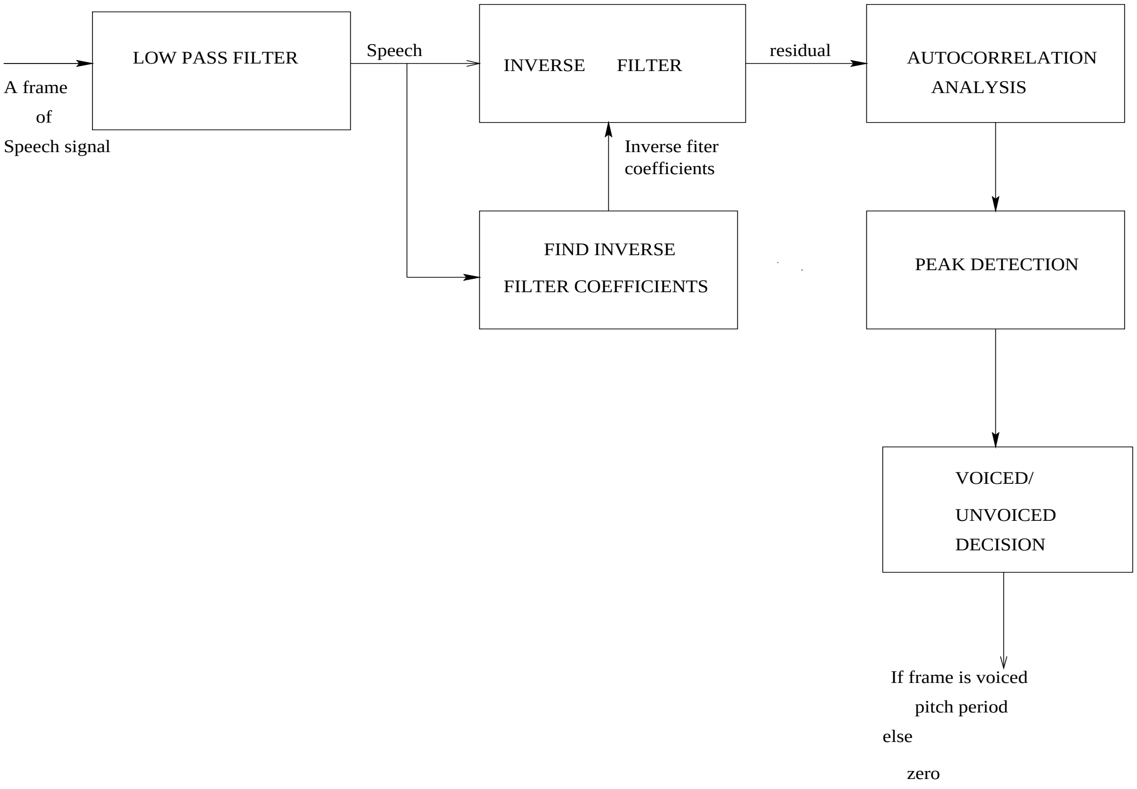


Fig. 4.4: Block Diagram of the SIFT algorithm

4.2.2 Simple inverse filter tracking algorithm

This method is based on inverse filter formulation which retains the advantage of the autocorrelation and cepstral analysis techniques. The block diagram of the SIFT algorithm is shown in Fig 4.4 [72]. The speech signal is prefiltered by a low pass filter with a cut off frequency of 800 Hz, and the output of the filter is sampled at 2 kHz. The sampled speech signal is then segmented into analysis frames of suitable size. The frame size and frame shift here are 20 msec and 6.4 msec, respectively. Short term autocorrelation is performed on the data of each frame, and the first 11 autocorrelation terms are obtained. The inverse filter coefficients a_i are obtained by solving a set of equations using Durbin algorithm [22].

The inverse filter is defined as

$$A(z) = 1 + \sum_{i=1}^{10} a_i z^{-1} \tag{4.2}$$

Autocorrelation analysis is performed over the output of the inverse filter. The distance between the largest peak of the autocorrelation sequence and the next largest peak within the specified limits of 2.5 msec - 12.5 msec corresponds to the pitch period. The F_0 contour is post-processed to eliminate low voiced regions which are likely to have erroneous pitch period values. This is done by forcing the pitch period of frames, which have energy value below 70% of the maximum energy over the entire utterance, to zero.

4.2.3 Method of discriminating speakers using pitch information

The basic system uses the DTW technique to align the reference and test patterns using the cepstral feature vectors. In this process, we get the matching frames of the reference and test utterances. The pitch contour of the reference and test utterances are computed using the SIFT algorithm. The difference of the pitch frequencies of a few selected matching frames in the reference and test utterances are summed up to get the pitch score (P_s). In fact the score obtained, gives the information regarding variation of the pitch frequency. A few pairs of matching frames, say 20, are selected such that the Euclidean distance between the spectral feature vectors of these pairs are lower among all the points in the warping path, and also it should be ensured that these pairs should have a non-zero pitch frequency for both the reference as well as the test frames. By doing this we ensure that the sound units taken are those which are similar to the extent possible in both the reference and test utterances, and those sound units are voiced. Computation of the pitch score can be expressed as follows:

$$P_s = \sum_{i=1}^{N} |F_0(x(i)) - F_0(y(i))| \tag{4.3}$$

where,

$F_0(x(i))$ = Pitch frequency of the frame $x(i)$ of the test utterance,

$F_0(y(i))$ = Pitch frequency of the frame $y(i)$ of the reference utterance,

N = Number of points (about 20) that are taken for computing the pitch score, which are least distant among the K points (total number of points in the warping path), and which satisfy the condition $F_0(x(i)) \neq 0 \quad \text{and} \quad F_0(y(i)) \neq 0$

4.2.4 Results of the system using pitch information

The speech database and the number of genuine and imposter speaker test are the same as described in Section 3.3.1 and 3.3.2. The test utterance is matched against the three target models and nine background models. The pitch score as mentioned in equation (4.3) is computed for the 12 tests. The scores are arranged in ascending order, and the decision logic used is that if the target speaker comes in the first and second position the claim is accepted, else it is rejected. The results of the text-dependent speaker verification system using pitch information alone for microphone and telephone speech are given in Table 4.2.

The experiments show that the pitch information is a useful feature for speaker verification task. The degradation in the performance of the system for telephone speech data may be attributed to the pitch frequency extraction algorithm. SIFT algorithm is likely to fail for noisy speech data such as telephone speech. It has been observed that the evidence produced in many tests are complementary in nature to the evidence obtained from duration and spectral information. Hence this information can be used to improve the performance of the basic system.

Table 4.2: **Results of the text-dependent speaker verification based only on pitch information for microphone and telephone speech data**

Speech Data	False Acceptance Rate (%)	False Rejection Rate (%)	Equal Error Rate (%)
Microphone	7.64	16.8	12.22
Telephone	7.8	24.4	16.1

4.3 SUMMARY

In this chapter, we have explored the use of suprasegmental features such as pitch and duration for speaker verification. The distortion of the DTW path is shown to give evidence regarding the mismatch in the relative durations of the units. The intonation information is captured using the knowledge of the matching frames of the reference and test utterances. The effectiveness of using duration and pitch features individually has been studied. This information can act as additional evidence for improving the performance of the basic system, which uses only the short-term spectral features for speaker verification. The next chapter discusses the performance of a system which uses the evidence from excitation source, and chapter 6 examines the performance of a system which combines evidence from several sources.

CHAPTER 5

SOURCE INFORMATION FOR SPEAKER VERIFICATION

5.1 INTRODUCTION

The characteristics of the time-varying excitation source that is used to excite the vocal tract system for the production of the speech signal is considered to be an important cue for automatic speaker verification. In other words, the uniqueness of the speaker's voice can also be attributed to the uniqueness in the excitation source signal. In the previous chapter, we have attempted to extract the high level information such as intonation and duration, which are present in longer duration of the speech signal. Information regarding the excitation source characteristics or the nature of vibration of the vocal folds are reflected in the Linear Prediction (LP) residual of the speech signal. It was shown that Auto-Associative Neural Network (AANN) models can be used to capture the speaker-specific information in the residual signal [16]. In this chapter, we describe studies made on the source characteristics of the speech signal for text-dependent speaker verification.

This chapter is organized as follows: In Section 5.2, the nature of the signal (LP residual) used for capturing the information regarding the excitation source is described. Section 5.3 gives a description of the AANN models used for learning the speaker-specific information present in the residual signal. The method of training and testing the AANN models and the result of experiments conducted are given in Section 5.4. Section 5.5 gives a discussion on the issue of the significance of the source characteristics for text-dependent speaker verification task. Section 5.6 gives a summary of the chapter.

5.2 SPEAKER-SPECIFIC INFORMATION CONTAINED IN THE LP RESIDUAL OF THE SPEECH SIGNAL

LP analysis is based on the assumption that speech production is modeled by excitation signal which is either white noise or quasi-periodic pulses fed through an all-pole filter. In the case of unvoiced speech, the excitation is a random noise and in voiced speech it is quasi-periodic pulses. For the all-pole model for the vocal tract, the signal s_n can be represented as a linear combination of the past p values and some input u_n, as shown below

$$s_n = -\sum_{k=1}^{p} a_k s_{n-k} + G u_n \tag{5.1}$$

where p is the order of prediction, and G is the gain factor.

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

$$\hat{s_n} = -\sum_{k=1}^{p} a_k s_{n-k} \tag{5.2}$$

The error between the actual value s_n and the predicted value $\hat{s_n}$ is given by,

$$e_n = s_n - \hat{s_n} \tag{5.3}$$

This error e_n is the LP residual of the speech signal, and it contains the source characteristics predominantly.

Experiments show that the LP residual of voiced speech reveals strong correlation. Hence, for voiced speech only the second order correlation is removed by means of the LP analysis, while most of the higher order correlation still remain in the LP residuals. This higher order correlation feature of the residual signal might be specific to the

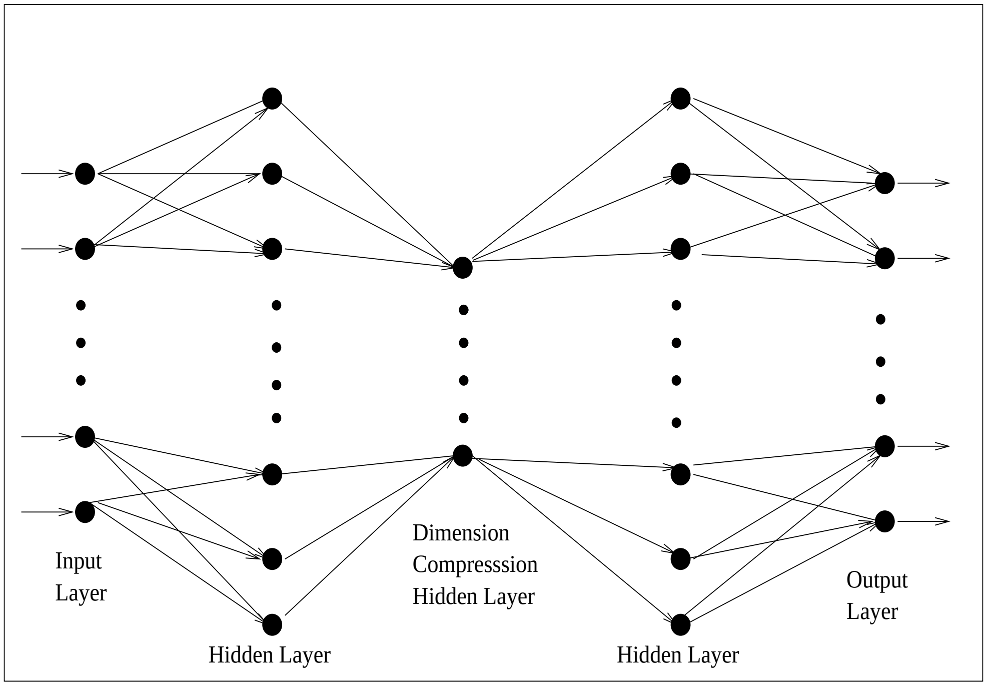


Fig. 5.1: General structure of AANN

speaker. This speaker-specific information present in the residual can be captured using AANN models [16]. The following sections describe the structure of the AANN model that is used for this purpose.

5.3 DESCRIPTION OF THE AANN MODELS

Autoassociative mapping or identity mapping is a technique that can be used to capture the behaviour or the characteristics of the system that generates a given pattern. Autoassociative neural network (AANN) models are a class of artificial neural networks that can perform the function of identity mapping. In other words, the AANN models after learning the system behaviour should be able to reproduce the input vector at the output with minimum error.

The general structure of AANN model is given in Fig 5.1. It consists of one input layer, an output layer and one or more hidden layers. The units in the input and output

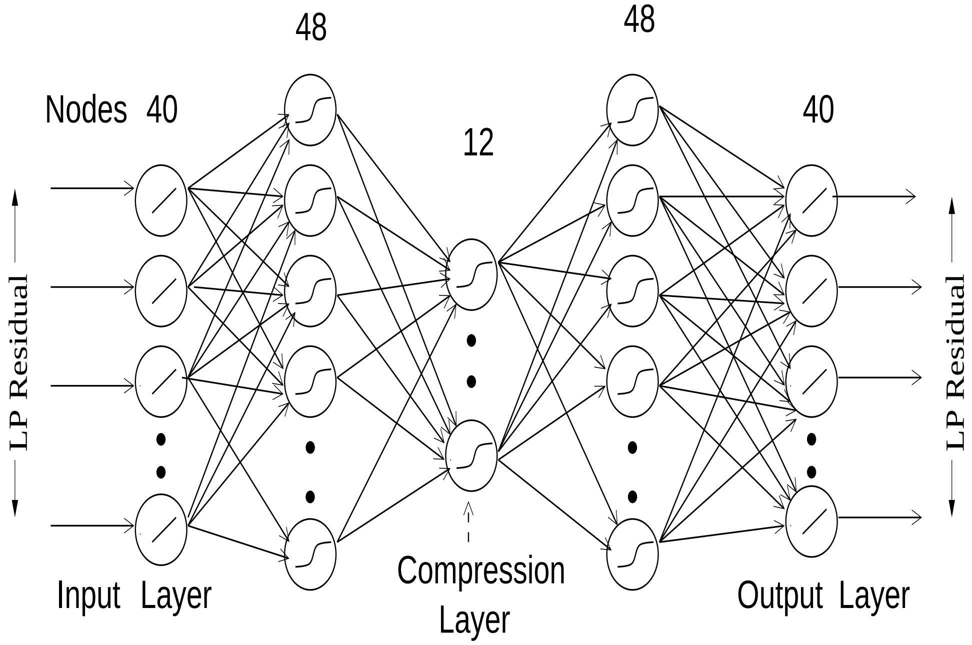


Fig. 5.2: Structure of AANN

layer are linear units, and the units in the hidden layer are nonlinear. The number of units in the input and output layers are equal to the dimension of the input data. One of the hidden layers have lesser number of units than the input and output layers. This layer is called the dimension compression hidden layer. The process of dimension compression is performed so that it may suppress any uncorrelated information. The autoassociative mapping function F can be separated into two parts, F_1 and F_2. F_1 is the dimension reduction process, and this transformation is done in the part of the network from the input layer upto the dimension compression hidden layer. F_2 is the dimension expansion function, and this transformation is done in the part of the network from the dimension compression hidden layer to the output layer.

5.4 SPEAKER VERIFICATION USING SOURCE FEATURES

The source characteristics which is unique to a speaker is seen to be present in the nature of the higher-order correlation among the samples of the LP residual signal. In order to capture the correlation among the samples of the residual, we need to give segments of the residual signal itself as the input vector to the AANN model. Though these input vectors are widely separated in the input space, the dominant information which is the correlation among the samples guides the learning process.
The AANN model used in our study has five layers with the structure shown in Fig 5.2. The structure of the network is $40L\,48N\,12N\,48N\,40L$, where L denotes linear units and N denotes nonlinear units. The training phase of the speaker verification system creates a model for each speaker. The testing phase compares the test utterance with the target model and makes a decision. These phases are discussed below in detail.

5.4.1 Training phase

The AANN model is trained by feeding blocks of 40 samples normalized residual samples shifted by 1 sample at the input layer. The target output is same as the input vector. The learning process is done by using backpropagation algorithm [74–76]. The network is given an initial set of random weights, and the weights are collected after 60 epochs. One epoch consists feeding all the frames of the residual signal in succession. The number of frames is almost equal to the number of the samples, except those belonging to the silence and the low energy regions. The training as well as testing utterances are approximately of 2 to 3 seconds duration, and the speech is digitized at 8 KHz. The final weights define the model for the speaker. Three models are created from three training utterances of a given speaker.

5.4.2 Testing phase

The test features are blocks of normalized residual samples from the test utterance. Each vector is given as input to the AANN model of the target speaker. The squared error (E_i) between the computed output and the target output (which is the same as the input vector) is obtained for each frame. The error E_i of the i^{th} frame is transformed into a confidence value c_i, where $c_i = exp(-\lambda E_i)$. Here $\lambda = 1$ throughout our studies [16]. By transforming the error E_i into a confidence value c_i, we attempt to de-emphasize large error values. The average of the confidence values over all the frames is computed as $c = \frac{1}{N} \sum_{i=1}^{N} c_i$, where N is the number of the frames used in testing. This average confidence values per frame is the score obtained by comparing the test utterance with the target model.

5.4.3 Method of analysis and results

In order to evaluate the effectiveness of the source features, we have conducted both genuine and imposter speaker tests. For consistency and comparison of the performance of the system using different features, the database used is the same as described in Section 3.3. The studies using the source features alone are conducted separately for microphone and telephone speech.

The test utterance is compared against 12 models which consists of 3 models belonging to the target or claimant speaker and 9 other background models. The background models used here are the same as those used in previous systems. The 12 scores obtained are arranged in descending order. If the target model comes in the first and the second position, then the claim is accepted, else it is rejected. Note that the decision logic used here is the same as that used in the other systems. The results of the speaker verification system using source features alone on speech data collected through microphone and telephone is given in Table 5.1.

It can be observed that source features give nearly equal performance for both microphone and telephone speech data. But when compared to spectral features, these features give poorer performance. But since the spectral features, suprasegmental features and source features are independent sources of information, the evidence from these sources can be combined to improve the performance of the system.

Table 5.1: **Results of the text-dependent speaker verification system based on source characteristics for microphone and telephone speech data**

Speech Data	False Acceptance Rate (%)	False Rejection Rate (%)	Equal Error Rate (%)
Microphone	6.78	10.8	8.79
Telephone	7.98	9.55	8.76

5.5 SIGNIFICANCE OF SOURCE INFORMATION FOR TEXT-DEPENDENT SPEAKER VERIFICATION

The source characteristics are extracted from the LP residual of the speech signal. Most of the information related to the text spoken is lost in the residual as the spectral information has been removed. But listening to the residual speech shows that the information about the text in terms of intonation and duration can still be perceived. These features are reflected in longer durations of the signal. The source characteristics that we exploit is the higher order correlations among the samples present in short (5 msec) segments of the residual. This feature may not be dependent on the sound unit or the text spoken. Moreover the training process smears the speech information if it is present. This is because the AANN model is fed with the patterns related to different sound units of the same speaker in succession.

Though the source characteristics extracted from short segments of the residual above are not sufficient for discriminating speakers in a text-dependent mode, they can be used in conjunction with other features that depend on the text information such as the spectral characteristics, intonation and duration information. Since these features are different in their physical meaning, the decision arrived at may be complementary in many cases. Hence the information from different sources may be combined to give better performance.

5.6 SUMMARY

In this chapter, the source characteristics of the speech utterance was examined for the speaker verification task. The complementary nature of the evidence produced by suprasegmental, segmental and source features indicates that the performance of the system may be improved by combining evidence from these features. The efforts made in this direction are described in the next chapter.

CHAPTER 6

SPEAKER VERIFICATION USING SEGMENTAL, SUPRASEGMENTAL AND SOURCE FEATURES

6.1 INTRODUCTION

Generally speaker verification systems use a single type of feature vector and a particular classification procedure to verify the identity claim of the speaker. The uniqueness in the voice of a speaker can be attributed to many features such as the size and shape of the vocal tract, the characteristics of the glottal vibrations, and the speaking style (i.e, the behavioural characteristics such as pitch accent, speaking rate and stress). In chapters 3, 4, and 5 the focus was on using the segmental (short-term spectrum), suprasegmental (intonation and duration) and source features separately for speaker verification. It was observed that individual classifiers could achieve a certain degree of success, but none of them are perfect for practical applications. Though the short-term spectral features gives a strong evidence for speaker verification task, they are sensitive to the characteristics of the transmission channel. The suprasegmental features do not depend on the frequency characteristics of the channel, but they are sensitive to the emotional and physical state of the speaker. This results in large intra-speaker variability in the features. The characteristics of the excitation source which are extracted from short segments of the LP residual signal are not sensitive to the characteristics of the transmission channel, but are affected by additive noise. Studies have shown that the features and classifiers of different types may complement one another in giving better classification performance when used together [17, 60, 64, 77]. Hence we have explored the possibility of combining the evidence from the segmental, suprasegmental and source features for speaker verification.

This chapter is organized as follows: Section 6.2 gives the need for multiple features and classifiers, and the various levels at which the evidence can be combined. Section 6.3 gives the description of the proposed method for combining the evidence at the measurement level. Section 6.4 gives the description of the approach used for combining the evidence from spectral, duration, pitch and source information. The results of analysis are also given in this section. Section 6.5 concludes with a summary of this chapter.

6.2 COMBINING EVIDENCE FROM MULTIPLE CLASSIFIERS

6.2.1 Need for multiple classifiers

Studies have shown that combination of several classifiers will improve the performance of a system [64]. The necessity to use multiple features or classifiers for a pattern recognition problem and to combine the evidence from these classifiers can be understood from the following two reasons:

1. For a specific pattern recognition problem, there are often numerous types of features which are either different in their physical meaning or in their representation. Since it is often difficult to lump these features for a single classifier to make a decision, it is often preferred to use multiple classifiers and then combine the evidence from these.

2. There are a number of classification algorithms that are developed from different theories and methodologies. Each of these are likely to provide complementary evidence in a few cases, which can be combined to obtain an improved performance.

6.2.2 Levels of classifier's output information

Consider a set of class labels $\Lambda = \{c_1, c_2, ..., c_M\}$ and a set of classifiers $k = \{e_1, e_2, ..., e_K\}$. For a given input sample x, the task of a classifier (e_k) is to assign a label $c_i \epsilon \Lambda$ to x. In other words, a classifier can simply be regarded as a function box that receives an input sample x and outputs a label c_i, irrespective of its internal structure, and can be denoted by $e_k(x) = c_i$. Although c_i is the only output information we require at the final stage of classification, practically many of the existing classification algorithms usually supply or are able to supply some other related information. This information can be divided into three levels such as the abstract level, rank level and the measurement level [17].

- The abstract level:

  A classifier e_k gives only a label $c_i \epsilon \Lambda$ as the output.

- The rank level:

  A classifier e_k ranks all the labels in a queue with the label at the top being the first choice.

- The measurement level

  A classifier e_k attributes each label in Λ a measurement value to show the confidence with which x has the label.

Among the three levels, the measurement level contains the highest amount of information and the abstract level contains the lowest. From the measurements attributed to each label, the rank for each label could be determined based on a rank rule (ascending or descending). By choosing the label at the top rank or by using an appropriate decision logic we can assign a unique label to x. In other words, from the measurement level to the abstract level an information reduction process or an abstraction process takes place.

Based on the level of information that is used for combining the evidence, the problem of combining multiple classifiers can be categorized into the following three types:
(a) Type 1: The combination is made based on the output information at the abstract level. Given K individual classifiers, $e_k, \;\; k = 1, ..., K$, each of which assigns the input x to a label c_i, the problem is to use these events to build an integrated classifier E, which gives x one definitive label $c_i \, \epsilon \, \Lambda$ [78].
(b) Type 2: The combination is based on the output information at the rank level. For an input x, each e_k produces a subset $L_k \subseteq \Lambda$, with all the labels in L_k ranked. The problem is to use these events $e_k(x) = L_k, k = 1, ...K$ to build an integrated classifier E, with $E(x) = c_i$, where $c_i \epsilon \, \Lambda$ [64].
(c) Type 3: The combination is made based on the output information at the measurement level. For an input x, each e_k produces a real vector $D(k) = [d_k(1), ...d_k(M)]^t$, where $d_k(i)$ denotes the degree with which e_k considers that x can be assigned the label c_i. The problem is to use these events $e_k = D(k)$ to build a classifier E, with $E(x) = c_i$, $c_i \epsilon \, \Lambda$.

The individual classifiers could be different in their theories and methodologies in the case of Type 1 problems. Hence this approach can be applied to all pattern recognition problems. In the case of Type 3 problems, it is necessary that all the classifiers should be able to supply output information at the measurement level. The measurement vectors should be of the same kind, or it should be possible to transform the vectors into the same kind. A reasonable combination operation on these measurements could be made only when they have the same measure scale.

6.3 COMBINING EVIDENCE AT THE MEASUREMENT LEVEL

The measurement level evidence provides maximum information regarding the output of the classifier. In this study, the classifiers that use different features such as spectral, duration, pitch and source information separately produce outputs which can be represented in the form of a measurement vector. The individual classifiers may have different degree of success, which makes it necessary to utilize the information present at the measurement level to combine the evidence obtained. Since different features produce different strengths of evidence in favour of each class, we have applied the features in a hierarchical manner in the combined system.

The method used for combining the evidence present at the measurement level is based on the technique of *averaged Bayes classifier* described in [17]. This method is suitable in cases where the individual classifiers are Bayes classifiers. However, it can be extended to other classifiers also provided that the measurement vectors can be transformed into a set of post-probability values such that they obey the basic axioms of probability theory. The principle of an average Bayes classifier is given here.

Assuming that the individual classifiers are Bayes classifiers, the classification of an input sample x is based on a set of real value measurements or post-probabilities $P(c_i/x)$, $i = 1, 2, ..., M$. For a Bayes classifier, a definitive decision could be made as,

$$e_k(x) = c_i, \text{ with } P_k(c_i/x) = \max_i P_k\,(c_i/x)$$

To combine the evidence present at the measurement level we do not use the decision made by the individual classifier. Rather, we utilize the measurement values obtained by testing the sample x with all the K classifiers. In order to obtain a new estimation of the probability in favour of each class, for the combined classifier E, the average value is computed as follows:

$$P_E(c_i/x) = \frac{1}{K}\sum_{k=1}^{K} P_k(c_i/x) \tag{6.1}$$

The final decision made by the averaged Bayes classifier E, is based on the newly estimated probabilities, which can be given by,

$$E(x) = c_i, \text{ with } P_E(c_i/x) = max_i\, P_E\,(c_i/x)$$

6.4 AN APPROACH FOR THE COMBINED SYSTEM

In this work we have attempted to combine the evidence from four different classifier which use the spectral, duration, pitch and the source information individually. The features are applied in a hierarchical manner, as these produce different strengths of evidence. The different strengths of evidence can be attributed mainly to the intra-speaker variability, although the representation of information and the classification algorithm play a role in the performance of the system.

Each classifier produces a set of 12 scores by matching the given test utterance with the three target/claimant models and nine background models. The background speakers are the same for each speaker across all the classifiers. The classifier based on spectral information results in a set of 12 scores which can be represented as $S_1, S_2, ..., S_{12}$. Let $D_1, D_2, ..., D_{12}$ be the output scores of the classifier based on duration information. Similarly the output scores of classifier based on pitch and source information are $P_1, P_2, ..., P_{12}$ and $R_1, R_2, ..., R_{12}$, respectively. In order to make them useful for combining, these values are transformed into a set of post-probability values. For example, the output scores of the spectral classifier $S_1, S_2, ..., S_{12}$ can be transformed into $Sp_1, Sp_2, ..., Sp_{12}$ by the following operation,

$$Sp_i = \frac{S_i}{\sum_{i=1}^{12} S_i} \qquad\qquad i = 1, 2, ..., 12 \tag{6.2}$$

Sp_i indicates the probability with which the classifier selects the given speaker i as the true speaker. The same operation can be extended to output scores of other classifier to transform them onto the same scale.

6.4.1 Combining evidence from duration and spectral information

In order to study the effectiveness of incorporating duration information into the basic system, which uses the spectral information, we have used the method of averaged

Bayes classifier as described in previous section. The duration information is applied in a hierarchical manner for the speakers who have been accepted using the spectral information. The output scores of the spectral classifier are converted into a set of post-probability measurements as given in equation (6.2). Similarly, the output scores of the duration classifier $D_1, D_2, ..., D_{12}$ are transformed into $Dp_1, Dp_2, ..., Dp_{12}$ by the same transformation. The averaged probability scores of each of the 12 classes/speakers are computed as follows:

$$A_i = \frac{1}{2}(Sp_i + Dp_i) \qquad i = 1, 2, ..., 12 \tag{6.3}$$

The scores A_i are rearranged in ascending order. The decision logic used is that if the target model is at the first and the second position, and if these models have $S_i < S_{thr}$ and $D_i < D_{thr}$, the speaker is a genuine one, else it is rejected as an imposter speaker. The values S_{thr} and D_{thr} are taken to be the maximum scores obtained by the genuine or target models among the total 450 genuine speaker tests conducted based on spectral information and duration information, respectively. The results of the combined system is analyzed using microphone and telephone speech, and is shown in Table 6.1.

Table 6.1: **Results of the text-dependent speaker verification system based on duration and spectral information for microphone and telephone speech data**

Speech Data	False Acceptance Rate (%)	False Rejection Rate (%)	Equal Error Rate (%)
Microphone	1.09	0.2	0.645
Telephone	1.38	0.6	0.99

Compared to the performance of the basic system given in Table 3.2, it can be seen that the performance of the combined system using duration and spectral information has improved significantly.

6.4.2 Combined system using pitch, duration and spectral features

The pitch information is found to be a weak evidence from the studies shown in Section 4.2.4. This is mainly due to large intra-speaker variability. Hence this feature is applied in a hierarchical manner for the speakers whose identity claim has been accepted using the combined knowledge of duration and spectral information as mentioned in the previous section. Since the pitch information is a weak evidence, we first select a small subset of most probable speakers to which the test utterance belongs. We select the models whose $S_i < S_{thr}$ and $D_i < D_{thr}$, in order to ensure that the target models are present in the reduced list. The number of speakers in the reduced list may vary for each test. Let the number of speakers in the list be N. The measurement vectors corresponding to each classifier is transformed into the post-probabilities. For example, the output of the spectral classifier $S_1, S_2, ..., S_N$ can be transformed into $Sp_1, Sp_2, ..., Sp_N$ by the following operation,

$$Sp_i = \frac{S_i}{\sum_{i=1}^{N} S_i} \qquad i = 1, 2, ..., N \tag{6.4}$$

The duration and pitch based scores are similarly transformed to the same scale as shown above. For this reduced subset, the average probability score is computed as

$$A_i = \frac{1}{3}(Sp_i + Dp_i + Pp_i) \qquad i = 1, 2, ..., N \tag{6.5}$$

The average probability scores are then rearranged in the ascending order. The decision logic is that if the target model comes in the first and the second position, the claim is accepted else it is rejected. The results are shown in Table 6.2. It can be observed that the performance of the system has improved by incorporating the pitch information.

Table 6.2: **Results of the text-dependent speaker verification system based on pitch, duration and spectral information for microphone and telephone speech data**

Speech Data	False Acceptance Rate (%)	False Rejection Rate (%)	Equal Error Rate (%)
Microphone	0.86	0.2	0.53
Telephone	1.03	0.6	0.82

6.4.3 Combined speaker verification system using the source, pitch, duration and spectral information

The output of the classifier based on source information alone results in a set of 12 confidence scores $R_1, R_2, ..., R_{12}$ as mentioned in Section 5.4.3.

These values R_i are then transformed into the post-probability values $Rp_1, Rp_2, ..., Rp_{12}$ to make them suitable for combining. The source information being a weak evidence, it is also applied in a hierarchical manner. Only the speakers who are accepted in the combined system which using spectral, duration and pitch information are further subjected to the test using the source information. The same set of reduced subset selected as mentioned in the previous section is used here. The average probability score of each of the speakers in the reduced subset in computed as

$$A_i = \frac{1}{4}(Sp_i + Dp_i + Pp_i + Rp_i) \qquad i = 1, 2, ..., N \tag{6.6}$$

The A_i's are rearranged in the ascending order and the decision logic of the combined system is that if the target model comes in the first and the second position, the claim is accepted, else it is rejected. The performance of the system for microphone and telephone speech data is shown in Table 6.3. It can be seen that the performance of the system has improved by using the evidence from characteristics of the excitation source of the speech signal.

Table 6.3: **Results of the text-dependent speaker verification system based on source, pitch, duration and spectral information for microphone and telephone speech data**

Speech Data	False Acceptance Rate (%)	False Rejection Rate (%)	Equal Error Rate (%)
Microphone	0.74	0.2	0.47
Telephone	0.86	0.6	0.73

6.4.4 Combining the evidence from measurement level and abstract level output of the classifiers

In addition to the above method of combining evidence from the measurement scores of each classifier, the abstract level information can also be used as an additional check. This additional check is carried out for a speaker who is accepted using the combined knowledge of the four different features. This is done by checking the condition that the speaker should be accepted in at least two of the four individual classifiers which are based on spectral, duration, pitch and source features. The performance of the system obtained for microphone and telephone speech is as shown in Table 6.4. It can be seen that the performance of the system based on telephone speech shows an improvement.

Table 6.4: **Results of the text-dependent speaker verification system based measurement level and the abstract level information from the four classifiers**

Speech Data	False Acceptance Rate (%)	False Rejection Rate (%)	Equal Error Rate (%)
Microphone	0.74	0.2	0.47
Telephone	0.74	0.6	0.67

6.5 SUMMARY

In this chapter, the evidence present in the short-term spectral characteristics, the duration knowledge, the intonation pattern and the source characteristics of the speech signal is combined for speaker verification. Since the features produce different strengths of evidence, the combination is done in a hierarchical manner using the evidence present at the measurement level. The study shows that an improvement in the performance of the system can be achieved by effective combination of the features.

CHAPTER 7

SUMMARY AND CONCLUSIONS

The objective of this work is to develop a robust system for text-dependent speaker verification by combining evidence from multiple features such as short-term spectral features, suprasegmental features (pitch and duration) and source features. The basic system for text-dependent speaker verification uses the traditional approach of template matching using the short-term spectral features, which are represented in the form of cepstral coefficients. The performance of this basic system was analyzed on a speech database which has been collected through telephone as well as microphone for 30 speakers. The system was observed to fail in cases where the end-points were not detected correctly. The performance of the DTW algorithm used for template matching depends critically on the accuracy of the detected end-points. The basic system uses the energy of the speech signal for begin and end detection of the utterance. This approach fails when the speech data is noisy, especially in the case of a telephone speech, where the additive channel noise and high amplitude spikes or glitches are present. Hence, we have used the knowledge of speech features such as vowel onset point for detecting the end-points of the utterance. This method is robust to noisy data to some extent. In order to evaluate the effectiveness of this technique for end-point detection, the basic system was analyzed using this approach for end-point detection. The performance of the basic system, using microphone and telephone speech data, was found to improve significantly. It was observed that the performance of the basic system which uses short-term spectral features, degrades for telephone speech data. This may be due to the fact that the spectral features are sensitive to the frequency characteristics of the transmission channel. Suprasegmental features such as pitch and duration are known to be less affected by the transmission channel characteristics, and these features are ignored in the basic system. It is also interesting to note that human beings implicitly integrate several features for recognizing a speaker. These facts have motivated us to exploit the information present at the suprasegmental level.

A novel method of extracting the duration and intonation information was proposed. The distortion of the dynamic time warping path is used to derive the mismatch in the relative durations of the units in the utterance. The pitch frequency was matched using the information in the matching frames of the test and reference utterances obtained from the DTW algorithm. The speaker verification systems using these features individually were evaluated for the speech data collected. It was observed that these features produce reasonably good performance, although they are weak evidences when compared to the short-term spectral features. The reason could be attributed mainly to the large intra-speaker variability compared to the short-term spectral features, which depend on the shape of the vocal tract. The method used for extracting the duration and pitch information also plays a role in the performance of the system that use these features individually. The performance of the system was analyzed for microphone as well as telephone speech data. In the case of system based on pitch information, a degradation is observed for telephone speech data. This degradation may be attributed to errors in the estimation of pitch. Although the data may be insufficient to draw a solid conclusion, it can be observed that pitch and duration are not critically dependent on the frequency characteristics of the channel.

It is shown that the characteristics of the excitation source is also speaker-specific, and it is different in its physical meaning from the other features such as pitch, duration and spectral information. Source characteristics extracted from the LP residual were modeled using autoassociative neural network. This system was analyzed for microphone and telephone speech data. The studies show that the source features give more or less equal performance for both microphone as well as telephone speech data. Although the suprasegmental and source features provide weak evidence when compared to short-term spectral features, they are less affected by channel characteristics. Since these features provide different sources of information regarding the identity of a speaker, they may complement the evidence in many cases. The evidence from these features was combined in a hierarchical manner, taking into consideration the strength of the evidence produced by each feature. The information present at the measurement level and abstract level was used for combining the evidence. The approach for combining the evidence at the measurement level is based on the principle of average Bayes classifier. The performance of the system was analyzed by incorporating each of these features. The performance of the combined system based on pitch, duration, source and spectral features was found to be better than any feature taken individually. It may also be concluded that with a combined system, the performance for telephone speech can be comparable to that of microphone speech data. Table 7.1 shows the summary of the results obtained for microphone and telephone speech.

7.1 MAJOR CONTRIBUTIONS OF THE PRESENT RESEARCH WORK

- An alternative approach for end-point detection based on the knowledge of the vowel onset points was proposed.
- An effective way of using the pitch and duration information individually for speaker verification was proposed.
- Analysis of the text-dependent speaker verification system based on source features was conducted.
- A hierarchical approach was proposed for combining the evidence from multiple features such as pitch, duration, source and spectral features, using the information present at the measurement level.

Table 7.1: Performance of the text-dependent speaker verification system with microphone and telephone speech. Scores obtained for telephone speech are shown in parentheses

Feature used	False Acceptance Rate (%)	False Rejection Rate (%)	Equal Error Rate (%)
duration information	7.18 (6.49)	6.00 (6.00)	6.59 (6.24)
pitch information	7.64 (7.8)	16.8 (24.4)	12.22 (16.1)
source information	6.78 (7.9)	10.8 (9.55)	8.79 (8.76)
spectral information (energy based end-point detection)	4.9 (5.8)	2.0 (5.7)	3.45 (5.75)
spectral information (VOP based end-point detection)	3.9 (4.19)	0.2 (0.6)	2.05 (2.4)
spectral + duration	1.09 (1.38)	0.2 (0.6)	0.645 (0.99)
spectral + duration + pitch	0.86 (1.03)	0.2 (0.6)	0.53 (0.82)
spectral + duration + pitch + source	0.74 (0.86)	0.2 (0.6)	0.47 (0.73)
spectral + duration + pitch + source +(support from atleast two classifiers)	0.74 (0.74)	0.2 (0.6)	0.47 (0.67)

7.2 SCOPE FOR FUTURE WORK

- Effectiveness of the features for mismatched data conditions need to be studied.
- The performance of the system for remote speaker verification can be studied. Robust pitch extraction algorithm is needed for this study.
- Text-prompted system with multiple hierarchical levels may improve the robustness of the speaker verification system.

APPENDIX A

DYNAMIC TIME WARPING

Time Normalization Constraints: For the alignment process to be meaningful in terms of time normalization for different renditions of an utterance, some constraints on the warping functions are necessary. Typical warping constraints that are considered necessary and reasonable for time alignment between utterances include the following:

1. **Endpoint constraints:** The endpoints of the speech signal are detected usually by a speech-detection operation. For time normalization, the endpoints are the fixed temporal limits of the utterances, leading to a set of constraints for the warping functions of the form,

$$\begin{aligned} x(1) = 1, y(1) = 1 \\ x(K) = X, y(K) = Y \end{aligned} \tag{A.1}$$

   There are K points $C(1), C(2), ..C(k), ...C(K)$, which constitute the warping path, where $C(k) = (x(k), y(k))$. $x(k), y(k)$ are the frames of the test and reference utterance respectively. (We assume that the first frame of the reference as well as test utterance is labeled as frame 1 and the last frame of the test and reference utterance are labeled as frame X and Y, respectively).

2. **Monotonicity constraints:** The temporal order of the spectral sequence in a speech pattern is of crucial importance to linguistic meaning. To maintain the temporal order while performing time normalization, it is necessary to impose constraint of the form

$$
\begin{aligned}
x(k-1) &\le x(k) \\
y(k-1) &\le y(k)
\end{aligned}
\tag{A.2}
$$

3. **Local continuity constraints:** The constraint is used to avoid the omission of any sound segment while time normalization is performed. The constraints are as follows,

$$
\begin{aligned}
x(k) - x(k-1) &= 1 \\
y(k) - y(k-1) &\le 2
\end{aligned}
\tag{A.3}
$$

4. **Global path constraints:** Certain portions of the (x, y) plane are excluded from the region the optimal warping path can traverse due to the local path constraints. Also the fluctuations in speech signal never causes excessive timing difference. So global path constraint is placed to save computation. This is carried out by defining a region in $x - y$ plane, inside which the warping path is allowed to traverse.

Practical Dynamic Time Warping Algorithm

The practical dynamic time warping algorithm is described here. Before applying this algorithm decide upon the local path constraints and the global region. There are three main steps in this algorithm.

- Local distance computation

  Compute the local distances of the grid points in the global region. The local distance between any two frames, $L(x, y)$ is defined as

$$L(x, y) = \sum_{k=1}^{N_{dim}} (a_{x,k} - b_{y,k})^2 \tag{A.4}$$

  N_{dim} is the dimension of the feature vector and $\{a_1, a_2, .., a_k, ., a_X\}$ is the sequence of vectors of the test utterance and $\{b_1, b_2, ., b_k, .., b_X\}$ is the sequence of vectors of the reference utterance.

For the points lying outside the global region a very high local distance score close to infinity is assigned.

- Cumulative distance computation

  The cumulative distance $D(x(k), y(k))$ is the minimum cost to reach from $(1,1)$ to $(x(k), y(k))$. This cost is the sum of all the local distances of the points through which the warp path passes to reach $(x(k), y(k))$ from $(1,1)$. Initialize the cumulative distance $D(1,1)$ with $L(1,1)$. Now for all the other points lying in the global region compute the cumulative distance using the local path constraints, i.e., the point $(x(k), y(k))$ can only be reached from the points $((x(k)-1), y(k))$ or $((x(k)-1), (y(k)-1))$ or $((x(k)-1), (y(k)-2))$. The path to be chosen is as follows.

$$D(x(k), y(k)) = \min \begin{cases} D((x(k)-1), y(k)) + L((x(k), y(k)) \\ D((x(k)-1), (y(k)-1)) + L(x(k), y(k)) \\ D((x(k)-1), (y(k)-2)) + L(x(k), y(k)) \end{cases} \quad \text{(A.5)}$$

  For all the grid points lying outside the global region assign a high value close to infinity. The $D(X, Y)$ gives a measure of similarity between the two speech patterns.

- Backtracking

  Using the cumulative distance matrix and the local path constraints backtrack the path of the warping function from (I, J) to $(1,1)$.

APPENDIX B

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:

$$\begin{aligned} E^{(0)} &= r[0] && \text{(A.7)} \\ k_i &= \frac{r[i] - \sum_{j=1}^{L-1} \alpha_j^{(i-1)} \cdot r[|i-j|]}{E^{(i-1)}} \qquad 1 \leq i \leq p && \text{(A.8)} \\ \alpha_i^i &= k_i && \text{(A.9)} \\ \alpha_j^i &= \alpha_j^{(i-1)} - k_i \cdot \alpha_{i-j}^{(i-1)} && \text{(A.10)} \\ E^{(i)} &= (1 - k_i^2) \cdot E^{(i-1)} && \text{(A.11)} \end{aligned}$$

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.

$$\begin{aligned} c_0 &= \ln \sigma^2 && \text{(A.13)} \\ c_m &= a_m + \sum_{k=1}^{m-1} \frac{k}{m} \cdot c_k \cdot a_{m-k} \qquad 1 \leq m \leq p && \text{(A.14)} \\ &= \sum_{k=1}^{m-1} \frac{k}{m} \cdot c_k \cdot a_{m-k} \qquad\qquad m > p && \text{(A.15)} \end{aligned}$$

BIBLIOGRAPHY

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

[2] Joseph P. Campbell, “Speaker recognition: A tutorial,” *Proc. IEEE*, vol. 85, no. 9, pp. 1436–1462, Sept. 1997.

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

[4] Douglas O’Shaughnessy, “Speaker recognition,” *IEEE ASSP Magazine*, pp. 4–17, 1986.

[5] Aaron E. Rosenberg, “Automatic speaker verification: A review,” *Proc. IEEE*, vol. 64, no. 4, pp. 475–487, Apr. 1976.

[6] A.E. Rosenberg and F.K. Soong, “Recent research in automatic speaker recognition,” in *Advances in Speech signal processing*, Sadaoki Furui and M. Mohan Sondhi, Eds., number 22, chapter 3, pp. 701–740. MARCEL DEKKER, New York, 1992.

[7] Sadaoki Furui, “An overview of speaker recognition technology,” in *Automatic Speech and Speaker Recognition*, Chin-Hui Lee, Frank K. Soong, and Kuldip K. Paliwal, Eds., chapter 2, pp. 31–56. Kluwer Academic, Boston, 1996.

[8] Jayant M. Naik, “Speaker verification: A tutorial,” *IEEE Communications Magazine*, pp. 42–48, Jan. 1990.

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

[10] Sadaoki Furui, “Cepstral analysis technique for automatic speaker verification,” *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 29, no. 2, pp. 254–272, Apr. 1981.

[11] G. Doddington, “Speaker recognition-identifying people from their voices,” *Proc. IEEE*, vol. 73, no. 11, pp. 1651–1664, 1985.

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

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

[14] M. T. Lin C. Y. Tseng S. S. Yu 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*, 1990, pp. 1273–1276.

[15] J. C. Junqua, "The lombard reflex and its role on human listeners and automatic speech recognizer," *J. Acoust. Soc. Amer.*, vol. 93, pp. 510–524, 1993.

[16] B. Yegnanarayana, K. Sharat Reddy, and S. P. Kishore, "Source and system features for speaker recognition using aann models," *Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*, 2001.

[17] A. Kryzak L. Xu and C. Y. Suen, "Methods of combining multiple classifiers and their applications to handwriting recognition," *IEEE Trans. Systems, Man and Cybernetics*, vol. 22, pp. 418–435, May 1992.

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

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

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

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

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

[23] B. S. Atal and S. L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," *J. Acoust. Soc. Amer.*, vol. 50, pp. 637–655, 1971.

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

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

[26] A.E. Rosenberg, Chin-Hui Lee, and F.K. Soong, "Cepstral channel normalisation technique for hmm-based speaker verification," in *Proceedings of Int. Conf. Spoken Language Processing*, 1994, pp. 1835–1838.

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

[28] J. P. Oppenshaw, Z. P. Sun, and J. S. Mason, "A comparison of composite features under degaded speech in speaker recognition," in *Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*, 1993, vol. 2, pp. 371–374.

[29] C. Liu , M. Lin , W. Wang and H. Wang , "Study of line-spectrum pair frequencies for speaker recognition," in *Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*, 1990, pp. 277–280.

[30] H.Hermansky and N.Malayath, "Speaker verification using speaker-specific mapping," in *RLA2C*, Avigon,France, Apr. 1998.

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

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

[33] Hynek Hermansky, "Perceptual linear predictive analysis of speech," *J. Acoust. Soc. Amer.*, vol. 87, no. 4, pp. 1738–1752, 1990.

[34] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 2, no. 4, pp. 578–579, 1994.

[35] 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*, 1986, pp. 877–890.

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

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

[38] J. D. Markel and S. B. Davis, "Text-independent speaker recognition from a large linguistically unconstrained time-spaced data base," .

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

[40] R. W. Schafer and L. R. Rabiner, "System for automatic formant analysis of voiced speech," *J. Acoust. Soc. Amer.*, vol. 47, no. 2, pp. 634–648, 1970.

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

[42] J. D. Markel, B. T. Oshika, and A. H. Gray, "Long-term feature averaging for speaker recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 25, no. 4, pp. 330–337, Aug. 1977.

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

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

[45] Larry Heck Kemal Sonmez, Elizabeth Shriberg and Michael Wintraub, “Modeling dynamic prosodic variation for speaker verification,” *Proceedings of EUROSPEECH’97*, 1997.

[46] S. P. Wagh, “Intonation knowledge based speaker recognition using neural networks,” *M.Tech project report, Dept. of Computer Science and Engineering, IIT Madras, Chennai*, Jan 1994.

[47] B. Yegnanarayana and B. Madhukumar and V. Ramachandran , “Robust features for application in speech and speaker recognition,” in *Proc. ESCA-ETRW on Speech Proc. in AD. CON.*, Cannes, 1992.

[48] B. Yegnanarayana, S. P. Wagh, and S. Rajendran, “A speaker verification system using prosodic features,” in *Proceedings of Int. Conf. Spoken Language Processing*, 1994, pp. 1867–1870.

[49] M. Mathew, “Combining evidendnces from multiple classifiers for text-dependent speaker verification,” *M. S. Thesis, Dept. of Computer Science and Engineering, IIT Madras, Chennai*, June 1999.

[50] S. Rajendran and B. Yegnanarayana, “Word boundary hypothesisation for continous speech in hindi based on f_0 patterns,” *Speech Comm.*, vol. 18, pp. 21–46, 1996.

[51] H. Sakoe, “Two level DP-matching - A dynamic programming based pattern matching algorithm for connected word recognition,” *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-27, pp. 588–595, 1998.

[52] C. C. Tappert and S. K. Das, “Memory and time improvements in dynamic programming algorithm for matching speech patterns,” *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 26, pp. 583–586, 1978.

[53] A. E. Rosenberg L. R. Rabiner and S. E. Levinson, “Considerations in dynamic time warping algorithms for discrete word recognition,” *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 26, pp. 575–582, 1978.

[54] R. M. Gray, “Vector quantization,” *IEEE ASSP Magazine*, pp. 4–29, Apr. 1984.

[55] 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*, 1986, pp. 873–876.

[56] Y. Gong and J.P. Haton, “Nonlinear vector interpolation for speaker recognition,” in *Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*, SanFrancisco, California, USA, Mar. 1992, vol. 2, pp. 173–176.

[57] Y. Bennani and P. Gallinari, “Neural networks for discrimination and modelization of speakers,” *Speech Comm.*, vol. 17, pp. 159–175, 1995.

[58] Y. Bennani and P. Gallinari, “On the use of tdnn-extracted features information in talker identification,” *Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*, pp. 385–388, 1991.

[59] S. P. Kishore, *Speaker Verification Using AutoAssociative Neural Network Models*, MS dissertation, Indian Institute of Technology, Department of Computer Science and Engg., Madras, Dec 2000.

[60] Lan Wang Ke Chen and Huisheng 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.

[61] Z. Sun and J. Mason, "Order analysis of combined features in speaker recognition," ICSP-93 Proceedings, 1993.

[62] T. Matsui and S. Furui, "Text-independent speaker recognition using vocal tract and pitch information," in *Proceedings of Int. Conf. Spoken Language Processing*, Kobe, Japan, Nov. 1990, vol. 1, pp. 137–140.

[63] Sun Mason Department, "Combining features via lda in speaker recognition," citeseer.nj.nec.com/365388.html.

[64] Jonathan J. Hull and Sargur N. Srihari, "Decision combination in multiple classifier systems," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 16, no. 1, pp. 66–75, Jan. 1994.

[65] P. Vermeulen S. Sharma and H. Hermansky, "Combining information from multiple classifiers for speaker verification," in *RLA2C*, Avigon,France, Apr. 1998.

[66] T. V. Ananthapadmanabha and B. Yegnanarayana, "Epoch extraction from linear prediction residual for identification of closed glottis interval," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 27, no. 4, pp. 309–319, 1979.

[67] S. R. M. Prasanna, S. V. Gangashetty, and B. Yegnanarayana, "Significance of vowel onset point for speech analysis," *Signal Processing and Communication (IISc, Bangalore, India)*, July 2001.

[68] J. E. Atkinson, "Inter- and intraspeaker variability in fundamental voice frequency," *J. Acoust. Soc. Amer.*, vol. 60, pp. 440–445, 1976.

[69] A. S. Madhukumar, *Intonation knowledge for Speech Systems for an Indian Language*, Ph. D dissertation, Indian Institute of Technology, 1993.

[70] L. R. Rabiner, "On the use of autocorrelation analysis of pitch detection," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 25, pp. 24–33, 1977.

[71] A. M. Noll, "Cepstrum pitch determination," *J. Acoust. Soc. Amer.*, vol. 41, pp. 293–301, 1967.

[72] J. D. Markel, "The sift algorithm for fundamental frequency estimation," *IEEE Trans. Audio and Electroacoustics*, vol. 20, pp. 367–377, 1972.

[73] David H. Friedman, "Pseudo-maximum-likelihood speech pitch extraction," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 29, no. 2, pp. 254–272, Apr. 1981.

[74] 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, chapter 14. MIT Press, Cambridge, 1986.

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

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

[77] F. Bimbot D. Genoud, G. Gravier and G. Chollet, "Combining methods to improve speaker verification decision," Technical Report IDIAP-RR-96-02, IDIAP, Martginy, Switzerland, 1996.

[78] I. krivonogov D. Mazurov and S. Kazantsev, "Solving optimization and identification problems by the committee methods," *Pattern Recognition*, vol. 20, no. 4, pp. 371–378, 1987.

LIST OF PUBLICATIONS

PRESENTATION IN CONFERENCES

- S. R. Mahadeva Prasanna and Jinu Mariam Zachariah, "Detection of Vowel Onset Point in Speech," to appear in *Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*, (Orlando, U.S.A), May 2002.