EXPLORING FEATURES FOR TEXT-DEPENDENT SPEAKER VERIFICATION IN DISTANT SPEECH SIGNALS

by

B AVINASH

200402006

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

Master of Science (by Research)
in
Computer Science and Engineering

Speech and Vision Lab.

Language Technologies Research Center

International Institute of Information Technology

Hyderabad, India

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY

Hyderabad, India

CERTIFICATE

It is certified that the work contained in this thesis, titled "**Exploring Features for Text-Dependent Speaker Verification in Distant Speech Signals**" by **B Avinash (200402006)**, has been carried out under my supervision and is not submitted elsewhere for a degree.

Date

Adviser: Prof. B. Yegnanarayana

Abstract

Automatic speaker verification (ASV) is the task of verifying a person's claimed identity from his/her voice using a digital computer. The existing ASV systems perform with high accuracy of verification when the speech signal is collected close to the mouth of the speaker ($<$ 1 ft). However, the performance of the ASV systems reduces significantly for speech signals collected at a distance from the speaker (2-6 ft). The objective of this work is to address some research issues in the processing of speech signals collected at a distance from the speaker, for text-dependent ASV system. The distant speech signal is collected using single channel microphone. An acoustic feature derived from short segments of speech signals is proposed for ASV task. The key idea is to exploit the high signal-to-noise nature of short segments of speech in the vicinity of impulse-like excitations. We demonstrate that the proposed feature suffers lesser degradation with distance when compared to the widely used Mel-frequency cepstral coefficients (MFCCs), and also yields better performance of speaker verification than MFCCs. We propose a method of begin-end detection based on the strength of the spectral peaks. A score normalization method is proposed by considering only the robust regions of speech signal. In addition, the regions of speech signal with high signal-to-reverberation ratio are identified, and greater weightage is given to these regions. These modifications are shown to result in a systematic improvement in the performance of the speaker verification system. The use of additional features of duration and pitch is shown to further improve the performance of speaker verification system for distant speech.

Keywords: *automatic speaker verification, text-dependent, distant speech, high signal-to-noise ratio, pitch, duration.*

Contents

List of Tables

List of Figures

Chapter 1

Introduction to speaker recognition systems

Speech is the output of a dynamic vocal tract system, which is excited by a time-varying input. It is a signal produced as a result of several transformations occurring at several levels, namely semantic, linguistic, articulatory and acoustic [1]. It can transmit several classes of information. Speech signal contains information about the text (speech) that is spoken, the language in which it is spoken, the speaker who uttered the text, the gender and the emotional state of the speaker. Speaker recognition is a generic term that refers to any task which discriminates between people based upon their voice characteristics [2]. In other words, recognizing a person solely from his/her voice is known as speaker recognition [3]. *Automatic speaker recognition* is the task of recognizing a person's voice by a machine from the information obtained from his/her speech signal. It is an example of biometric personal identification like face identification, fingerprint identification etc. The advantage of speaker recognition using voice, over the other biometrics lies in the fact that speech is the most widely used means of communication. Also speech signal is easy to collect when compared with other means.

The voices of two persons differ due to physical differences between their vocal organs and the manner in which they use them during speech production [4] i.e.,

anatomical and learned differences. Anatomical differences are the result of variations in sizes and shapes of components of vocal tract: larynx, pharynx, tongue, teeth, oral and nasal cavities. Anatomical differences lead to differences in fundamental frequency, laryngeal source spectrum, formant frequencies and bandwidth. Learned differences lead to variations in dynamics of vocal tract like co-articulation effects and rate of formant transitions [5].

Generally, we use multiple levels of speaker information conveyed in the speech signal. At the lowest level, we recognize a person based on the sound of his/her voice (e.g., low/high pitch, nasality, etc.). But we also use other types of information in the speech signal to recognize a speaker, such as a unique laugh, usage of a particular phrase, or rate of speech, among other factors [2].

In many speech applications, performance of human beings is far better than that of machines. But for speaker recognition, it has been observed that the performance of machines exceeds that of human listeners at times [6]. But human listeners have been robust speaker recognizers when presented with degraded speech.

Though a good performance has been achieved for speaker recognition systems using close-speaking speech signal, it has not been the same for speech signals collected at a distance. Humans on the other hand are able to achieve the same quiet significantly. In addition to channel mismatch (mismatch between training and testing data) and speaker variability, noise and reverberation are the degradations that are introduced in case of speech signals collected at a distance. The state-of-art speaker verification systems use spectral features for verification which get affected by noise and reverberation. Hence, the performance of systems decrease as we are dealing with corrupted features. The objective of this thesis is to address some research issues in the processing of speech signals collected at a distance for text-dependent speaker verification task.

1.1 Speaker recognition principles

Speaker recognition is broadly classified into speaker identification and speaker verification. The speaker identification task is to classify an unlabeled voice token as belonging to one of a set of N reference speakers whereas speaker verification task is to decide whether or not an unlabeled voice token belongs to a specific reference speaker [2]. The fundamental difference between identification and verification is the number of decision alternatives. When it comes to the performance, verification system performance is better than identification system due to the fact that the result is independent of population size. For identification, performance generally degrades with the increase in the number of users.

Speaker recognition can also be categorized based on the usage of text used to extract the features of the speaker. They are text-independent (free-text) and text-dependent (fixed-text) systems. Fixed text systems require recitation of predetermined text whereas free-text systems accept speech utterances of unrestricted text [7]. Fixed-text is better than free-text because with adequate time alignment, one can make reliable comparisons between two utterances of same text which is not possible for free-text systems. Since the main aim in the thesis is in the development of a speaker verification system in distant speech, the general modules of a speaker verification system are explained below.

A general speaker verification system has two stages: (a)Enrollment and (b) verification. Enrollment or training phase is the task of creating models for each speaker. Speaker models refer to the set of features extracted from the speech signal that can distinguish between speakers. The features might correspond to the shape and size of the vocal tract system or represent some higher-level information. Verification or testing phase is the main task where a claimant is declared accepted or rejected. This is done by comparing the claimant reference model and the test utterance. The main modules of a speaker verification system are: (a) feature extraction, (b) pattern comparison and (c) decision logic. Each of these are discussed below.

(a) **Feature extraction**: The main aim of feature extraction is to reduce the dimensionality of the original measurement space while at the same time preserving or enhancing speaker discrimination. In the context of speaker verification, this step involves the processing of speech signal to derive some acoustic parameters.

(b) **Pattern comparison**: The pattern comparison step outputs a matching score after comparing a test feature set and a speaker model.

(c) **Decision Logic**: This is the final step in the verification part. After calculating the matching score from the pattern comparison step, a decision of accept or reject is given to the claimed speaker. It is generally based on thresholds on the scores obtained from the pattern comparison step.

Two types of errors are possible in speaker verification systems: false acceptance of an imposter and false rejection of a genuine. Generally, false acceptance is considered a bigger error than false rejection. A threshold can be set a posteriori so that equal error rate (EER) is achieved. EER corresponds to the threshold at which false acceptance (FA) is equal to the false rejection (FR). A cost function can also represent the error of the system. This is represented as a weighted sum of the FA and FR. The best performance of the system is achieved at the minimum value of the cost function.

Speaker verification needs to handle large as well as small speaker populations and has many applications in the practical world. These can be used in the bank and credit authorizations, access to secure information or premises and carrying out transactions from remote locations by telephone or other voice communication links. The basic assumption of speaker verification systems is that the users of the system are cooperative i.e., the speaker is willing to utter the text consistently during the verification attempts.

1.2 Review of speaker recognition systems

The two main modules of speaker recognition systems are: extraction of features and their classification. This section presents a review of approaches used in speaker recognition. The approaches in extraction of features and classification for general speaker recognition systems are reviewed. This is followed by a review of speaker recognition in distant speech.

1.2.1 Acoustic features for speaker recognition

The features extracted for speaker recognition generally represent the physical attributes like vocal tract system and excitation source, or may represent the high-level features like accent and speaking rate. It is described in [5] that the speaker-specific features should occur naturally and frequently, be easily measurable, not change over time or be affected by speaker's health, background noise or transmission characteristics and not be consciously modifiable by the speaker. The features which exhibit high speaker discrimination power, high inter-speaker variability and low intra-speaker variability are desired [1].

As mentioned earlier, the speaker-specific characteristics are of two types: physical and learned. The vocal tract shape which is a physical characteristic can be estimated from the spectral shape of the voice signal because the frequency content (spectrum) of the acoustic wave is affected as it passes through the vocal tract system. The frequency of oscillation i.e., fundamental frequency depends on the length, tension and mass of the vocal folds. So, it is another characteristic that is physical. Speaking rate, prosodic effects and dialect come under the learned speaker-specific characteristics [1].

Short-term spectral features carry the information regarding the shape and size of the vocal tract. The cepstrum, which is the Fourier transform of the log magnitude spectrum, is commonly used in speaker recognition systems because of its ability to capture formant structure information [8]. It was reported by Atal that LP cepstral

coefficients performed better than LP coefficients and autocorrelation coefficients [3]. For speaker verification systems, spectrum based Mel-frequency cepstral coefficients (MFCC) have been used the state-of-art features [9]. The list of other features include segment durations, formant frequencies, and fundamental frequency. As a part of reducing the effects of channel, cepstral mean subtraction was suggested in [10].

Pitch as a feature in speaker recognition was used in many cases [11, 12]. Earliest use of pitch in speaker recognition was reported in [4] where temporal variations of pitch were used. In [13], speaker identification was performed using the mean and variance of pitch period in voiced sections of an utterance, which contained useful speaker discriminative information. In [14], speaker's F_0 movements were modeled by fitting a piecewise linear model to the F_0 track to obtain a stylized F_0 contour. Though several standard statistical approaches like mean, minimum, maximum and slope of F_0 contour can be applied in the representation of pitch as a feature, the time-dependent information has an advantage. An imposter may be able to mimic average pitch of the speaker but not the variation of pitch as a function of time [15].

1.2.2 Modeling speaker characteristics

There are two ways in which a speaker can be modeled: (a) using stochastic models and (b) using template models. In stochastic models, pattern matching is probabilistic and results in a measure of likelihood of the observation given the model. Pattern matching is deterministic in template models. Pattern matching methods like dynamic time warping (DTW), hidden Markov model (HMM), and vector quantization (VQ) are used for speaker verification whereas Gaussian mixture models (GMM) and artificial neural networks (ANN) are used for speaker identification purposes [1, 16]. Template models are used in DTW, whereas statistical models are used in HMM and codebook models are used in VQ.

1.2.3 Speaker recognition in distant speech

Some attempts have been made in literature about speaker recognition for distant microphone speech. The major areas include compensation of standard features, enhancement of distant speech and use of multiple microphones for recording purposes. In [17], a reverberation compensation method for a GMM based speaker identification system was proposed along with a multiple channel combination and feature normalization. In another study, a beam-forming enhancement was used to improve a single channel GMM based identification system [18]. SVM-GMM supervector, pitch and formant frequency histograms and cross-channel adaptation were used in [19]. Artificially reverberated data was used to train speaker models for reducing channel mismatch [20]. Spectro-temporal features extracted by filtering the speech signal with gamma-tone filter-bank and applying a modulation filter-bank to the temporal envelope of each gamma-tone filter output were proposed in a text-independent speaker identification [21]. A combination of four different subsystems(based on GMM's and SVM's) were used in a text-independent speaker verification [22]. Microphone arrays were used to improve the performance of the system in [23]. Speech enhancement techniques like echo cancellation have also been proposed [24]. A method to compensate channel mismatch by utilizing a room reverberation model was proposed in [25].

1.3 Issues addressed in the thesis

Most of the previous attempts of speaker recognition in distant speech focus on speech enhancement, feature compensation and microphone array methods. Generally, the speech enhancement methods depend on room-transfer functions which are time varying. The feature compensation methods are limited because they try to improve the features which are corrupted. Microphone arrays which use multiple microphones have shown some improvement in the performance but limit the portability of a system and are also expensive.

There is still a need for new features which are robust to distance i.e., whose variation is lesser (or ideally zero) as distance increases, and which retain the speaker information at the same time. The present work focusses on developing a text-dependent speaker verification system based on a single channel microphone speech collected at a distance. The main issues include channel mismatch, speaker variability, noise and reverberation.

1.4 Organization of the thesis

The present work is organized as follows.

In Chapter 2, the characteristics of distant speech are studied along with the description of the database used for performance evaluation. The effect of distance is studied on the variation of a standard feature Mel-frequency cepstal coefficients (MFCC).

Chapter 3 describes the development of the modules of a baseline speaker verification system which can process speech signals collected at a distance.

In Chapter 4, a speaker-specific feature based on short segments of speech signal is proposed which improves the performance of the system in distant speech.

In Chapter 5, performance of the system is enhanced by using frame weights extracted from high signal-to-reverberation component ratio (SRR) regions, duration and pitch information.

Chapter 6 presents a summary of the present work and outlines the scope for further research.

Chapter 2

Characteristics of distant speech

Distant speech (DS) is the speech signal which is collected with a microphone kept at a distance (>1 ft) from the speaker. Close speech (CS) on the other hand, corresponds to the speech signal collected close (<=5 cm) to the speaker. This terminology will be used throughout the thesis unless otherwise specified.

The main difference between CS and DS is the additional effects of noise and reverberation in the case of DS. Signal-to-noise ratio (SNR) is higher for CS and gradually decreases as distance increases because of the background noise. Also due to the decrease in the strength of direct component, the reverberant component will be significant as distance increases. Due to the binaural nature of hearing, human beings overcome the effect of noise and reverberation. But a speech signal collected at a distance is different from what a human perceives.

This chapter attempts to explain a mathematical formulation of distant speech. A distinction between distant speech and close speech is made in Section 2.1. A database collected for the speaker verification system is explained in Section 2.2. Section 2.3 illustrates a study on the variation of a commonly used acoustic feature with distance.

2.1 Mathematical model of distant speech

A mathematical model of a speech signal collected at a distance can be given by

$$y_d[n] = x_d[n] + \sum_{i=1}^{p} \alpha_i x_d[n - n_i] + \vartheta[n] \tag{2.1}$$

where $y_d[n]$ is the signal collected at distance d, $x_d[n]$ is the direct component collected at distant d at time n, α_i is the gain factor and $\vartheta[n]$ is the background noise i.e., distant speech is a summation of the direct component and reverberant component at that point along with the additive background noise.

For close-speech, $x_d[n] \gg \sum_{i=1}^{p} \alpha_i x_d[n - n_i] + \vartheta[n]$, so $y_d[n] \approx x_d[n]$. As the distance increases, the value of $\sum_{i=1}^{p} \alpha_i x_d[n-n_i]+\vartheta[n]$ increases and $x_d[n]$ decreases. Hence the effect of reverberation and noise is more in the case of distant speech.

An illustration of how the magnitude spectrum gets affected because of the distance is given in Fig. 2.1. The figure shows the magnitude spectrum of a segment of speech signal of 30 ms collected at close, 2ft, 4ft and 6ft distance which are time aligned. The main factors, namely, noise and reverberation affects the magnitude spectrum across distance. Noise tends to decrease the dynamic range of the magnitude spectrum whereas reverberation will distort the information due to pitch harmonics.

Since the spectrum is varying, the spectrum based features will also vary as a function of distance. So, by using such features in the speaker verification task, the performance of the system decreases. The performance of the speaker verification system is evaluated on a database which is collected in a lab environment. The data collection process is explained in the next section.

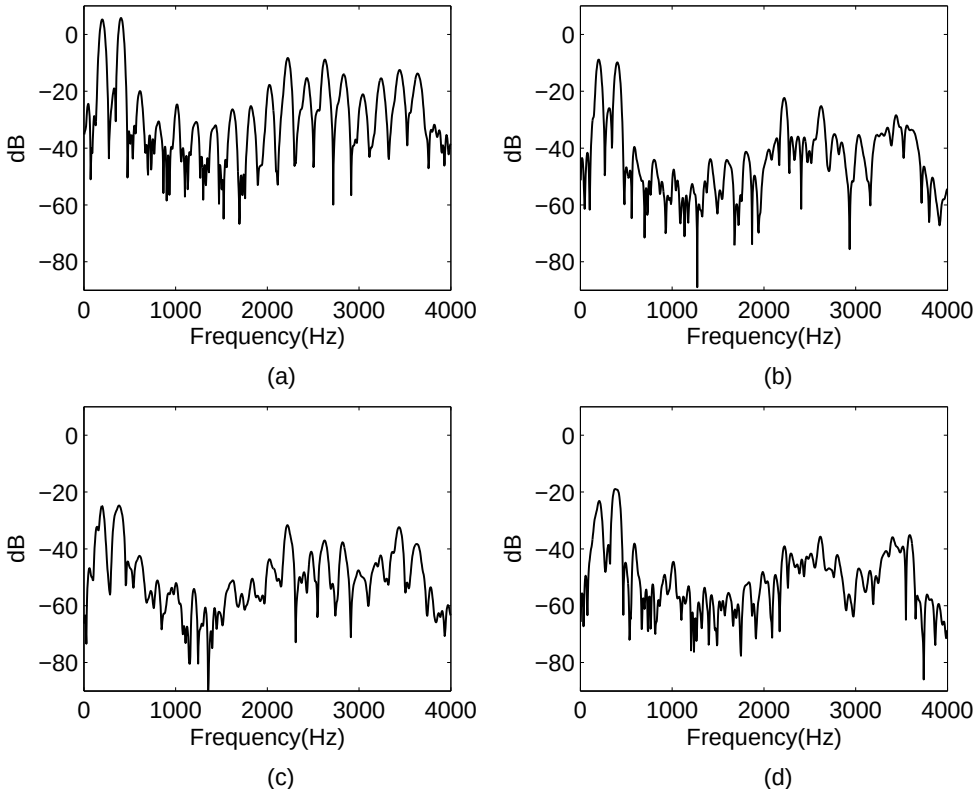

Figure 2.1: Short-time spectra of a segment of time aligned speech signals showing the effect of noise and reverberation. The spectra are derived from (a) close-speaking speech signal, (b) distant speech collected at 2 ft, (c) distant speech collected at 4 ft and (d) distant speech collected at 6 ft.

2.2 Data collection for speaker verification process

A database of 45 speakers is collected in a lab environment with dimensions of 20ft $\times$ 15ft $\times$ 10ft. The data is collected using 30 male and 15 female speakers. The microphones are placed on a plastic rod at distances of 4-5 cms, 2 ft, 4 ft and 6 ft from the position of the speaker. The microphones used are 4 omni-directional microphones. The plastic rod is placed such that the speaker is at the same horizontal level as the microphone setup. The environment has noise from the computer systems and fans. The same environment is used for all the speakers. An illustration of the microphone setup is shown in Fig. 2.2. The points 1, 2, 3 and 4 correspond to the close (4-5 cms), 2 ft, 4 ft and 6 ft microphone respectively and the speaker is placed just before the close-speaking microphone. Two Edirol

Table 2.1: Summary of the data collection process.

Session	No: spkrs	No: repetitions	Collected in	Total # of utterances
Session 1	45	5	April 2009	45×5 = 225
Session 2	45	10	June 2009	45×10 = 450
Session 3	45	10	October 2009	45×10 = 450

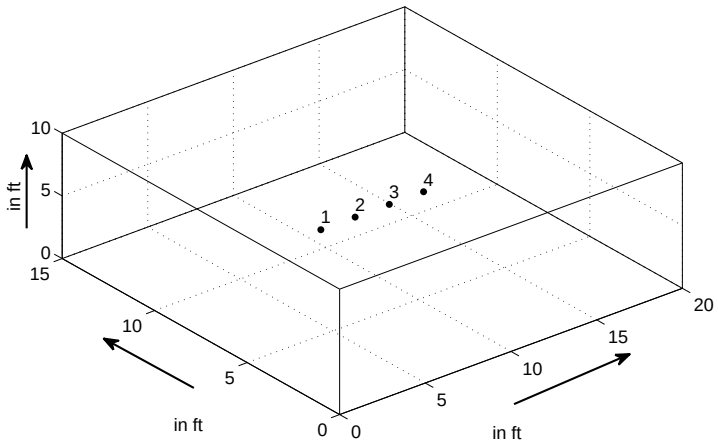

Figure 2.2: Microphone setup used in the data collection process.

[26] devices are used to collect the data simultaneously at all the distances which is recorded at 48000 Hz and stored as 16 bits/sample. However the speech signals are resampled to 8000 Hz for processing purposes. The sentence used as the text in the data collection process is "We were away a year ago". The data is collected in three different sessions within a span of 7 months. The number of repetitions of the sentence varied from session to session. It is five, ten and ten repetitions for the first, second and third session respectively. A summary of the data collection process is given in Table. 2.1

The total number of repetitions of the sentence by each speaker is 25. Three utterances of session 1 are used for enrolling purposes. Twenty utterances of session 2 and session 3 are used for verification. So, the total number of genuine tests for 45 speakers will be 20×45 = 900. A speaker can be used as an imposter for the remaining 44 speakers. So, the number of imposter tests will be 20×45×44 = 39600. This speech data is used for the development of a baseline speaker verification system and for experiments to enhance the performance of the system.

2.3 Variations in the feature MFCC across distance

In this section, we examine the variation of an acoustic feature with distance. We consider the Mel-frequency cepstral coefficients (MFCC), which are a standard feature used in many speech and speaker recognition applications [9, 27]. To illustrate that this well known feature varies across distance, a study is made which is explained below.

20 MFCC coefficients are extracted for each frame with a size of 20 ms and a shift of 5 ms. 25 triangular filters are used within the range of 0-3300 Hz. The speech signals used are sampled at 8000 Hz. The extraction process of MFCC feature is explained in detail in Section 3.1.1. A direct comparison of a close speaking and distant speaking microphone is made on the MFCC features after the two signals are time-aligned. Euclidean distance is calculated between frames as

$$\chi(i) = ||\mathbf{x_{c,i}} - \mathbf{x_{d,i}}|| \tag{2.2}$$

where $\mathbf{x_{c,i}}$ and $\mathbf{x_{d,i}}$ represent the unit vectors of the feature vectors for frame i, of close and distant speech respectively.

A framewise variation of the MFCC feature with respect to distance for time-aligned speech signals is shown in Fig. 2.3. Fig. 2.3(a), (b) and (c) shows the variation between close - 2 ft, close - 4 ft and close - 6 ft case, respectively. Notice the increase in the variation of the Euclidean distance as a function of distance.

For each distance d, mean χ_d is calculated as

$$\chi_d = \frac{1}{N}\sum_{i=1}^{N}\chi(i) \tag{2.3}$$

where N is the total number of frames.

The mean χ_d reflects the deviation of the feature as a function of distance.

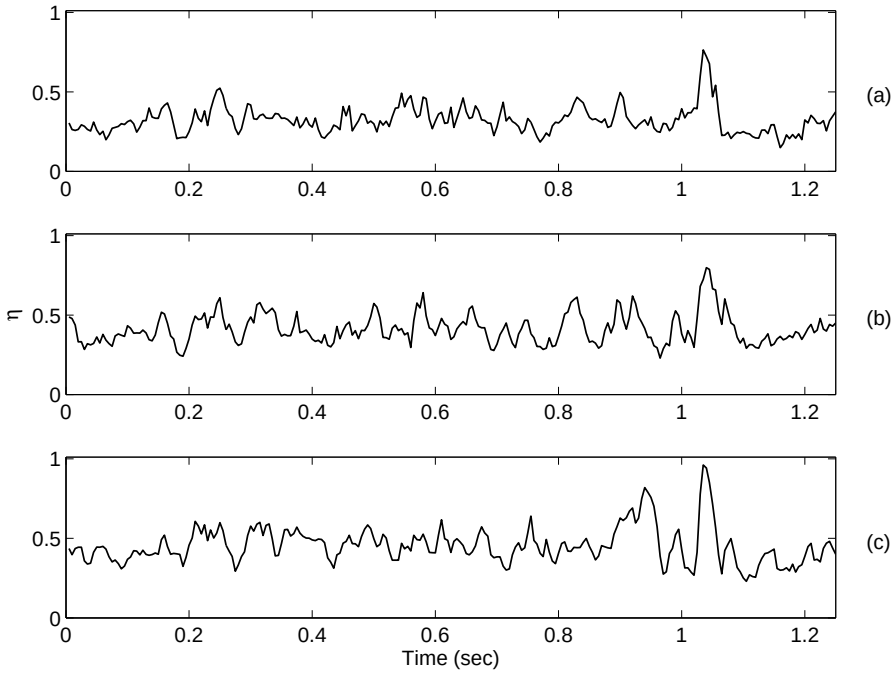

Figure 2.3: Effect of distance on MFCC feature. (a) Variation between close-speaking signal and distant speech collected at 2 ft, (b) variation between close-speaking signal and distant speech collected at 4 ft and (c) variation between close-speaking signal and distant speech collected at 6 ft.

Fig. 2.4 shows the average variation of the feature MFCC across speakers as a function of distance. The number of speakers used is 10 and the number of utterances used per speaker is 5. Only voiced frames are considered in the calculation of average χ_d. It can be seen from the figure that the mean χ_d increases with the distance, indicating that the MFCC features are varying. This underlines the importance of deriving robust features from distant speech, which do not vary significantly with distance.

2.4 Conclusions

This chapter has given a background on the main differences between distant speech and close speaking speech. We have seen that there is a variation with distance in the standard features like MFCC. The next chapter describes the evolution of a baseline system which attempts to address some of the issues posed by distant speech.

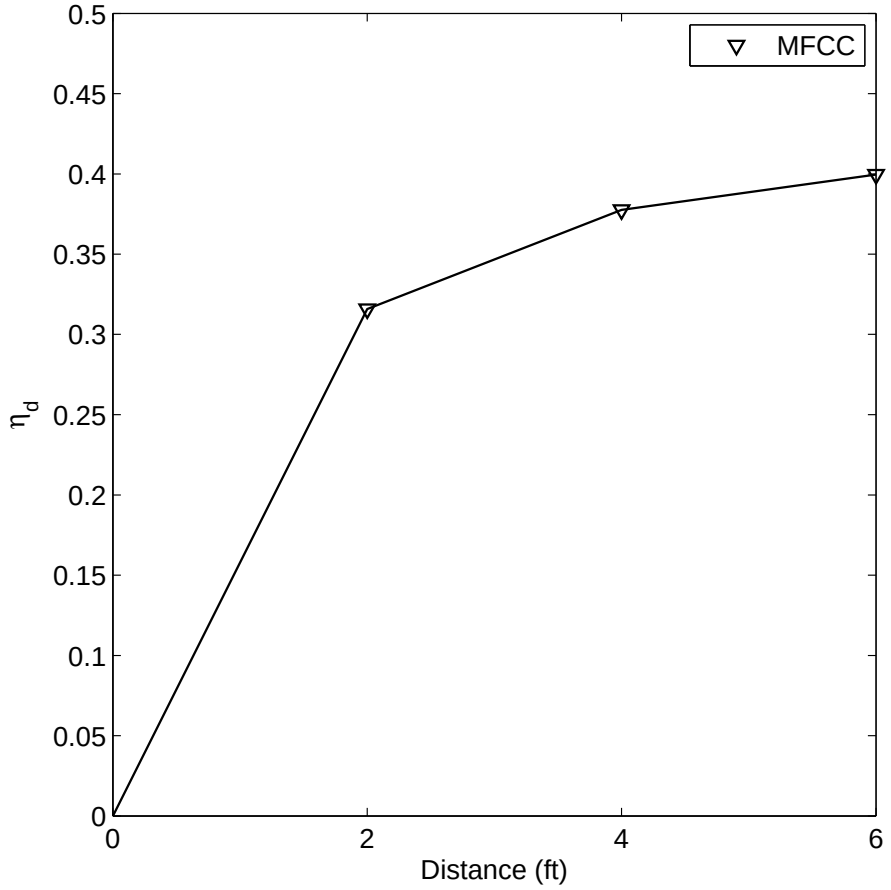

Figure 2.4: Variation of acoustic feature (MFCC) with distance.

Chapter 3

Baseline speaker verification system

A text-dependent speaker verification system based on MFCC features is developed and used as a baseline system. The text used is "We were away a year ago". Enrollment phase uses three utterances of the speaker to create three speaker models for each speaker. Each speaker will get a unique ID after enrolling in the system which represents his claim information. Verification stage compares a test utterance with the three reference models of the claimant along with background models and results in a binary decision of accepted or rejected. The basic flow of the system is given in Fig. 3.1. This chapter deals with the development of the baseline system for distant speech.

The chapter is organized as follows. Section 3.1 deals with the components of the baseline system. The improvements made in each component, to handle distant speech is also explained. In Section 3.2, the performance of the baseline system is evaluated for the database. Section 3.3 concludes the present chapter.

3.1 Components of the system

The main components of a speaker verification system are

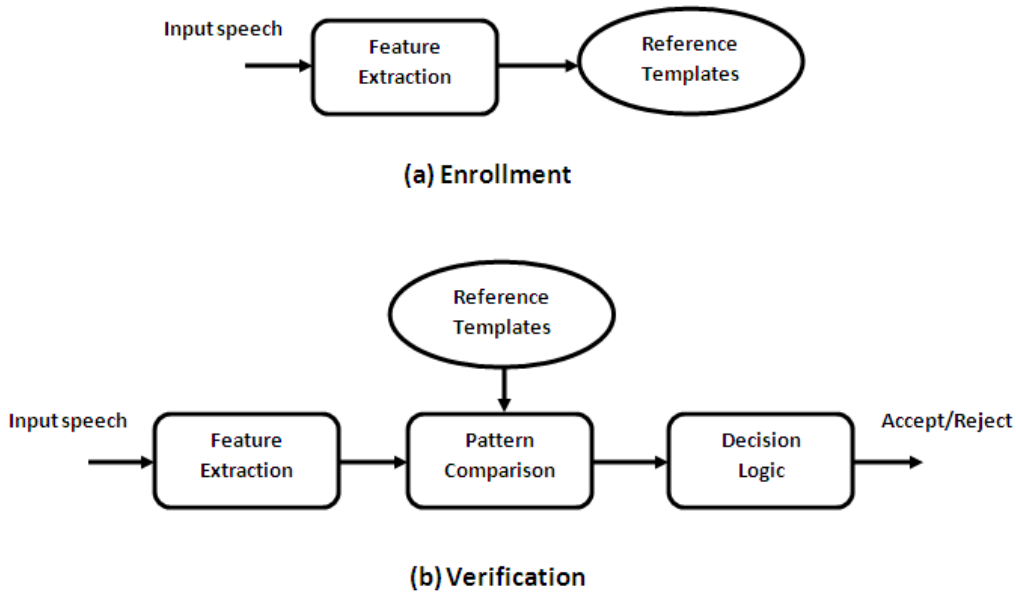

Figure 3.1: The flow of the speaker verification system. (a) Enrollment and (b) verification.

- Extraction of acoustic features
- Pattern comparison
- Decision logic

This section explains the components of the system in detail.

3.1.1 Extraction of acoustic features for speaker verification

The main aim of feature extraction module is to reduce the dimensionality of the input speech signal while at the same time preserving or enhancing speaker discrimination. Feature extraction generally involves a begin-end detection as a preprocessing step, which identifies the regions of speech activity. Existing methods for begin-end detection are based on the energy of the signal. Such methods do not work satisfactorily in distant speech, due to the presence of degradations. We propose a method for begin-end detection which works better for speech signals collected at a

distance. It is based on the strength of spectral peaks in the short-time spectrum of voiced speech. The method is described below.

Begin-End detection based on the strength of spectral peaks

Low signal-to-noise ratio (SNR) and reverberation are the main issues in the processing of distant speech signals. Standard methods for begin-end detection are based on energy of the signal but the performance degrades with distance. Low SNR also affects the spectrum in voiced regions by reducing the dynamic range of the spectrum. But the peaks in the magnitude of the short-time spectrum are less affected by noise. These peaks correspond to the locations of formants and pitch harmonics. Such peaks are not present in the non-speech regions. Fig. 3.2(a) shows a narrow band spectrogram of a speech signal collected at 6 ft where the spectral peaks (i.e., the dark regions in the spectrogram) are much more prominent in the voiced regions than in the non-speech regions. So, the strength of peaks of short-time spectrum is a good indication of voicing even in case of distant speech. Reverberation slightly distorts a few pitch harmonics, but not all the peaks. Hence only few strong peaks are taken into consideration. The steps involved in begin-end detection using this feature are described below.

1. Narrow band spectrogram of the speech signal is computed with a frame size of 30 ms and frame shift of 5 ms. Each frame in the spectrogram corresponds to the logarithm of the magnitude of the short-time spectrum, which is given by.

$$E_l = \log_{10}|X_l[k]|^2, \;\; k = 0, ..., N-1, \tag{3.1}$$

where $X_l[k]$ is the discrete Fourier transform (DFT) of the frame $x_l[n]$, and is given by

$$X_l[k] = \sum_{n=0}^{N-1} x_l[n] e^{-j\frac{2\pi}{N}nk}, \; k = 0, 1, ..N-1 \tag{3.2}$$

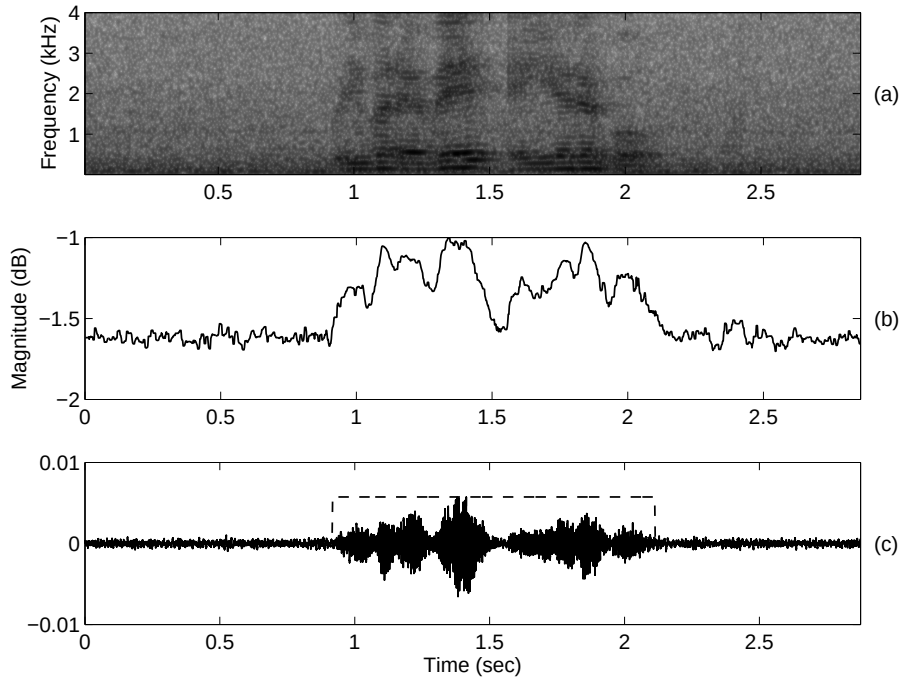

Figure 3.2: Begin-end detection for distant speech based on peaks of spectrogram. (a) Spectrogram of speech signal, (b) mean of magnitude of peaks extracted from short-time spectrum and (c) begin-end points plotted over speech signal collected at 6 ft distance.

Here, l indicates the frame number and N is the number of points in the DFT. For frames of 30 ms at a sampling frequency of 8000 Hz, $N = 512$ is used.

2. The peaks are then identified for each frame of the log magnitude spectrum. Let

$$\delta_l[k] = E_l[k] - E_l[k-1], \;\; k = 0, 1...N-1 \tag{3.3}$$

A positive value of δ_l corresponds to an increasing curve at that location and a negative value of δ_l corresponds to a decreasing curve of E_l. Thus, the locations of positive-to-negative zero crossings in δ_l indicates a peak location. All such peak locations are identified.

3. The mean of magnitudes of the strongest P peaks is calculated for each frame. This energy measure is used to select voiced regions in distant speech. A value of $P = 15$ is used in the system.

An automatic threshold is extracted from the energy measure which is then used to select the begin-end points. Fig. 3.2 shows a speech signal collected at 6 ft, along with the spectrogram of the signal and the energy measure based on the strength of spectral peaks. The speech region detected using the energy measure is also shown.

Extraction of short-time spectral features

As mentioned earlier, Mel-frequency cepstral coefficients (MFCC) have been the state-of-art features that are used to capture speaker-specific information [9, 28, 27]. It is a representation of a short-term power spectrum of a signal based on a linear cosine transform of a log power spectrum. This power spectrum is computed on a nonlinear Mel scale of frequency. Mel scale [29] is guided by the human auditory processing of speech signal by giving more importance to the low frequency part of the spectrum than the higher frequency part. The baseline system uses MFCC as the feature to capture the speaker-specific information. The speech signal after the begin-end detection is used to extract the short-time spectral features. The conventional MFCC extraction algorithm is explained below[30, 31].

1. The speech signal after the begin-end detection is sampled at 8000 Hz and separated into frames which are segments of 20 ms. An overlap of 5 ms between two adjacent frames is used.

2. Each frame $s[n]$ of 20 ms is multiplied by a 160-point Hamming window $w[n]$, to reduce the effect of abrupt truncation of the frame $s[n]$. The windowed signal is given by

$$x[n] = s[n]w[n], \tag{3.4}$$

where the Hamming window is given by

$$\begin{aligned} w[n] &= 0.54 - 0.46\cos[\frac{2\pi n}{N-1}], n = 0, 1, ...N-1 \\ &= 0, \text{otherwise} \end{aligned} \tag{3.5}$$

where $N = 160$ denotes the number of samples in one frame. The log magni-

tude spectrum E_l of each frame $x[n]$ is calculated which is given in equation 3.1.

3. The log magnitude spectrum is then filtered using 25 triangular shaped band-pass filters which are centered on equally spaced frequencies in Mel domain between 0 Hz and 3300 Hz. The linear frequency (f) to Mel-frequency (m) mapping is given in equation 3.6. Log energy output Q_k of each filter is then calculated.

$$m = 1127\log_e(1 + \frac{f}{700}) \tag{3.6}$$

Mel-frequency cepstral coefficients are computed as

$$C_i = \sum_{k=1}^{25} Q_k cos[i(k - \frac{1}{2})\frac{\pi}{25}], \;\; i = 1, 2, ...M, \tag{3.7}$$

where M is the number of cepstral coefficients and Q_k is the log-energy output of the k^{th} filter. We have used $M = 20$ for the baseline system.

To reduce the effect of channel variations, cepstral mean subtraction (CMS) is performed after extracting the coefficients [10]. So, the feature used for the baseline system is MFCC + CMS.

3.1.2 Pattern Comparison

A template-based pattern comparison approach, namely dynamic time warping (DTW) algorithm is used in the system. Dynamic time warping is an algorithm used for measuring optimal match between two sequences which may vary in time or speed. The sequences are warped non-linearly in the time dimension to determine a measure of their similarity independent of certain non-linear variations in the time dimension. It was developed for isolated word recognition application [32] and was adapted by Furui for text-dependent speaker verification in [10]. For speaker verification task, this approach compares the acoustic features derived from the speech

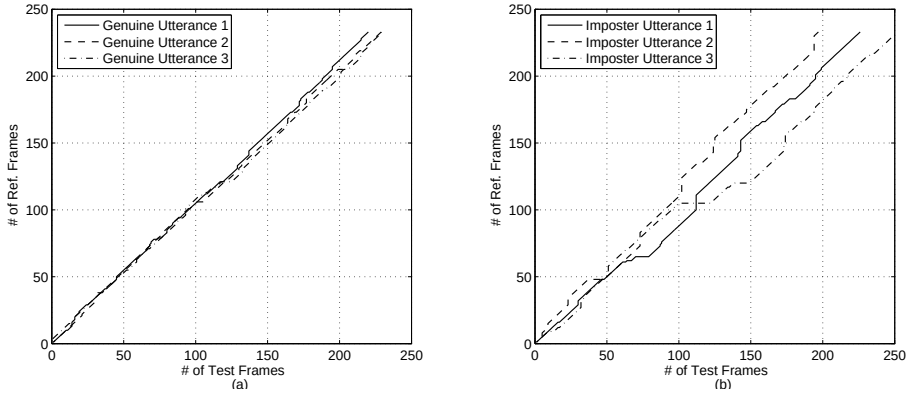

Figure 3.3: Optimal warping path of three different test utterances matched against a reference template. (a) Genuine case and (b) imposter case.

signal collected during enrollment, with the acoustic features derived from the speech signal collected during verification. The result of this comparison is a dissimilarity measure. It has been used in many text-dependent speaker verification applications [33, 34, 35]. A detailed explanation of DTW algorithm is given below.

Let us consider two sequences X and Y, each consisting of differing number of feature vectors, given by

$$X = \mathbf{x_1}, \mathbf{x_2}...\mathbf{x_i}...\mathbf{x_M} \tag{3.8}$$

$$Y = \mathbf{y_1}, \mathbf{y_2}...\mathbf{y_j}...\mathbf{y_N} \tag{3.9}$$

Here X and Y represent sequences of feature vectors derived from the test utterance and reference utterance, respectively.

1. A Euclidean distance is computed for each set of frames between the feature vectors.

$$d(i,j) = ||\mathbf{x_i} - \mathbf{y_j}||, \quad i = 1, 2, .., M, j = 1, 2, .., N \tag{3.10}$$

2. A cumulative distance D is calculated from the Euclidean distance values

$d(i,j)$ as follows:

$$D(i,j) = d(i,j) + \min([D(i-1,j), D(i-1,j-1), D(i,j-1)]) \quad (3.11)$$

with an initial condition of

$$\begin{aligned} D(1,1) &= d(1,1) \\ D(i,1) &= d(i,1) + D(i-1,1) \\ D(1,j) &= d(1,j) + D(1,j-1) \end{aligned} \quad (3.12)$$

$D(M,N)$ gives a dissimilarity measure between the two vectors. An optimal warping path can be found from the cumulative distance matrix D by backtracking from (M,N) to $(1,1)$ using the same constraints used for calculating the cumulative distance matrix. The points in the optimal warping path represent the best mapping between the test and reference feature vectors, and are used in the decision logic for score normalization and other purposes. An example of the optimal warping paths for genuine and imposter cases are shown in Fig. 3.3. The scores obtained for the genuine case are generally lesser when compared to the imposter case. Notice that the optimal warping path for the genuine case traverses almost along the diagonal path and obtains a small score showing the similarity between the feature vectors. On the other hand, though the score obtained for the imposter case is higher, optimal warping path grossly traverses along the diagonal path because the text used is the same.

The advantage of DTW algorithm lies in the fact that it eliminates the differences in time between the two sequences by nonlinear warping. But the disadvantage of DTW algorithm is the large storage requirement and computation time.

3.1.3 Decision Logic

The performance of the system is affected by the decision logic module, since it generates the decision to accept/reject the claim based on the scores obtained from

the pattern comparison step. Two important techniques can be used for decision logic: (a) A threshold based technique and (b) a technique based on background speakers. The first technique applies a threshold on the scores generated during the pattern comparison module (generally cumulative score), to declare the claim as accepted or rejected. The second technique compares the scores of the claimant with the scores of background speakers to declare the claim as accept/reject. A background speaker can be any other speaker other than the claim.

In the case of distant speech, performance of the DTW algorithm is affected because of the variation in channel, speaker and distance. The first technique, using a cumulative score to verify a speaker, is not a good measure since it is affected by the change in the feature vectors. However, the gross temporal characteristics are preserved in the sequence of feature vectors, and this is reflected in the optimal warping path obtained from DTW. This can be used in the second technique to declare a claim as accept/reject. Also, the set of scores obtained will differ from one speaker to another speaker, because of the variability in the speaker from one session to other. There is a need for score normalization which will attempt to reduce these variabilities so that the ranges of the scores are uniform among different speakers and a common threshold can be set for the given speaker verification system. The score normalization technique used for the baseline system is explained below.

Score normalization

The methods of score normalization can be classified as model normalization and test utterance normalization. In model normalization, a speaker's model is tested against example imposter utterances and the resulting scores are used to estimate speakers-specific statistics. In test utterance normalization, the test utterance is compared against the model of a claimant speaker, and also against background models. The scores of background speakers are used to normalize the speaker's score for that utterance. Test utterance based normalization of scores is used in the baseline system which is explained below.

As mentioned earlier, the cumulative score is affected because of the channel, speaker variations and distance (noise and reverberation). It can be observed that the number of frames which gets affected by distance increases as the distance increases because of the effect of noise and reverberation in feature extraction. So, instead of using the cumulative score, it is advantageous to use those frames which are less affected by the above mentioned factors. It is accomplished with the help of framewise Euclidean scores which is a function of test frames. The main idea is to use the fact that the variation of acoustic features of a genuine speaker is lesser, when compared with a background speaker even in degradations. The framewise Euclidean scores are obtained from the optimal warping path, which gives the mapping between the reference and test frames. Optimal warping path is a non-linear mapping where one test frame can be mapped to one or more reference frames and vice versa. The procedure for calculation of framewise Euclidean scores explained below.

Given $(k, h(k))$ for $k = 1, 2, ..K$ where K represents the number of points in the optimal warping path, $d(k, h(k))$ gives the Euclidean distance at that point. If the total number of test frames is L, the framewise Euclidean score $f(i)$ for the i^{th} frame is

$$f(i) = \min(d(k, h(k))) \ \forall k = i \tag{3.13}$$

and i ranges from 1 to L.

The sum of framewise Euclidean scores of the function f will approximately be equal to the cumulative score.

An illustration of the framewise Euclidean scores is shown in the Fig. 3.4 where a test utterance is compared with one reference utterance and three background utterances. Fig. 3.4(a) and Fig. 3.4(b) are the cases where the test utterance belongs to a genuine speaker and imposter speakers, respectively. From Fig. 3.4(a), we observe that the scores of the genuine test utterance are lesser than that of utterances of the background speakers for most of the frames. No such pattern can be observed

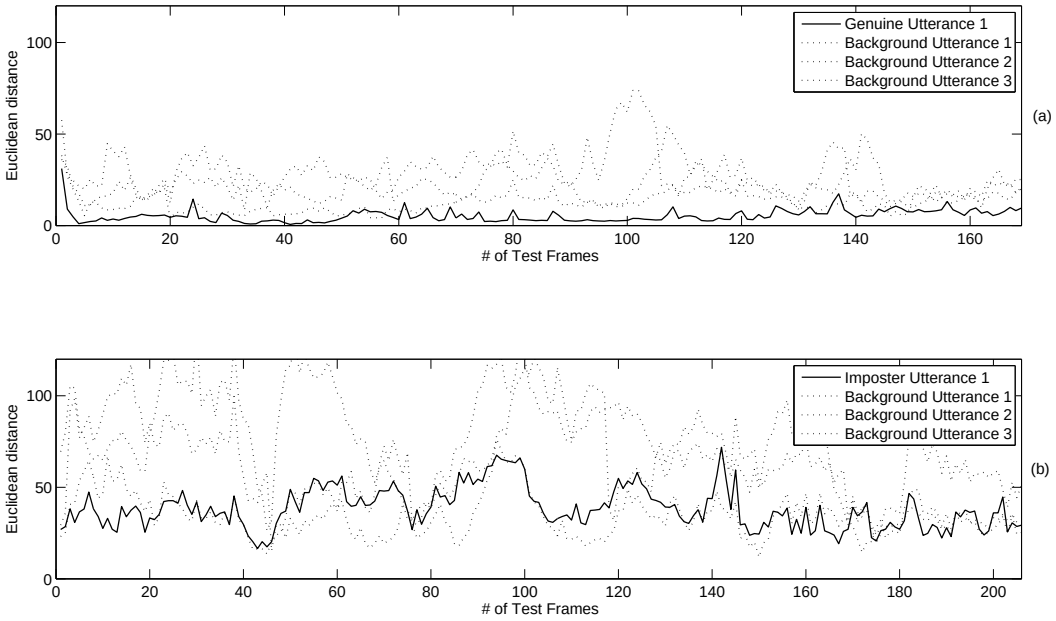

Figure 3.4: Frame aligned Euclidean scores of the claim. (a) Genuine case and (b) imposter case.

in the case of imposter utterance (Fig. 3.4(b)). Also, the scores in each frame are lesser for genuine utterance than imposter utterance. A genuine claim can be detected using these two observations from the framewise Euclidean scores. Since a single reference utterance cannot exactly characterize a speaker, three reference utterances are used in the baseline system. The extraction of information from the framewise Euclidean scores is explained below.

Consider the framewise Euclidean scores for the M claimant reference models and N background models, denoted by f_k where $k = 1, 2...(M+N)$. The N background models are chosen randomly from the available list of speakers. Each frame of the test utterance is assigned to that model which has the least score in that frame. This decision is based on the $M + N$ scores available for each frame. This results in a frame-to-model mapping g which gives the information about the relationship between a frame and a model. The expression for obtaining the frame-to-model mapping is

$$g(i) = \underset{k}{\operatorname{argmin}}(f_k(i)) \;\; \forall k = 1, 2...(M+N) \tag{3.14}$$

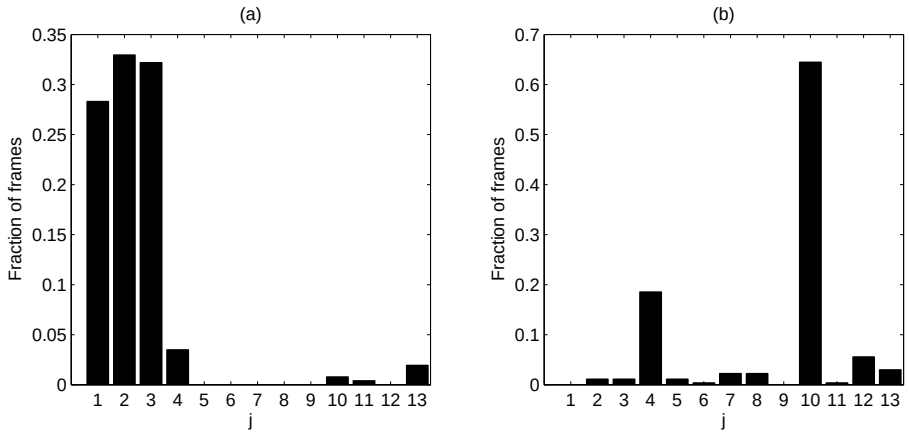

Figure 3.5: Normalized histogram showing the fraction of frames for each reference template. (a) Genuine case and (b) imposter case.

where i represents the frame number and $i = 1, 2...L$.

A frame i is said to be assigned to j^{th} reference model if $g(i) = j$ and $j <= M$. Similarly the frame i is said to be assigned to j^{th} background model if $g(i) = M + j$. For the baseline system, $M = 3$ and $N = 10$, that is, 3 references models and 10 background models are used.

A histogram can be plotted with bars representing the number of frames assigned to a particular model. Examples of such histograms are shown in Fig. 3.5 with $M = 3$ and $N = 10$. Fig. 3.5(a) represents the genuine case and fig. 3.5(b) represents the imposter case. It can be seen that most of the frames are assigned to the reference models in the genuine case, where as no such pattern is found in the imposter case. That is, the number of frames allocated to reference models is similar to that allocated to the background models for imposter case. Using the histograms and the frame-to-model mapping we have developed two decision-making mechanisms which are described below.

- **Difference between the number of frames of claimant and the background (α)**: The main task of score normalization is to create a measure which distinguishes the genuine speaker from the imposters. The pattern which can be observed from Fig. 3.4 is that the discrimination between the claim and background is better in case of genuine speaker than the imposters.

So, to capture the discrimination, the difference between the sum of frames assigned to the claimant and the sum of top k background models is computed. This value is divided by the number of frames so that the normalized scores range between -1 and 1. This is a confidence score for speaker verification purposes i.e., a high score indicates a greater chance of getting accepted.

$$\alpha = \frac{A - B}{L} \tag{3.15}$$

where A is sum of the number of frames assigned to the claimant models, B is sum of the number of frames assigned to the top k background models and L is the total number of frames of the test utterance.

- **Stable score (β)**: When the test speech utterance is collected at a distance, the percentage of frames affected by noise, reverberation and channel is more than that in the case of close-speaking speech. These factors affect the computation of cumulative score. Hence, it is advantageous to consider only those frames which are not affected much by the above mentioned factors. A threshold applied on the framewise Euclidean scores to select the frames does not provide a solution because there exists a variability between different utterances of the same speaker. Hence, we use the background speakers to select the frames. Also, the percentage of frames assigned to the claimant is an indication of the genuineness of the claim. So, a score is computed as the ratio of the sum of scores of frames assigned to the claimant and the fraction of the frames assigned to the claimant.

$$\beta = \frac{\mathrm{C}}{\mathrm{D}} \tag{3.16}$$

where C is the sum of scores of the frames assigned to the claimant and D is the fraction of the frames assigned to the claimant. This score (β) is a dissimilarity score i.e., the low score indicates a greater chance of getting accepted.

A combination of the above two scores is used for decision making during speaker verification.

3.2 Performance evaluation on close and distant speech

The database described in Chapter 2 is used for the evaluation process. As explained earlier, for each channel, the number of genuine test cases and imposter tests cases are 900 and 39600 respectively. A combination of the scores of α and β is used as a measure for discriminating between genuine and imposter claims. The α value is a confidence measure and β is a dissimilarity measure. Given the 45000 α and β values, for calculating the combination of the two measures, the measure β is converted into a confidence measure by subtracting each value with its maximum. All the scores of α and β are then normalized in the range of 0 to 1. A weighted sum of α and β is used as the actual measure for verifying speakers and is given by

$$S = 0.1 \times \alpha + 0.9 \times \beta \tag{3.17}$$

The weights of 0.1 and 0.9 are calculated empirically. Since it is a confidence measure, the higher scores indicate greater chances of getting accepted.

The cost function used for the system is given by

$$\eta_C = 0.1 \times \eta_R + 0.9 \times \eta_A, \tag{3.18}$$

where η_R and η_A denote the false rejection and the false acceptance rates respectively. The same cost function is used in later sections unless specified.

The threshold that yields the least cost for close-speaking speech signals is set as the threshold of the baseline speaker verification system. The same threshold is used for the speech signals collected at distances of 2 ft, 4 ft and 6 ft. Fig. 3.6 shows the cost function of the baseline system for each distance as a function of threshold. It can be observed from the plot that the minimum value of the cost function increases as the distance increases. If the variation of the feature is less, then the cost functions would be similar for all the distances. Due to the variation of

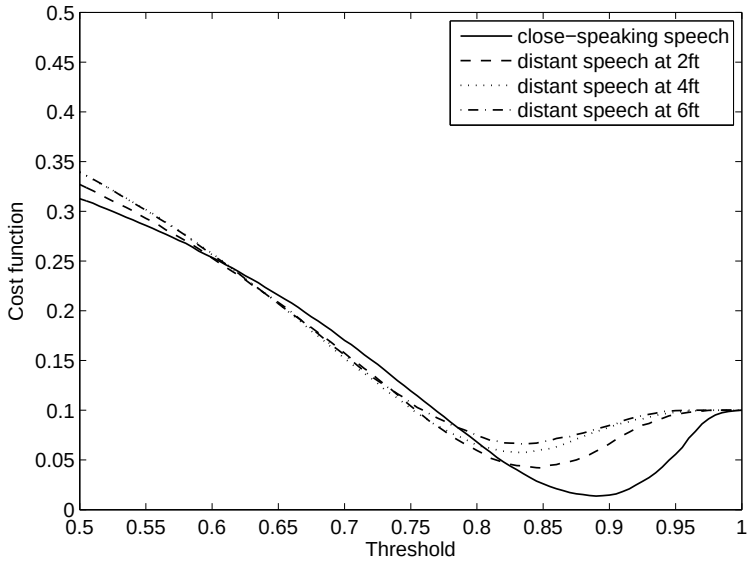

Figure 3.6: Cost function of the system using MFCC + CMS as feature.

the acoustic feature w.r.t distance, there is a decrease in the confidence scores, which results in a reduction of the performance of the system. The minimum cost for each distance is not the actual cost of the system at that distance since the threshold chosen for the system corresponds to a minimum cost of the close-speaking speech signals. Moreover, distance is not a parameter used during the verification phase. With the available features, an additional information of distance will increase the performance of the system but finding distance from the speech signal is a difficult task.

Table. 3.1 shows the performance of the baseline system. The term *best cost* shows the minimum cost function value for each distance. *System cost* corresponds to the cost value at the threshold of the system. In an ideal case, both the *best cost* and *system cost* will be the same for all distances. The genuine acceptance rate and imposter rejection rate presented are the values at *system cost*. In later sections, the same convention is followed for reporting the performance of the speaker verification system using other features. Note that the performance is shown as a function of the distance only for analysis but no information of the distance is used during the verification phase.

When the same threshold is set for all the distances, the genuine acceptance

Table 3.1: Performance of speaker verification system using MFCC + CMS as features.

Channel	Genuine acceptance(%)	Imposter rejection(%)	Best cost	System cost
close	91.7	99.3	0.0137	0.0137
2 ft	43.2	99.7	0.0420	0.0588
4 ft	22.6	99.8	0.0575	0.0784
6 ft	20.5	99.8	0.0659	0.0811

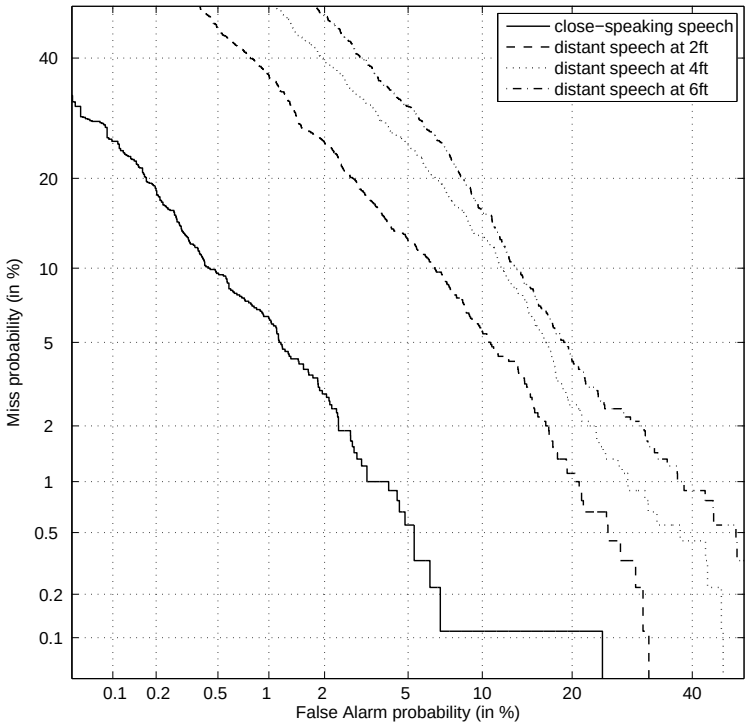

Figure 3.7: DET curve showing the system performance using MFCC + CMS as feature.

rate decreases from 91.7% for close-speaking speech to 20.5% for speech collected at 6 ft. Since same threshold is used for all the distances with a high priority for false acceptance, the speaker verification system progressively gets to the case where it rejects all the inputs. So, not much of a difference can be observed in imposter rejection rate for different distances. For an ideal system, both the genuine acceptance and imposter rejection rate should be 100%.

The performance of the speaker verification system using MFCC for speech signals collected at different distances is also shown in the form of a detection error trade-off (DET) curves [36] in Fig. 3.7. The point where the line $y = x$ meets the curve gives us the equal error rate (EER) of the performance of the system for that

distance. The EER represents the performance of the system with equal weights given to false rejection and false acceptance rates. The lesser the EER, the better is the performance of the system. Notice that the EER increases for distant speech signals showing a decrease in the performance.

3.3 Conclusions

In this chapter, the components of the baseline speaker verification system were explained. A robust method for begin-end detection was also proposed for distant speech. The DTW algorithm for comparison of patterns of varying lengths was explained. For score normalization, we have used additional information from the optimal warping path of DTW algorithm to extract and represent speaker-specific information from the features. It was seen that the performance of the system decreases significantly as the distance increases, giving further evidence of variation of acoustic features. This leads us to the conclusion that robust acoustic features, which suffer lesser variation with distance are required to develop robust speaker verification system for distant speech.

Chapter 4

Short segment features derived from high SNR regions of speech

In Chapter 2 we have seen that the standard spectral features used in the feature extraction process are modified due to reverberation and noise. In Chapter 3, it was observed that this variation led to a decrease in the performance of a speaker verification system even with improvements in begin-end detection and score normalization. Human beings on the other hand can recognize speakers even at a distance because of the binaural nature of hearing. Though compensation of acoustic features have been used to overcome the effect of distance, the improvements have been minimal. Moreover, such methods are specific for a given ambience. Hence, there is need for new features for speaker verification which are robust to distance. This chapter deals with the extraction and processing of a newly proposed feature based on processing short-segments of speech signal.

The chapter is organized as follows. In Section 4.1, the significance of short-segments is explained. In Section 4.2, we describe the extraction of a short-segment feature. In Section 4.3, an analysis of the proposed feature is made and the results of speaker verification system using the proposed feature are shown.

4.1 Significance of short-segments

The speech production mechanism for producing voiced speech involves a sequence of impulse-like excitations being modulated by a time-varying vocal tract system. A short segment of speech signal is a block of speech signal within a pitch cycle. The blocks in the neighborhood of the impulse-like excitations have a higher SNR because of the impulse-like nature, when compared to other regions. It is also likely that human beings extract information from high SNR regions in speech, which helps them perceive speech even when spoken at a distance. Thus, features derived from short segments of speech signal in the regions of high SNR may be robust in case of distant speech.

The effect of reverberation on the regions of speech signal near the impulse like excitations is lesser when compared with other regions because of the higher SNR of these regions. It is the regions with low SNR that are more affected by reverberation. The standard methods use blocks of 10-30 ms for extraction of features, which typically consist of 2-3 pitch cycles of voiced speech. By considering a frame of 2-3 pitch cycles in case of distant speech, the percentage of low SNR regions is more. So, by considering a bigger block in case of distant speech, the contribution of the degraded speech samples of the low SNR regions in the computation of acoustic features might be more. So, it is better to use short segments for the degraded speech.

The significance of short segments over blocks of 2-3 pitch cycles can be shown through an experiment which calculates normalized energies of the speech signal for different window lengths both for a close-speaking speech as well as distant speech. The close-speaking speech signal and the distant speech signal are time-aligned, by compensating for the delay between the two signals.

Fig. 4.1(a) shows a segment of close-speaking speech signal followed by the energies calculated using different window lengths in Fig. 4.1(b),(c) and (d). The window lengths used are 4 ms, 10 ms and 30 ms respectively. It can be seen that the energies in the vicinity of the impulses are higher in the case of the short segments

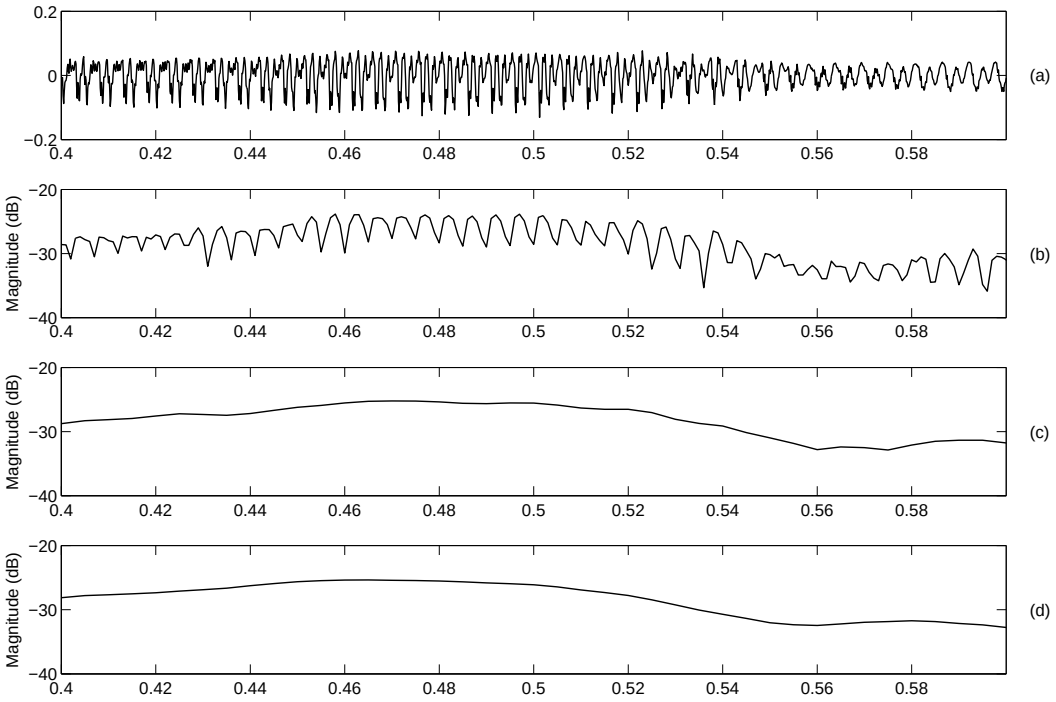

Figure 4.1: Energy of a segment of close-speaking speech signal with different frame sizes. (a) Frame size of 4 ms, (b) frame size of 10 ms and (c) frame size of 30 ms.

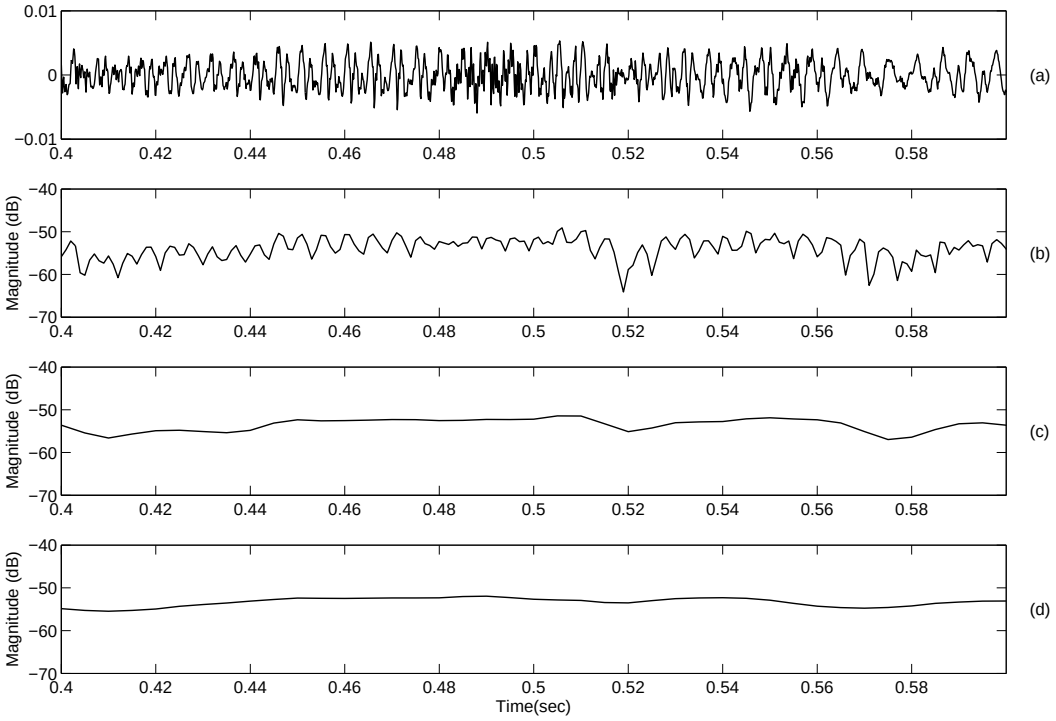

Figure 4.2: Energy of a segment of speech signal collected at 6 ft with different frame sizes. (a) Frame size of 4 ms, (b) frame size of 10 ms and (c) frame size of 30 ms.

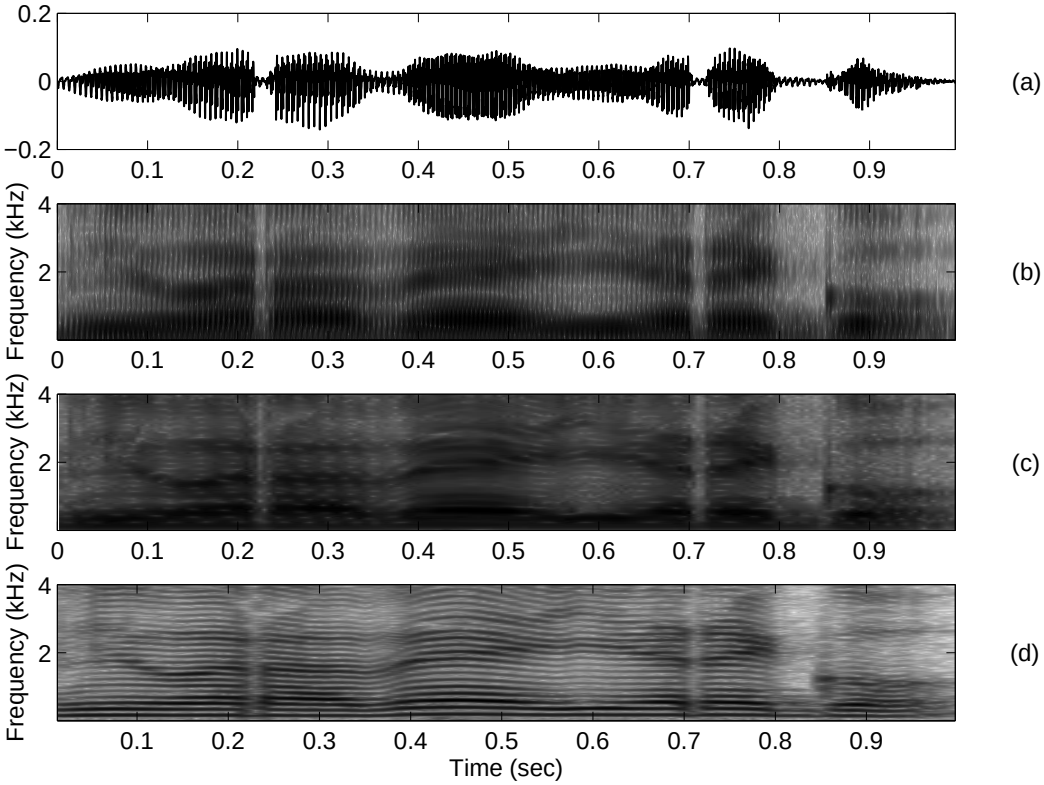

Figure 4.3: Spectrogram of a close-speaking speech signal computed using different frame sizes. (a) Close-speaking speech signal, (b) spectrogram with frame size of 4 ms, (c) spectrogram with frame size of 10 ms and (d) spectrogram with frame size of 30 ms.

(i.e., with window length of 4 ms) when compared to the other window lengths. Similar trend is observed in Fig. 4.2 which shows the energies for different windows, computed from a speech signal collected at 6 ft. Though the overall energy degrades from close to distant speech, the regions around impulses suffer from lesser degradation when compared to other regions. These short segments are likely to preserve acoustic features even in distant speech.

Another observation can be made from the spectrogram of speech signals. Fig. 4.3(a) shows a close-speaking speech signal, along with a spectrograms computed using frame sizes of 4 ms, 10 ms and 30 ms in Fig. 4.3(b),(c) and (d) respectively. Similarly Fig. 4.4 shows the corresponding spectrograms of the same speech collected at 6 ft. The formant contours can be clearly seen in Fig. 4.3(d) which is a spectrogram computed using a window length of 30 ms for the close-speaking speech signal. However, due to the effect of noise and reverberation on the signal in the case of

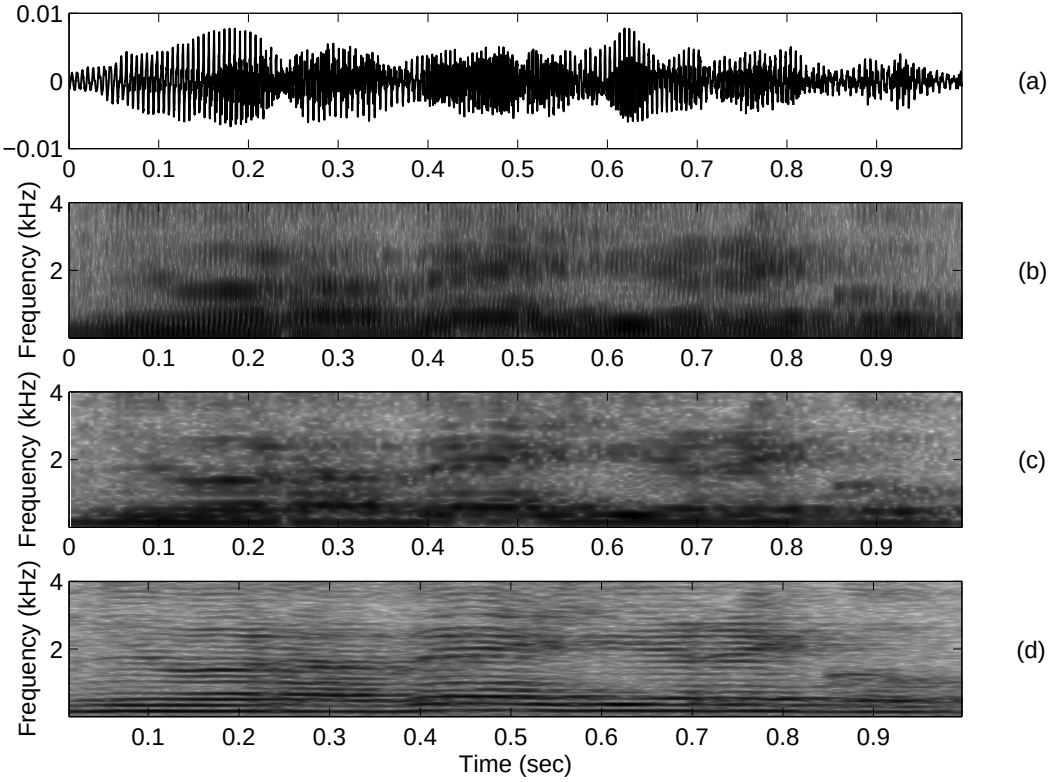

Figure 4.4: Spectrogram of a speech signal collected at 6 ft, computed using different frame sizes. (a) Distant speech signal, (b) spectrogram with frame size of 4 ms, (c) spectrogram with frame size of 10 ms and (d) spectrogram with frame size of 30 ms.

distant speech, no such clear contours can be seen in Fig. 4.4(d). But Fig. 4.4(b) which shows a spectrogram of speech signal computed using a window length of 4 ms, shows a better formant contour than the spectrogram computed using a window length of 30 ms shown in Fig. 4.4(d). This shows the robustness of using short segments in the computation of spectral features. From this illustration, we can observe that in case of distant speech, formant contours can be tracked better when a short segment is considered for analysis.

Formant contours are known for containing both speech and speaker specific properties and have been used in many speaker recognition applications [37, 38, 39]. A speaker-specific feature is extracted using the key idea of short segment processing, which is explained in the next section.

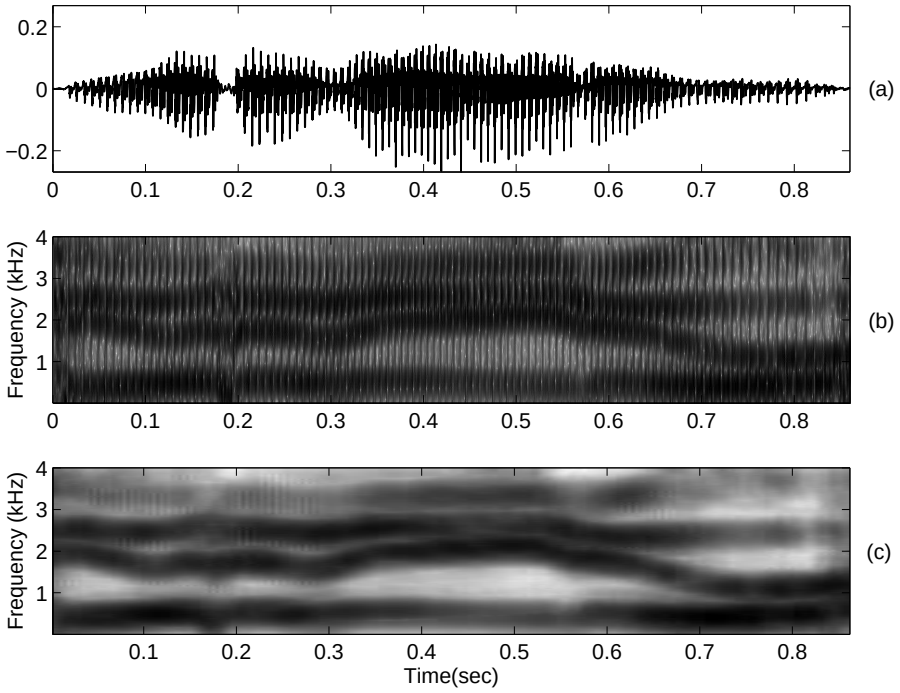

Figure 4.5: Illustrating the significance of short-segment analysis: (a) Close-speaking speech signal, (b) wideband spectrogram of the speech signal and (c) time-averaged wideband spectrogram.

4.2 Extraction of short segment cepstral coefficients

The extraction of the short segment cepstral coefficients involves computation of a wideband spectrogram. A spectrogram is a visual representation of the spectral density of a signal which varies with time. A wideband spectrogram is a representation which uses short segments for analysis typically less than a pitch cycle with a small shift (such as 1 ms). We have used segments of 3 ms for computing the short-time spectrum, and a shift of 1 ms. Each frame of the speech signal is normalized so that the effect of noise on the range of spectrum is minimized. Fig. 4.5(b) shows the wideband spectrogram of the speech signal captured from a close-speaking microphone. The sentence uttered is "We were away a year ago". Notice that the formant contours are clearly visible in the spectrogram.

To elevate the peaks of the formant contours, the wideband spectrogram is averaged over time, by using 10 successive frames. This results in a time-averaged

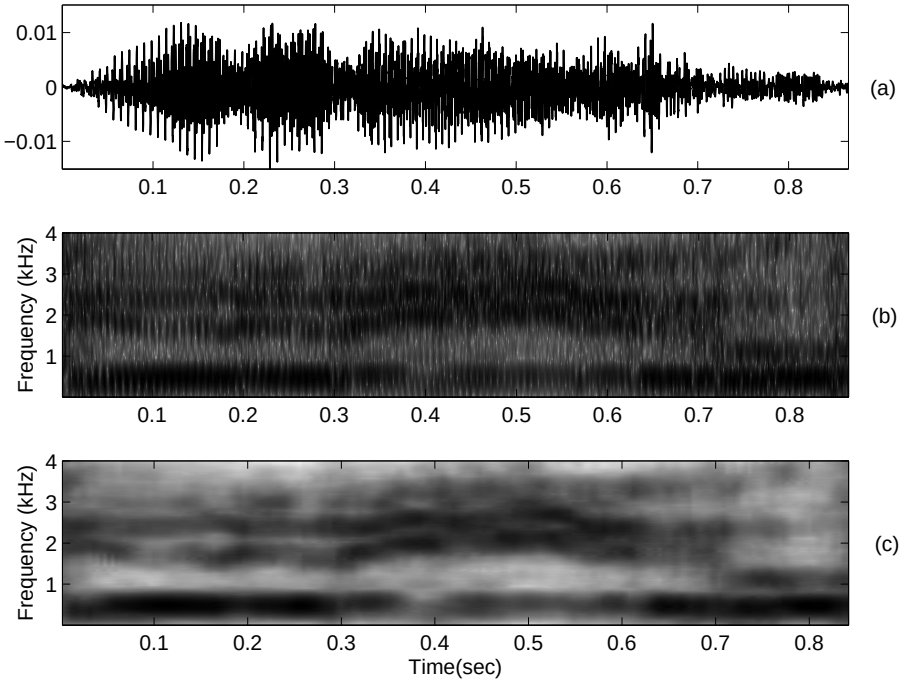

Figure 4.6: Robustness of short-segment analysis in distant speech. (a) Distant speech signal collected at 6ft, (b) wideband spectrogram of the speech signal and (c) time-averaged wideband spectrogram.

wideband spectrogram, which is shown in Fig. 4.5(c).

Fig. 4.6 shows the extraction of short-segment feature for the same signal collected at a distance of 6 ft. Notice the similarities between the time-averaged wideband spectrograms of close-speaking and distant speech signals. The formant contours are clearly visible for close speech whereas they slightly deteriorate in the case of distant speech. In fact, this variation is also speaker specific i.e., the similarities between the time-averaged wideband spectrograms varies across speakers.

The cepstral coefficients are calculated for each frame to reduce the dimensionality of the feature [8]. This is done by calculating the inverse discrete Fourier transform (IDFT) of the time-averaged wideband spectrogram for each frame. 20 cepstral coefficients are derived to represent the spectral information in each frame. So, the extracted features are termed as short segment cepstral coefficients (SSCC).

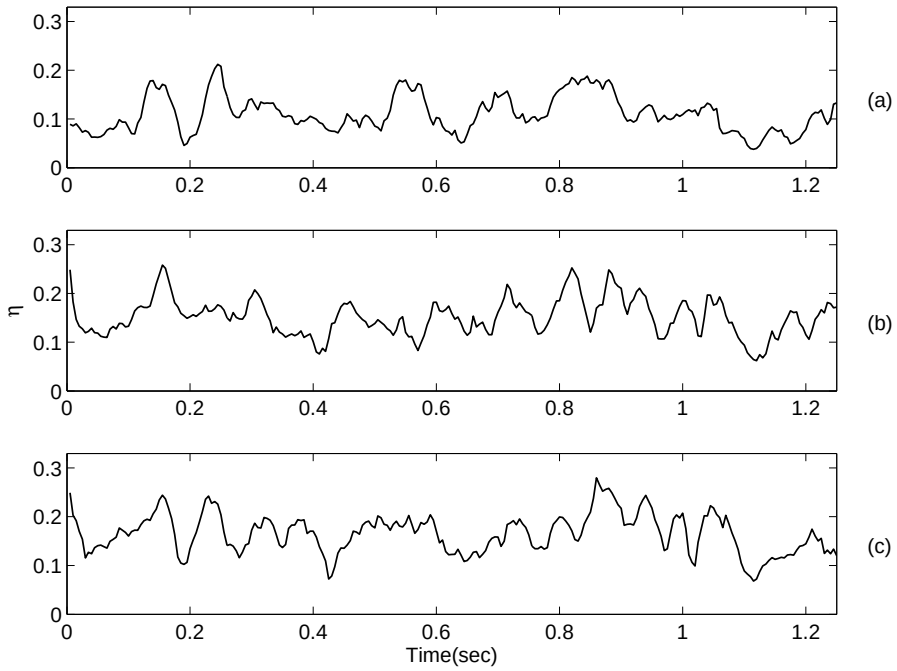

Figure 4.7: Variation of SSCC feature with distance. Variation between SSCCs extracted from close-speaking speech and distant speech for speech signals collected at (a) 2 ft, (b) 4 ft and (c) 6 ft.

4.3 Analysis of the variation of short-segment features

In this section, we observe the variation of short-segment cepstral coefficients between close-speaking speech and distant speech. We follow a procedure similar to the one described in Section 2.3. That is, Euclidean distance is computed between the SSCCs derived from the corresponding frames of close-speaking and distant speech signals. The Euclidean distance is given by

$$\chi(i) = ||\mathbf{x_{c,i}} - \mathbf{x_{d,i}}|| \tag{4.1}$$

where $\mathbf{x_{c,i}}$ and $\mathbf{x_{d,i}}$ represent the unit vectors of the SSCCs derived from the close-speaking speech and distant speech, respectively, for i^{th} frame of the signals.

The effect of distance on the SSCC feature is shown in Fig. 4.7 where the variation of SSCC between a close-speaking speech signal and its corresponding time aligned

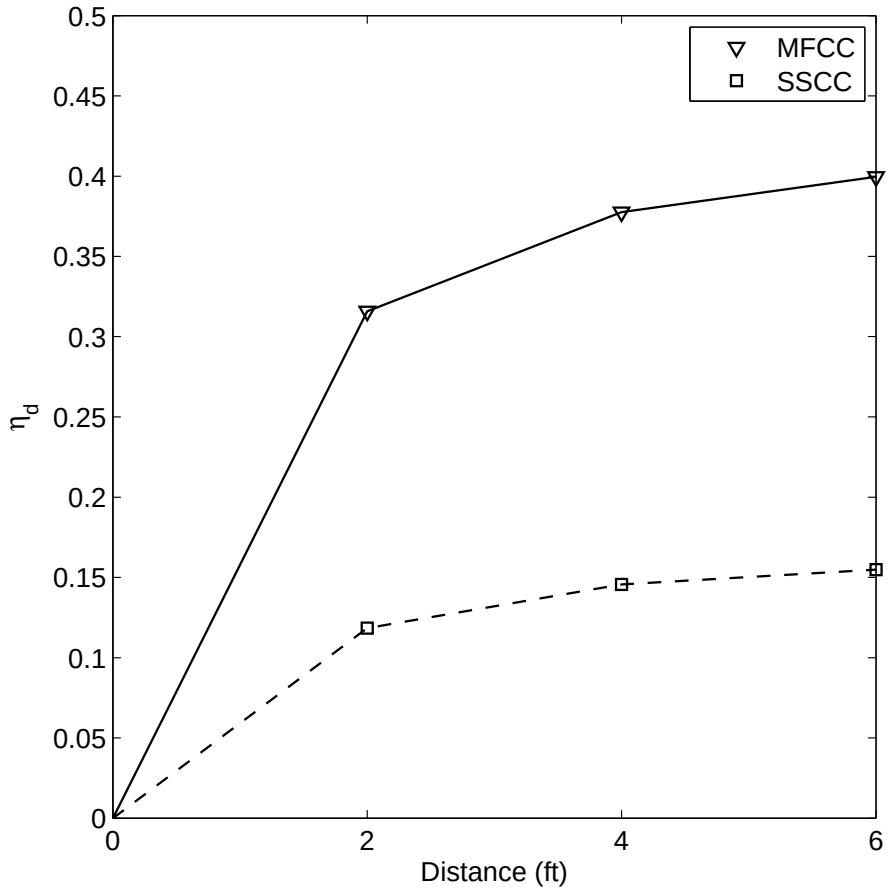

Figure 4.8: Variation of acoustic features with distance.

distant speech signal is shown. For each distance d, mean χ_d is calculated as

$$\chi_d = \frac{1}{N} \sum_{i=1}^{N} \chi(i) \tag{4.2}$$

which gives the deviation of the features as a function of distance.

The similarity between the curves in Fig. 4.7 indicates that the SSCC feature does not degrade appreciably with distance. We have also observed that the variation of SSCC feature with distance is speaker-dependent.

Fig. 4.8 shows the average variation of the features across speakers as a function of distance. The number of speakers used is 10 and the number of utterances used per speaker is 5. Only voiced frames are considered in the calculation of average χ_d.

We observe from Fig. 4.8 that the average variation of SSCC across distance

Table 4.1: Performance of speaker verification system using SSCC feature.

Channel	Genuine acceptance(%)	Imposter rejection(%)	Best cost	System cost
close	93.4	99.3	0.0121	0.0121
2 ft	69.8	99.5	0.0280	0.0344
4 ft	43.0	99.7	0.0486	0.0596
6 ft	38.3	99.6	0.0553	0.0648

is significantly less when compared to MFCC features. This is attributed to the fact that SSCCs primarily represent the information specific to formants (spectral peaks), while MFCCs represent information specific to gross spectral envelope.

The time-averaged wideband spectrogram represents the information of formant contours, so the extracted SSCC feature captures both speech and speaker specific properties. The speaker-specific property of the feature is reflected in the performance of the modified baseline system where the MFCC feature is replaced with SSCC feature. The database has 900 genuine tests and 39600 imposter tests for each channel. The α and β explained in section 3.1.3 are extracted from the SSCC feature and represent the measures for discriminating between speakers. The scores are then normalized in the range of 0 to 1 and are combined as explained in Section 3.2. The weights of 0.1 and 0.9 for α and β are empirically found. The best performance threshold obtained for close-speaking speech is set as system threshold of the system and is used for 2 ft, 4 ft and 6ft cases. The cost function used is given by

$$\eta_C = 0.1 \times \eta_R + 0.9 \times \eta_A, \tag{4.3}$$

where η_R and η_A denote the false rejection and the false acceptance rates respectively. Table. 4.1 shows the performance of the system using SSCC as feature.

With the same threshold for all cases, the genuine acceptance rate is 93.4% for close-speaking speech, 69.8% for speech collected at 2 ft, 43% for speech collected at 4 ft and 38.3% for speech signal collected at 6 ft. The performance of the system shows an improvement of 1.7%, 26.6%, 20.4% and 17.8% in the genuine acceptance rate for speech signal collected at 2 ft, 4 ft and 6 ft, respectively, when compared with the baseline system. The imposter rejection rate is maintained around 99%

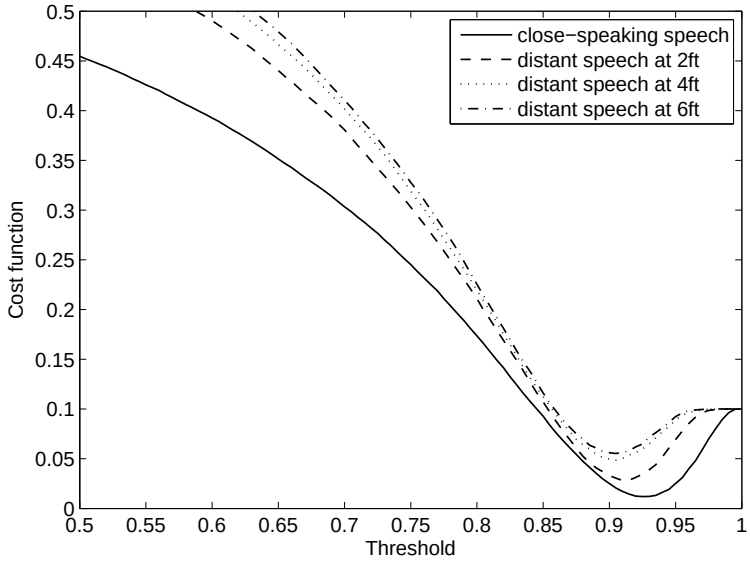

Figure 4.9: Cost function of the system using SSCC as feature, for speech collected at different distances.

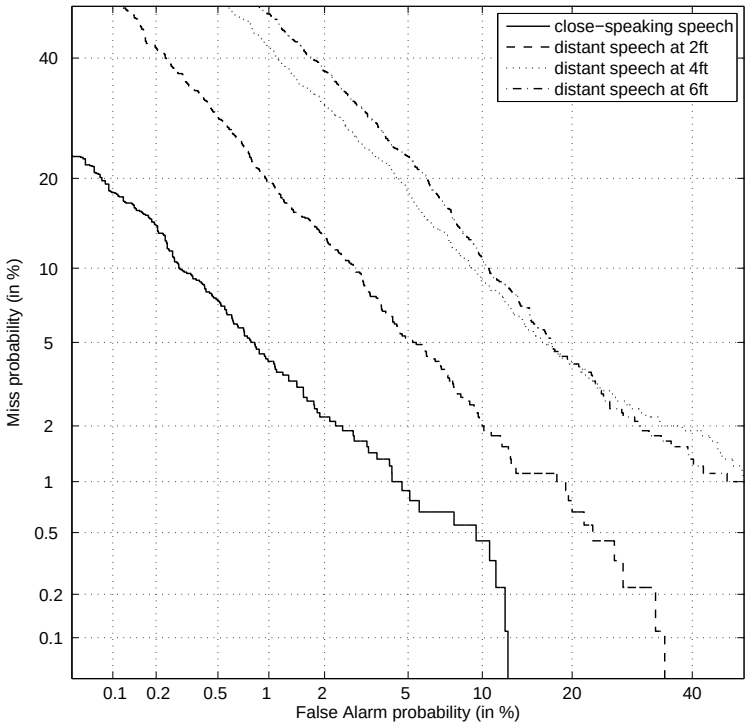

Figure 4.10: DET curve showing the system performance using SSCC as feature.

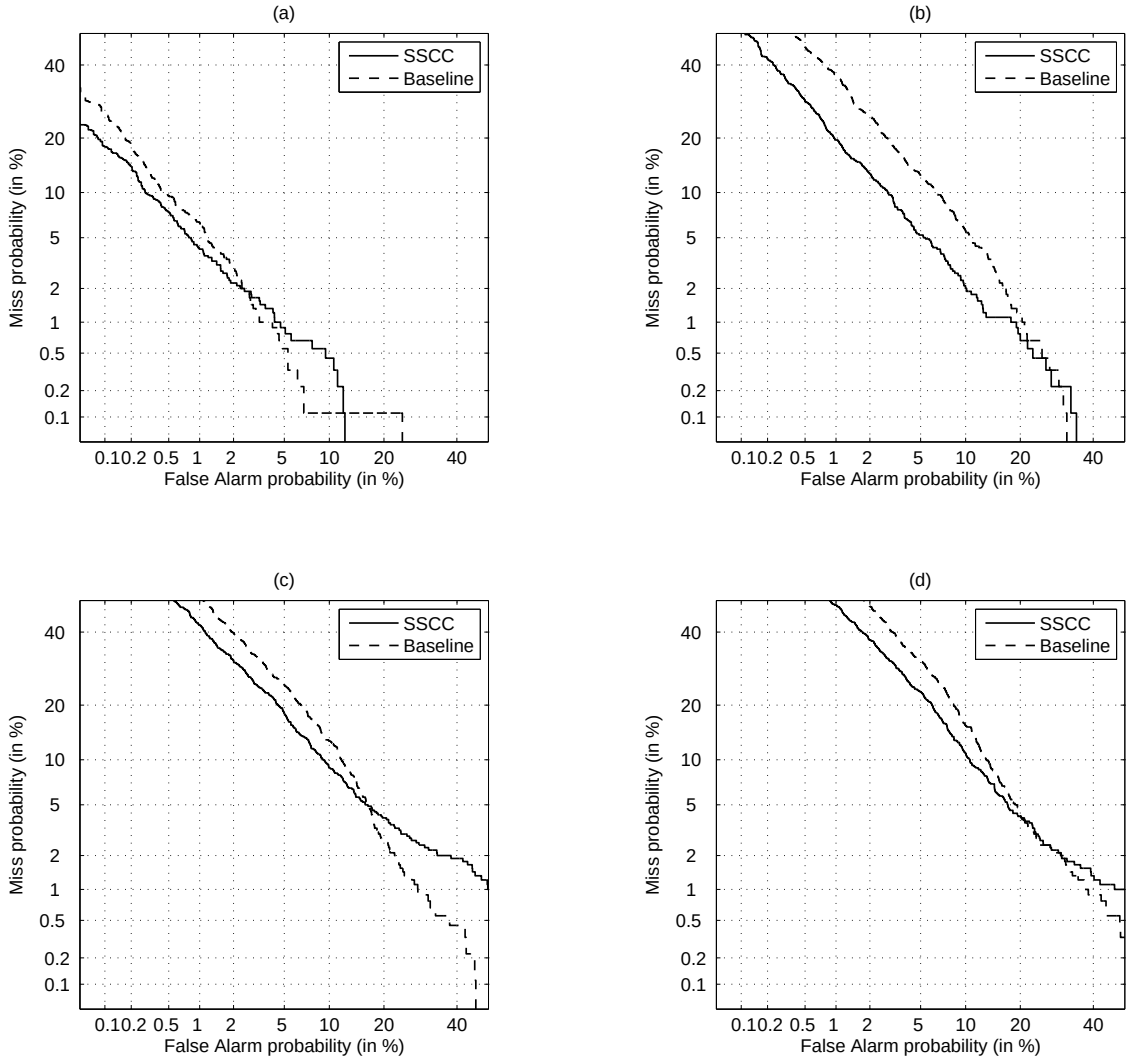

Figure 4.11: Channelwise comparison of the system performance using (MFCC + CMS) and SSCC as features for (a) close-speaking speech, (b) distant speech collected at 2 ft, (c) distant speech collected a 4 ft and (d) distant speech collected at 6 ft.

for all distances because the high weightage given to the false acceptance rate. Also notice that the best cost at each distance is close to the system cost showing the robustness of the feature. The cost function for the system is shown in Fig. 4.9. It is observed that the cost functions for distant speech case are converging to the close-speaking case indicating an advantage of the SSCC feature compared to MFCCs. The performance of the speaker verification system using SSCCs is also shown in the form of DET curves in Fig. 4.10, for speech signals collected at different distances. A similar trend in the decrease of EER is observed for each distance showing an improvement in the performance.

A channel wise comparison between the system performances using MFCC and SSCC as features is shown in Fig. 4.11. It is observed that in all the cases (Figs. 4.11(a), (b), (c) and (d)), the DET curves due to SSCC as feature are closer to the origin, relative to the DET curves due to MFCCs. Notice that the DET curves are similar for close-speaking speech but the separation can be observed between the curves leading to a difference in the EER values. This indicates the difference in the performance of the system for the two features.

4.4 Conclusions

In this chapter we showed the significance of extracting features from short-segments, and its advantages in case of distant speech. A feature based on short segments is extracted and applied on the baseline system, which improved the performance of the system slightly. Though the improvement is small, there is scope of improvement by using additional robust features like pitch and duration. The next chapter deals with those features, and the combination of the features to improve the performance of the system.

Chapter 5

Significance of suprasegmental features for speaker verification

In chapter 4, we examined the importance of a short-segment feature which was able to improve the performance of the speaker verification system for distant speech. The performance can be further improved by using additional robust features. Human beings use several features like pitch, duration, speaking rate and speaking style for recognizing speakers. These robust features have complementary information, but have not been used extensively because of the difficulty in extraction and usage of those features. This chapter discusses how some of those features can be extracted to improve the performance of the speaker verification system in distant speech.

The chapter is organized as follows. In Section 5.1, we use the significance of high signal-to-reverberation component ratio (SRR) for improving the performance of the speaker verification system. The performance of this system is further improved by using duration and pitch information which are explained in Section 5.2 and 5.3 respectively. Combination of all the features is explained in Section 5.4.

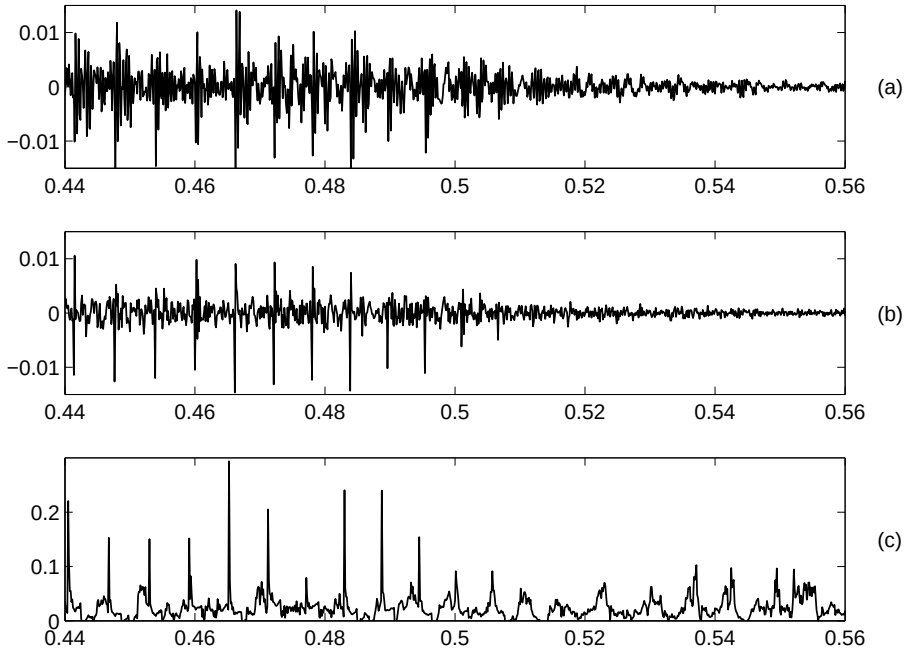

Figure 5.1: Effect of reverberation on distant speech. (a) Distant speech signal collected at 2 ft of a voiced segment, (b) LP residual and (c) smoothed normalized error.

5.1 Significance of signal-to-reverberation component ratio in the selection of speech frames

We have seen earlier that the speech signal collected at a distance from the speaker (distant speech) is affected by noise and reverberation. Reverberation is a result of the addition of the original signal and the reflected signal. The effect of the reflected component will be different in different segments because of its dependence on the energy of the segment and hence all the segments may not have the same degradation. By exploiting the segments with lesser degradation, the system performance can be improved. Signal-to-reverberation component ratio (SRR) was previously proposed to identify such regions [40]. It was shown that a region with high SRR can be used in enhancement of a reverberated speech. These regions are presently used during score normalization where frame weights are assigned. Note that no enhancement of distant speech is performed.

5.1.1 Extraction of frame weights

It has been shown that normalized error (η) of linear prediction analysis can be used in identifying high SRR regions [40]. Normalized error is a ratio of the energy of the LP residual and the speech signal, computed over a frame of 2-5 ms. This computation is performed for overlapping frames, with a frame shift of 1 sample. The extraction of high SRR regions involves computation of a normalized error (η) for every sample of the differenced speech signal for each frame of 2 ms duration. A 5^{th} order LP analysis using the autocorrelation method is used for the calculation of LP residual [41]. Time-averaged normalized error is then calculated by removing the trend in normalized error by averaging it with a 10 ms Hamming window and subtracting the smoothed function from normalized error.

Fig. 5.1(a), shows a voiced segment of a speech signal collected at 2 ft and the corresponding LP residual is shown in Fig. 5.1(b). The time-averaged normalized error is shown in Fig. 5.1(c). Notice that the distinction of the glottal cycles can be made in time region of 0.45 - 0.5 sec in Fig. 5.1(c) which corresponds to the high SRR region. In the region 0.5 - 0.56 sec, no such distinction can be seen due to low SRR or purely reverberant nature of the signal.

For the regions where the peaks are clearly visible in the time-averaged normalized error, the lag of autocorrelation is consistent. The frame weight information is calculated by using a threshold on the standard deviation of the lags extracted from autocorrelation of the time-averaged normalized error. Frame weight of 1 is given for high SRR regions and 0 is given to other regions. Fig. 5.2(a) shows a time-averaged normalized error along with the LP residual and speech signal collected at 2 ft, in Fig. 5.2(b). The weightage given to the frames is indicated in Fig. 5.2(c).

These frame weights are used in the score normalization step of calculating α and β by only considering the frames with high SRR.

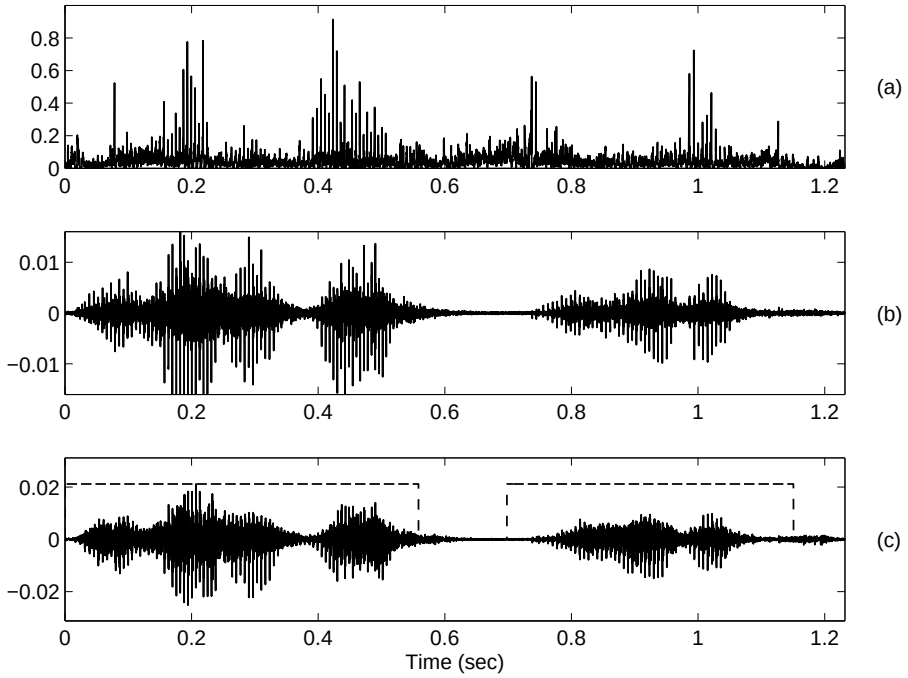

Figure 5.2: Derivation of frame weights. (a) Time-averaged normalized error, (b) LP residual and (c) distant speech signal collected at 2 ft along with frame weights.

Table 5.1: Performance of speaker verification system using SSCC features extracted from high SRR regions.

Channel	Genuine acceptance(%)	Imposter rejection(%)	Best cost	System cost
close	93.1	99.4	0.0114	0.0114
2 ft	71.5	99.5	0.0256	0.0321
4 ft	46.2	99.5	0.0512	0.0583
6 ft	41.3	99.4	0.0579	0.0640

5.1.2 Performance of the system using selected regions of speech

The procedure mentioned in Section 3.2 is applied on the scores obtained by using frame weights on SSCC features. The performance of the system with SSCC feature along with frame weights are shown in Table. 5.1. The main advantage of using frame weights can be seen in the case of distant speech, where genuine acceptance rate has increased by 1.7 % for 2 ft, 3.2% for 4 ft and 3% for 6 ft when compared with the performance of SSCC feature. For genuine speakers, though the framewise Euclidean scores are lesser than the background speakers, effects of distance is reflected in an increment in the framewise Euclidean scores for a few frames as shown in Fig. 5.3(a).

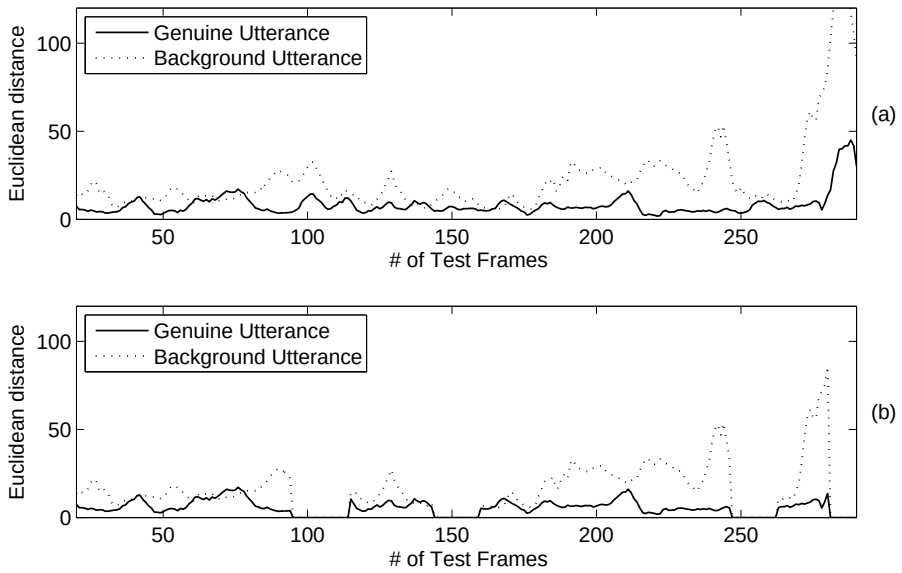

Figure 5.3: Illustration of the use of high SRR regions during score normalization. (a) Without frame weights and (b) with frame weights.

By only considering the regions with high SRR, the effect of reverberation is reduced in the sense that frames which contribute to acoustic mismatch are eliminated and is shown in Fig. 5.3(b). This in turn reduces the mismatch for genuine speakers. For imposter speakers, the probability that his scores are lesser than background speakers is less and so it will not affect the system performance. It is observed that percentage of frames recognized as high SRR reduces to approximately 50% of the total number of frames for 6 ft distant speech signals. For general purposes, if most of the speech signal is affected by noise and reverberation, then an option of "No information available" can be used instead of verifying a speaker with the corrupt speech. Alternatively, the speaker may be prompted to utter the text again with greater clarity and loudness. It is also observed that this percentage is speaker-dependent.

Due to the selection of high SRR regions, the overall system cost for each distance reduces as shown in Fig. 5.4. Also, the convergence of the cost functions for distant speech is better when compared with previous cases. The performance of the speaker verification system using SSCCs and the frame weights is shown in the form of DET curves in Fig. 5.5, for speech signals collected at different distances. Notice that

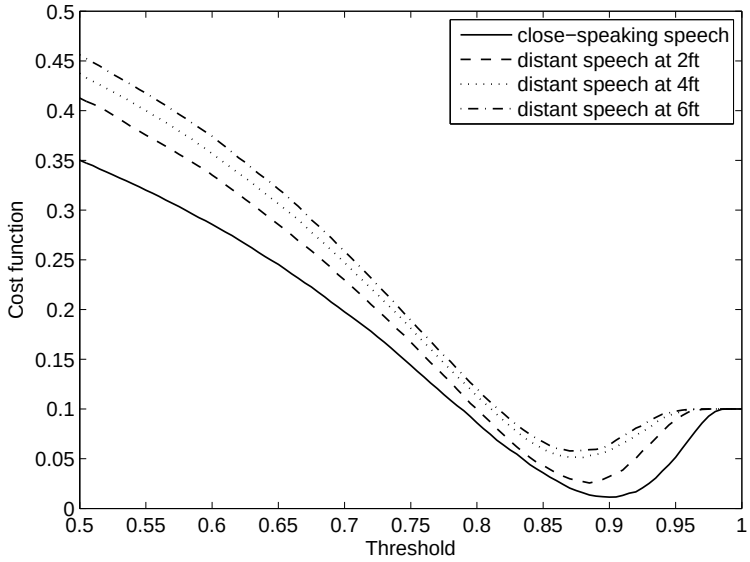

Figure 5.4: Cost function of the system using SSCC features extracted from high SRR regions.

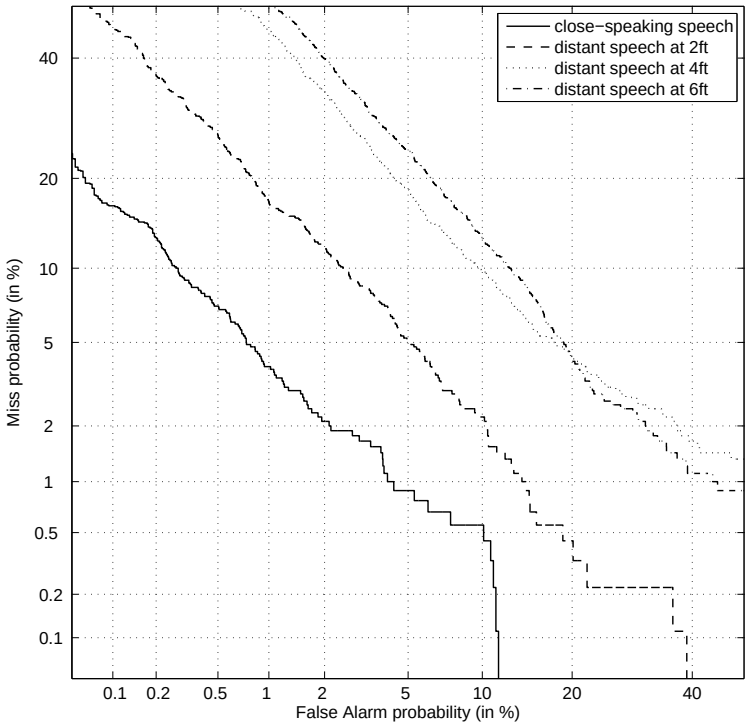

Figure 5.5: DET curves showing the system performance using SSCC along features extracted from high SRR regions.

the EER values are lesser in all the cases when compared to the speaker verification system using only SSCC features, indicating an improvement in the performance of this system.

5.2 Exploiting duration information for speaker verification

Duration information can be categorized as duration of sentences, words and syllables. Duration as a feature has be used previously in many speaker recognition systems [42, 43, 44]. But because of the difficulty in measuring the information, it has not been used much in text-dependent speaker verification applications. Duration information without explicitly locating boundary of the units has been proposed in [45]. This is achieved using the nature of the optimal warping path obtained in the DTW algorithm. Studies show that the optimal warping path closely follows the diagonal line ($y = x$) in the $x - y$ plane for genuine speakers whereas it deviates significantly from the diagonal line for imposter speakers. Fig. 5.6 shows an illustration of this pattern between a genuine and imposter case. Note that duration itself is not the best speaker-specific feature but when used along with other features, it provides additional evidence in the favor of the genuine speaker, and thus helps in rejecting the imposters. This contributes to an increase in the performance of the system. Though the total duration of a given text may vary between two utterances of the same speaker, the relative durations or the percentage durations of the units are usually consistent. An imposter might mimic the overall duration but not the relative durations in a sentence [45].

The duration information is extracted by determining a line (regression line) which is the best line fit for the optimal warping path curve, and then measuring the deviation of the warping path from that line. Let (x_i, y_i) be the points on the warping path, with $i = 1, 2, ...K$. The deviation of y_i from its regression line is an indication of duration mismatch. The regression line $\hat{y}_i = mx_i + c$ can be found

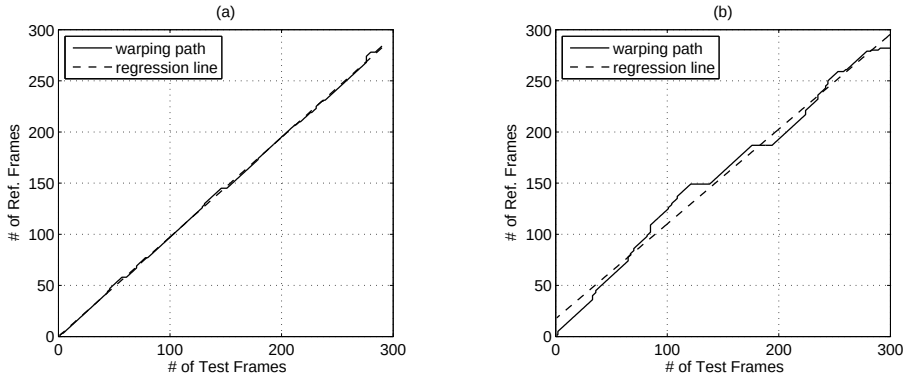

Figure 5.6: Duration information represented by the deviation between warping path and regression line. (a) Genuine case and (b) imposter case.

using the least squares method. Then, average sum of squared error (ξ) represents the duration mismatch between the training and the testing features. This is a dissimilarity measure and is calculated as

$$\xi = \frac{\sum_{i=1}^{K}(\hat{y}_i - y_i)^2}{K} \tag{5.1}$$

The regression line is determined using least squares method which is explained below.

5.2.1 Least squares method

Given the warping path (x_i, y_i), $i = 1, 2..K$ the objective is to find $\hat{y}_i = mx_i + c$ such that error $E = \sum_{i=1}^{K}(y_i - \hat{y}_i)^2$ is minimum.

So

$$E = \sum_{i=1}^{K}(y_i - (mx_i + c))^2 \tag{5.2}$$

Table 5.2: Performance of speaker verification system using SSCC features along with duration.

Channel	Genuine acceptance(%)	Imposter rejection(%)	Best cost	System cost
close	94.1	99.5	0.0103	0.0103
2 ft	73.4	99.7	0.0222	0.0284
4 ft	42.2	99.8	0.0409	0.0589
6 ft	38.3	99.8	0.0455	0.0629

To minimize error, $\frac{\partial E}{\partial m} = 0$ and $\frac{\partial E}{\partial c} = 0$

$$\frac{\partial E}{\partial m} = 2\sum_{i=1}^{K}(y_i - (mx_i + c))(-x_i) = 0 \tag{5.3}$$

$$\frac{\partial E}{\partial c} = 2\sum_{i=1}^{K}(y_i - (mx_i + c)) = 0 \tag{5.4}$$

Solving the equations, we get

$$m = \frac{K\sum x_i y_i - \sum x_i \sum y_i}{K\sum x_i^2 - (\sum x_i)^2} \tag{5.5}$$

$$c = \frac{1}{K}(\sum y_i - m\sum x_i) \tag{5.6}$$

Hence ξ can be calculated using a set of training and testing features, which represents a measure of duration.

5.2.2 Performance of the system using duration information

The measures α and β are calculated from the SSCC features as explained in section 3.1.3. All the scores (α, β and ξ) are normalized in the range of 0 to 1 as explained in section 3.2 and then a weighted sum of the these measures is used to verify a speaker. The weighted sum is given by

$$S = 0.09 \times \alpha + 0.81 \times \beta + 0.1 \times \xi \tag{5.7}$$

The weights for the measures are empirically found.

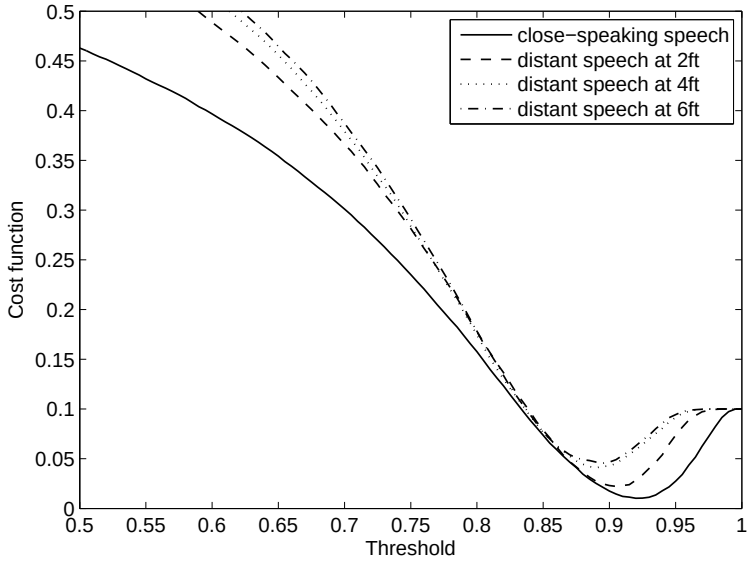

Figure 5.7: Cost function of the system using SSCC feature along with duration information.

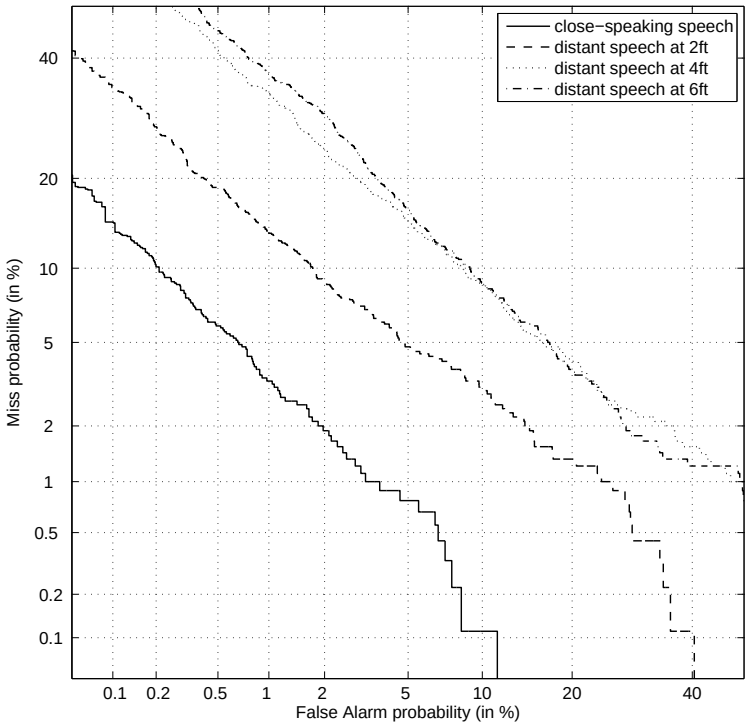

Figure 5.8: DET curve showing the system performance using SSCC feature along with duration information.

The best performance threshold obtained for close-speaking speech is set as threshold of the system and is used for 2 ft, 4 ft and 6ft cases. The results of the speaker verification system using the SSCC feature and duration information are shown in Table. 5.2. The results show an improved performance for the combined case showing the complementary information of duration. Duration helps in increasing the genuine confidence scores and thus decreasing the imposter acceptance. The performance of the speaker verification system has been improved considerably for speech collected at 2 ft. Though duration is a robust feature, the improvement is seen only because it is extracted from SSCC feature which is able to preserve speech information even for speech signals collected at a distance. In contrast, it is observed that MFCC features do not bring out the duration information significantly for distant speech.

The improvement in the system performance is reflected in the cost functions shown in Fig. 5.7 where the convergence is better in the combined case. The performance of the system is also shown in the form of DET curves in Fig. 5.8 where a similar trend in the reduction of EER is observed.

5.3 Pitch information

It was shown that pitch information has important advantages over spectral information for speaker recognition [46, 47, 48]. It was shown earlier that the standard spectral features are affected in case of distant speech. But pitch is unaffected by distance and channel variations [49]. Though pitch is a robust feature in distant speech, difficulty exists in the extraction process of the F_0 for distant speech. Studies on extraction methods of F_0 in adverse conditions was performed in [50][51]. A recent study on estimation of pitch period from distant speech was done in [52], which will be used for F_0 extraction.

5.3.1 Extraction of F_0 from distant speech

The estimation of pitch period from distant speech signals is a two step procedure. In the first step, a zero-frequency filtered signal is derived, which has a property that the interval between successive positive zero crossings gives an estimate of the pitch period. In the second step, the number of spurious zero crossings in the zero-frequency filtered signal is reduced by emphasizing the regularity in successive pitch periods.

The extraction of the zero-frequency filtered signal was proposed in [53], where it was shown that the information about the time instants of occurrence of the excitation impulses is reflected as discontinuities in the time domain. The effect of the discontinuities was highlighted by filtering the speech signal with a resonator at zero frequency which predominantly contains frequency component around zero. A zero-frequency resonator is a second-order infinite impulse response (IIR) filter with poles on the unit circle. The main advantage of using a zero-frequency resonator is that it removes the resonances of the vocal tract which are located at much higher frequencies than the zero frequency.

The following steps are involved to derive the zero frequency filtered signal [53].

1. The speech signal $s[n]$ is differenced to remove any time-varying low frequency bias introduced by recording devices.

$$x[n] = s[n] - s[n-1] \tag{5.8}$$

2. The signal $x[n]$ is passed through a cascade of two ideal resonators at zero frequency to result in $y_2[n]$ which grows/decays in polynomial time. The cascading is equivalent to successive integration four times.

$$y_1[n] = x[n] - a_1 y_1[n-1] - a_2 y_1[n-2] \tag{5.9}$$

$$y_2[n] = y_1[n] - a_1 y_2[n-1] - a_2 y_2[n-2] \tag{5.10}$$

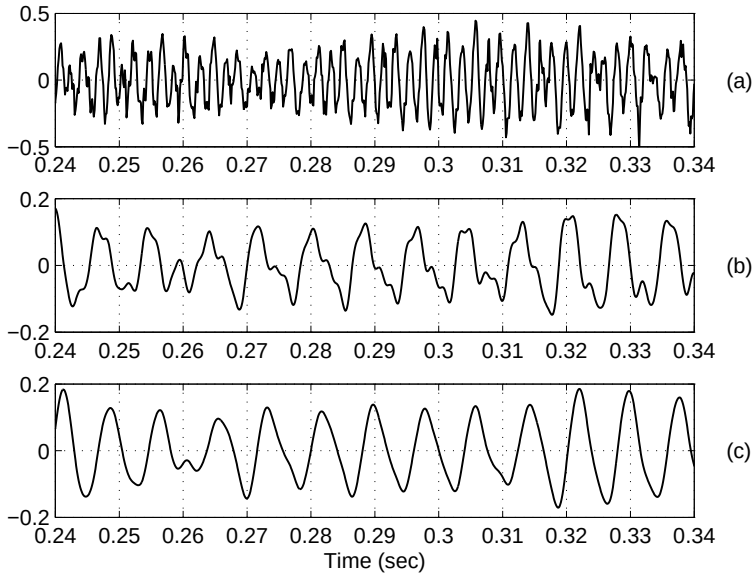

Figure 5.9: Step 1 of pitch extraction process. (a) Speech signal collected at 6 ft, (b) filtered signal $y[n]$ and (c) refined filtered signal $\tilde{y}[n]$.

where $a_1 = -2$ and $a_2 = 1$.

3. The trend in $y_2[n]$ is removed by subtracting the average over 10 ms at each sample. The resulting signal

$$y[n] = y_2[n] - \frac{1}{2N+1} \sum_{m=-N}^{N} y_2[n+m] \tag{5.11}$$

is called the zero-frequency filtered signal. Here $2N+1$ corresponds to number of samples in the 10 msec interval.

The positive zero crossings of the filtered signal $y[n]$ correspond to the instants of glottal closure in voiced speech and the interval between successive positive zero crossings gives us an estimation of the pitch period in voiced speech [53].

It has been observed that locations of positive zero crossings in the zero-frequency filtered signal are nearly same for close-speaking and distant speech but there were instances of spurious zero crossings which can be attributed to noise and reverberation [52]. Fig. 5.9(b) illustrates the spurious zero crossings of the zero frequency resonator output calculated from a segment of a speech signal collected at 6 ft which is shown in Fig. 5.9(a). These spurious zero crossings will lead to a wrong estimation

of pitch period which is assumed to be the time interval between two successive positive zero crossings. This has been reduced by using the property of regularity of the pitch period in successive intervals which is reflected in the narrowband spectrogram as the most dominant spectral component. A 2nd order linear prediction analysis [41] is used to estimate most dominant spectral peak ω_0 in short-time spectrum of $y[n]$ for each frame. The signal $y[n]$ is passed through an all-pole filter $P(z)$ given by

$$P(z) = \frac{1}{(1 - re^{j\omega_0}z^{-1})(1 - re^{-j\omega_0}z^{-1})} \tag{5.12}$$

where ω_0 is the location of the spectral peak, to result in $\tilde{y}[n]$ which is free of spurious zero crossings. Fig. 5.9(c) shows the refined signal of Fig. 5.9(b) where the spurious zero crossings are eliminated. The positive zero crossings of $\tilde{y}[n]$ are used for estimating the pitch period and a median filtering is performed. Fig. 5.10 shows the overall pitch extraction process where pitch is extracted from zero crossings of the output of zero frequency resonator both for close-speaking speech signal and its corresponding signal collected at 6 ft. Fig. 5.10(e) is the overall output after the refining of the signal to estimate $\tilde{F}_0$.

5.3.2 Incorporation of pitch information in speaker verification system

The pitch estimation method explained in the previous section is used for the speaker verification process. The similarity of the pitch contours of the reference and test utterances has been captured by using the optimal warping path obtained in the DTW algorithm [45], where absolute difference of the pitch frequencies for a few selected matching frames in the reference and test utterances are summed up to get the pitch score (τ). The optimal warping path is obtained from the DTW algorithm on the SSCC features. Fifty pairs of matching frames are selected such that the Euclidean distance between the feature vectors of the frame pairs are the lowest

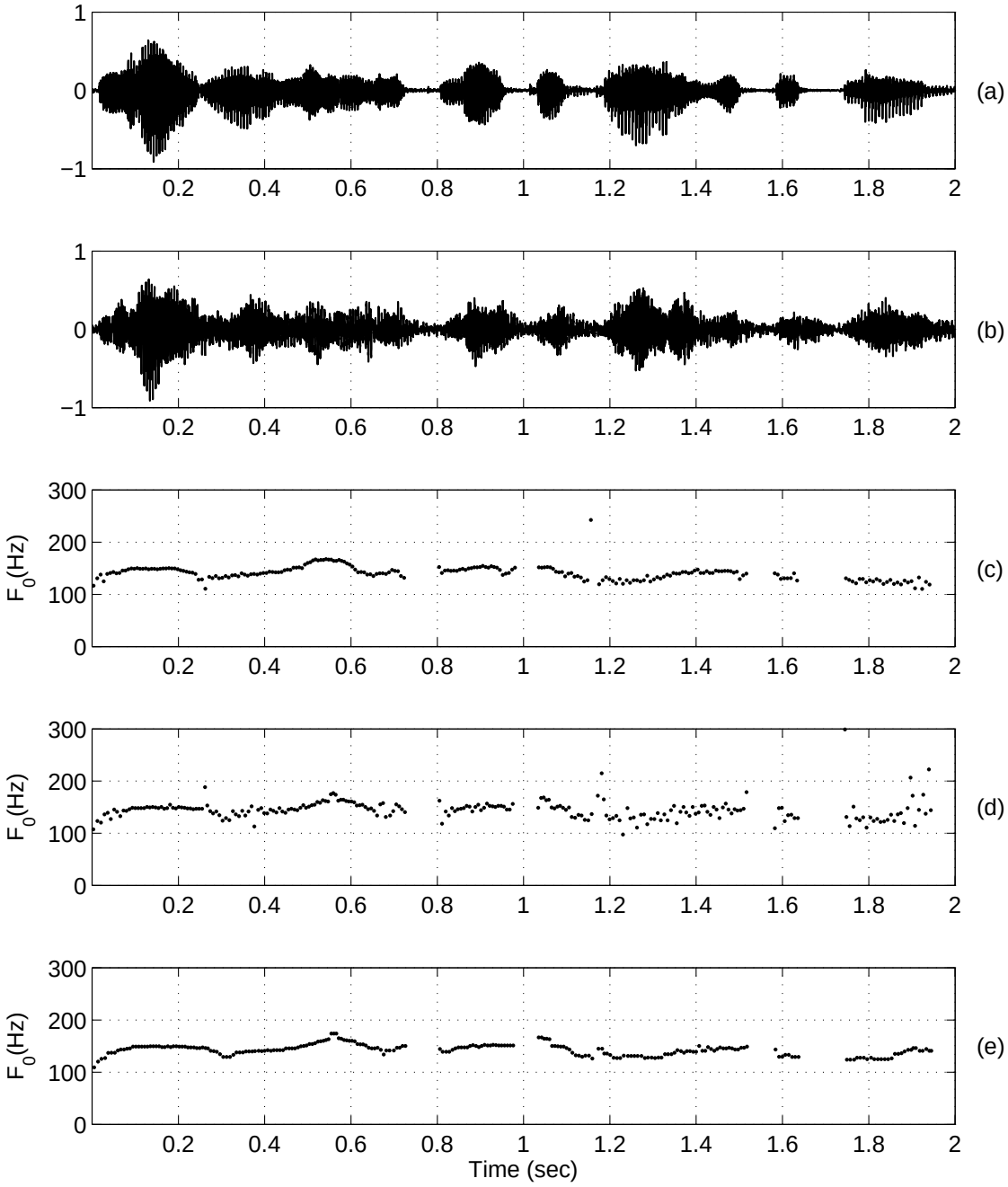

Figure 5.10: Pitch extraction process. (a) Close-speaking speech signal, (b) corresponding time aligned distant speech collected at 6 ft, (c) F_0 extracted from (a), (d) F_0 extracted from (b) and (e) refined estimate $\tilde{F}_0$ of the F_0 of distant speech.

Table 5.3: Performance of speaker verification system using SSCC feature along with pitch.

Channel	Genuine acceptance(%)	Imposter rejection(%)	Best cost	System cost
close	95.1	99.4	0.0096	0.0096
2 ft	72.6	99.7	0.0236	0.0300
4 ft	46.3	99.8	0.0399	0.0544
6 ft	42.2	99.7	0.0469	0.0598

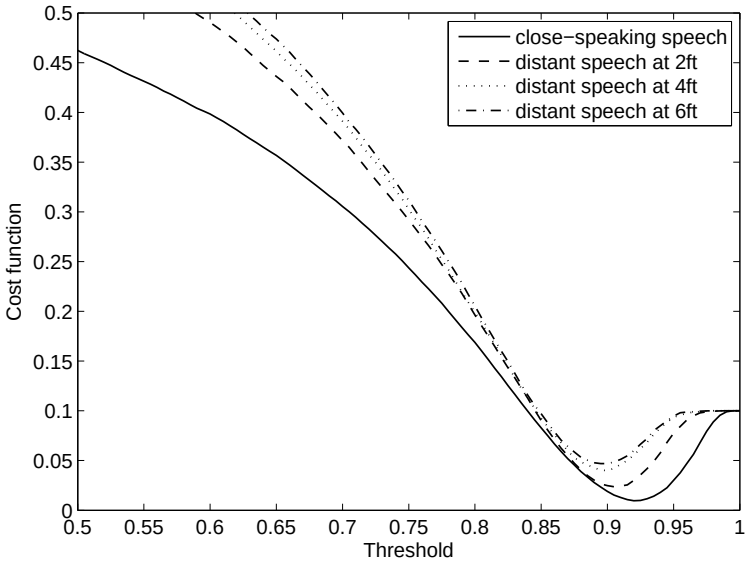

Figure 5.11: Cost function of the system using SSCC along with pitch as feature.

among all the points in the warping path. The pitch score is computed as

$$\tau = \sum_{i=1}^{L} |F_0(x(i)) - F_0(y(i))| \tag{5.13}$$

where $F_0(x(i))$ is the pitch frequency of the frame $x(i)$ of the test utterance, $F_0(y(i))$ is the pitch frequency of the frame $y(i)$ of the reference utterance and L is the number of points. This is a dissimilarity measure.

5.3.3 Performance of the system using pitch as information

A similar procedure explained in section 5.2.2 is used for the combination of SSCC and pitch information. All the scores (α, β and τ) are normalized in the range of 0

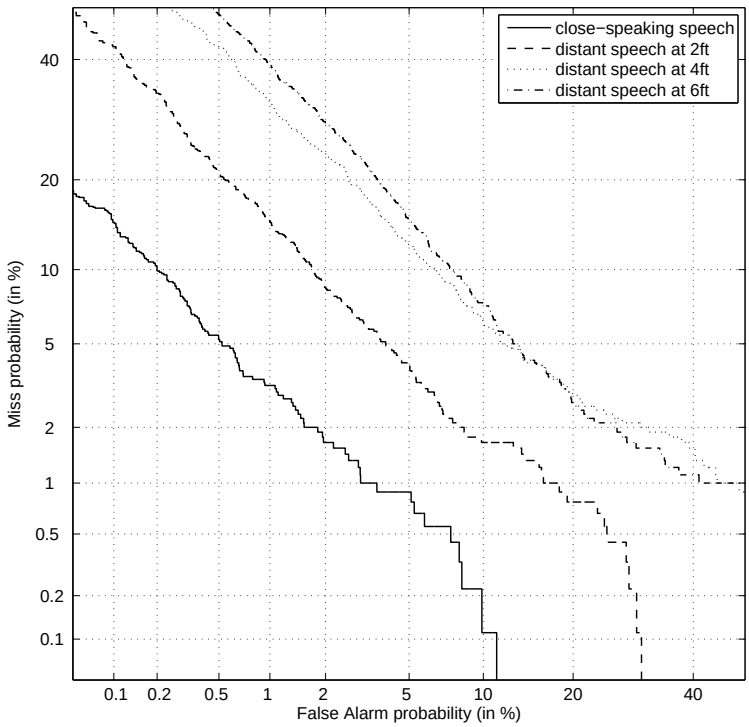

Figure 5.12: DET curve showing the system performance using SSCC along with pitch as feature.

to 1 as explained in section 3.2. The weighted sum is computed as

$$S = 0.09 \times \alpha + 0.81 \times \beta + 0.1 \times \tau \tag{5.14}$$

where α and β are the normalized scores representing SSCC feature and τ is the normalized score representing pitch information. The weights are found empirically. The best performance threshold obtained for close-speaking speech is set as threshold of the system and is used for 2 ft, 4 ft and 6ft cases. The performance of the system using SSCC feature and pitch information is shown in Table. 5.3. The significance of pitch as feature can be seen in the improved performance of the speaker verification system. An improvement in both the genuine acceptance rate and imposter rejection rate is observed.

The improvement in the system performance is reflected in the cost functions shown in Fig. 5.11. The performance is also shown in the form of a DET curve in Fig. 5.12.

Table 5.4: Performance of speaker verification system using SSCC features extracted from high SRR regions, duration and pitch information.

Channel	Genuine acceptance(%)	Imposter rejection(%)	Best cost	System cost
close	95.1	99.5	0.0085	0.0085
2 ft	76.3	99.8	0.0208	0.0252
4 ft	48.4	99.8	0.0376	0.0527
6 ft	44.3	99.8	0.0420	0.0571

5.4 Combination of features

The additional information of frame weights, duration and pitch information was able to improve the performance of the system for distant speech individually. A study is made on the combination of SSCC features extracted from high SRR regions, duration and pitch information. For this process, all the scores are normalized in the range of 0 to 1 as explained in Section 3.2. A weighted sum of all the measures is calculated as

$$S = 0.08 \times \alpha + 0.72 \times \beta + 0.1 \times \xi + 0.1 \times \tau \tag{5.15}$$

The weights are found empirically. As expected, the result of the combination of features is better than the individual features. The result is shown in Table. 5.4. Improvements can be seen both in the genuine acceptance and imposter rejection rates showing the complementary information of the features which has been extracted using SSCC feature. The convergence of the cost functions across distance has been the best in the case of combination of the features which is shown in Fig. 5.13. The performance of the speaker verification system using the combination of features is also shown in the form of DET curves in Fig. 5.14, for speech signals collected at different distances.

A channelwise comparison on the performance of the baseline system and the system using the combination of the features is shown in the form of DET curves in Fig. 5.15. Notice that the improvement in the performance is seen as the difference in the EER value of the curves. The system using the combination of features has a better performance over the baseline system in all the channels. The improvement

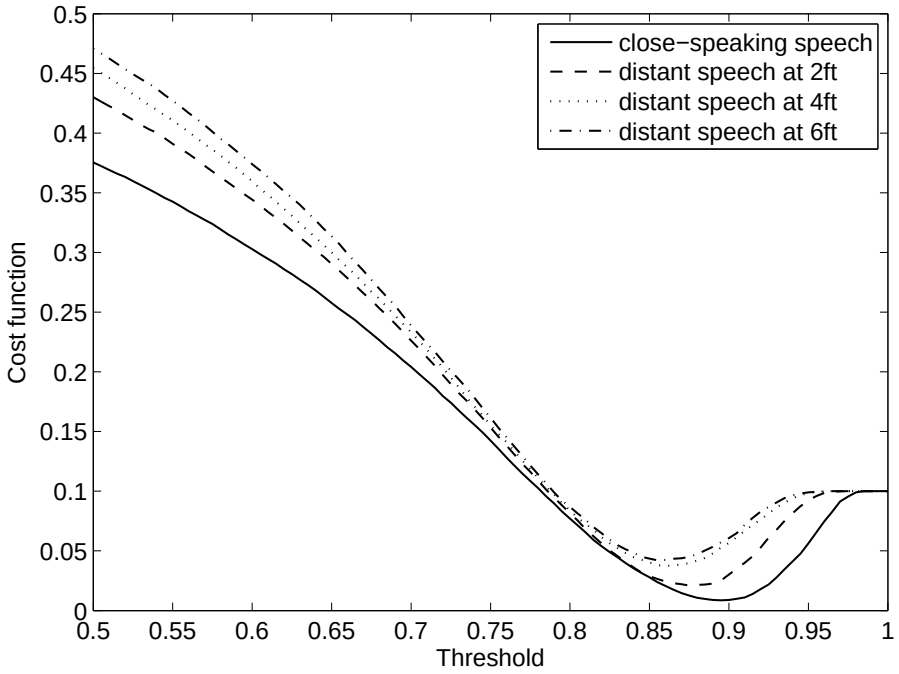

Figure 5.13: Cost function of the system using SSCC along with frame weights, duration and pitch as feature.

Table 5.5: Performance of speaker verification system using (MFCC + CMS) as features extracted from high SRR regions, duration and pitch information.

Channel	Genuine acceptance(%)	Imposter rejection(%)	Best cost	System cost
close	93.2	99.4	0.0113	0.0113
2 ft	48.1	99.7	0.0403	0.0538
4 ft	26.7	99.7	0.0546	0.0757
6 ft	23.6	99.7	0.0636	0.0786

over the baseline system in the genuine acceptance is 3.4%, 33.1%, 25.8% and 23.8% for speech signals collected at close, 2 ft, 4 ft and 6 ft respectively. The imposter rejection is maintained around 99.8% in all the cases.

A similar study is made on the combination of (MFCC + CMS) feature extracted from high SRR regions, duration information extracted from MFCC feature and pitch information extracted using MFCC feature, which is shown in Table. 5.5. It is observed that the improvement can only be seen in the case of close-speaking speech. Though duration and pitch are robust features, its information is derived from MFCC feature which has variation across distance and hence the improvement in the performance has been minimal for distant speech.

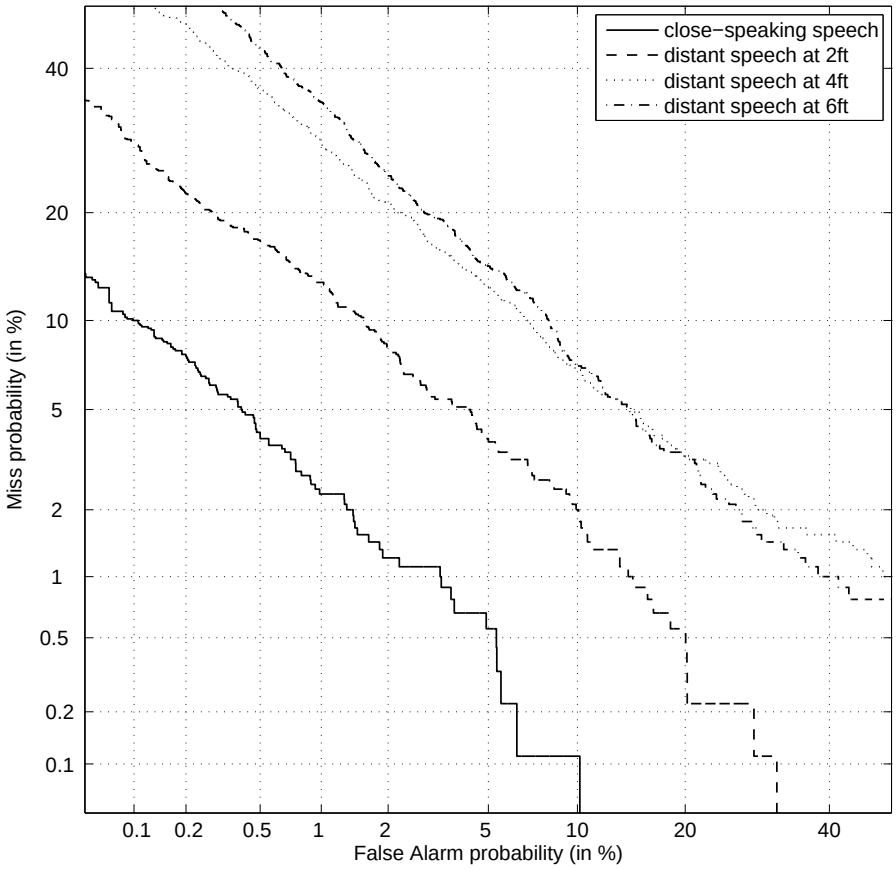

Figure 5.14: DET curve showing the system performance using SSCC along with frame weights, duration and pitch as feature.

5.5 Conclusions

In this chapter, we have seen an improvement in the performance of the system by the addition of robust features. Frame weights were derived using the high signal-to-reverberation component ratio (SRR) and only the regions with high SRR are used in the score normalization step. Duration information was calculated as the deviation from the optimal warping path of the DTW. The recent advances in the extraction of pitch in distant speech was exploited and used as an additional feature in the speaker verification system. Improvements were seen when the feature information was added individually to the system using SSCC feature. Finally a combination of all the features was studied which has shown a significant improvement in the performance of the system over the baseline.

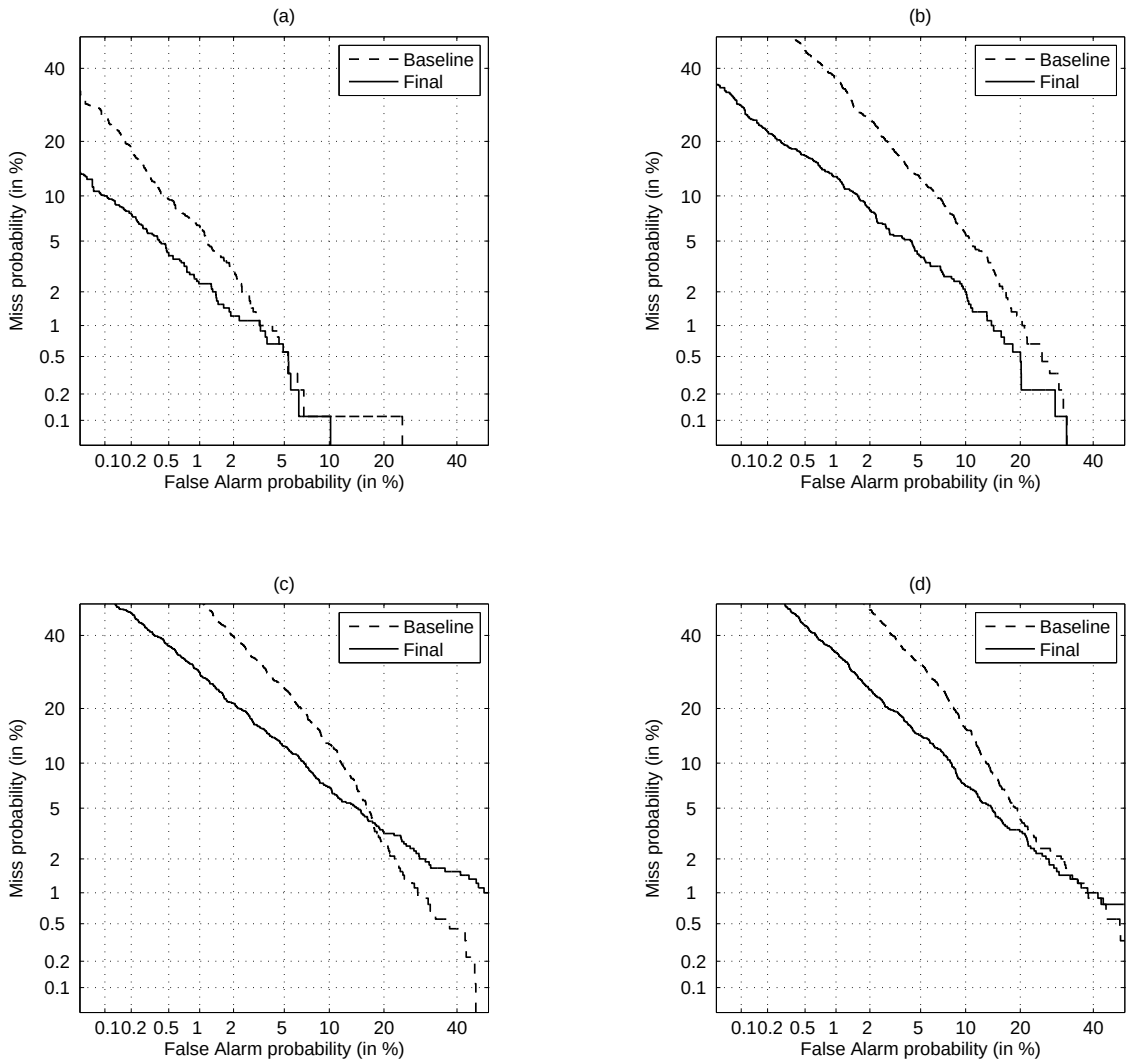

Figure 5.15: Channelwise comparison of DET curves between the baseline system and the system using combination of SSCC with frame weights, duration and pitch. (a) Close-speaking speech signal, (b) distant speech collected at 2 ft, (c) distant speech collected at 4 ft and (d) distant speech collected at 6 ft.

Chapter 6

Summary and conclusions

The aim of the present work is to develop a text-dependent speaker verification system for speech signals collected at a distance, without using any prior information of the distance. Speech signal at a given distance is collected using a single channel. Speaker verification is performed by the combining evidence from a spectral feature derived from short segments of speech signal, frame weights extracted from high SRR regions, duration and pitch. We have observed the effect of distance on the spectrum of the speech signal and it was shown experimentally that standard features like MFCC show significant variation with distance. By simultaneously recording speech signals at close, 2 ft, 4 ft and 6 ft, a database of 45 speakers is collected for an English sentence. This is later used in the performance evaluation of the speaker verification system. A standard baseline system using MFCC as feature and DTW for pattern comparison was built. It was seen that the standard methods for begin-end detection which are based on energy of the signal do not work satisfactorily for distant speech. A begin-end detection based on the strength of spectral peaks was proposed which performs well for distant speech. Improvements were made to a standard text-dependent speaker verification system by using the proposed begin-end detection. For further improvement, a method of score normalization was proposed which considers only the robust regions in speech signal. This was achieved by the fact that variation of acoustic features of a genuine speaker is lesser, when compared with a background speaker. But it was seen that performance of the

Table 6.1: Performance of speaker verification system on close-speaking speech.

Feature	Genuine acceptance(%)	Imposter rejection(%)	Cost of system
Baseline (MFCC + CMS)	91.7	99.3	0.0137
SSCC	93.4	99.3	0.0121
SSCC + high SRR	93.1	99.4	0.0114
SSCC + duration	94.1	99.5	0.0103
SSCC + pitch	95.1	99.4	0.0096
SSCC + high SRR + duration + pitch	95.1	99.5	0.0085

Table 6.2: Performance of speaker verification system on speech collected at 2 ft.

Feature	Genuine acceptance(%)	Imposter rejection(%)	Cost of system
Baseline (MFCC + CMS)	43.2	99.7	0.0588
SSCC	69.8	99.5	0.0344
SSCC + high SRR	71.5	99.5	0.0321
SSCC + duration	73.4	99.7	0.0284
SSCC + pitch	72.6	99.7	0.0300
SSCC + high SRR + duration + pitch	76.3	99.8	0.0252

baseline system decreased significantly with distance even with these improvements. To enhance the performance of the speaker verification system for distant speech, an acoustic feature derived from short segments of speech signal was proposed. The proposed feature which exploits the high signal-to-noise nature of speech signal, preserves both speech and speaker characteristics. It was shown that the proposed feature suffers lesser degradation with distance when compared to the widely used MFCC feature. It also yields a better result in speaker verification for distant speech compared to MFCCs.

To improve the performance of this system, a combination of features was used. Using the fact that the effect of distance on the speech signal is different in different regions, the performance of the current system using SSCC features was improved by considering only those regions which suffered lesser degradation. This was accomplished by using the high SRR regions to assign frame weights with higher weightage given to the frames with lesser degradation. It was seen that by considering only the regions with lesser degradation, a slight increment could be achieved in the performance of this system. Additional information of duration was applied along with the proposed feature to improve the performance. Also, the recent advance-

Table 6.3: Performance of speaker verification system on speech collected at 4 ft.

Feature	Genuine acceptance(%)	Imposter rejection(%)	Cost of system
Baseline (MFCC + CMS)	22.6	99.8	0.0784
SSCC	43.0	99.7	0.0596
SSCC + high SRR	46.2	99.5	0.0583
SSCC + duration	42.2	99.8	0.0589
SSCC + pitch	46.3	99.8	0.0544
SSCC + high SRR + duration + pitch	48.4	99.8	0.0527

Table 6.4: Performance of speaker verification system on speech collected at 6 ft.

Feature	Genuine acceptance(%)	Imposter rejection(%)	Cost of system
Baseline (MFCC + CMS)	20.5	99.8	0.0811
SSCC	38.3	99.6	0.0648
SSCC + high SRR	41.3	99.4	0.0640
SSCC + duration	38.3	99.8	0.0629
SSCC + pitch	42.2	99.7	0.0598
SSCC + high SRR + duration + pitch	44.3	99.8	0.0571

ments of pitch extraction methods in distant speech is used to further improve the performance.

Table. 6.1 shows the improvement in the system performance from the baseline system to the combination of all features for close-speaking speech. Similarly, the performance of the system for distances of 2 ft, 4 ft and 6 ft are shown in Table. 6.2, 6.3 and 6.4 respectively. It can be observed that, for each distance, a systematic improvement was achieved in the performance of the system with addition of features. A significant improvement can be seen in the performance for 2 ft distance, whereas it slightly decreases for 4 ft and 6 ft. It was seen that the extent of noise and reverberation was more in case of speech signal collected at 6 ft. For practical purposes, more focus on processing of speech signals collected at 2-4 ft is essential.

6.1 Major contributions of the present work

1. A begin-end detection algorithm is proposed on the basis of strength of the peaks in the short-time spectrum of speech signal.

2. An acoustic feature based on the short segments of speech in high signal-to-

noise ratio (SNR) regions is proposed, which performs better than traditional MFCC features for speaker verification in distant speech.

3. The feature of high signal-to-reverberation component ratio (SRR) is exploited for the extraction of acoustic features from distant speech signals.

4. The knowledge of duration and pitch is exploited to improve the performance of the speaker verification system.

5. A method of score normalization has been proposed by using the speaker-specific nature of the short-segment feature.

6.2 Scope for future work

1. It was shown that the proposed short-segment feature is able to retain the formant information. So, the time-averaged spectrogram can be used in the extraction of formant contours in distant speech.

2. Features of excitation source can also be exploited for robust speaker verification

3. The use of a binaural capturing device, similar to the humans can also be exploited for further studies.

Bibliography

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

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

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

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

[5] J. J. Wolf, "Efficient acoustic parameters for speaker recognition," *J. Acoust. Soc. Amer.*, vol. 51(6 (Part 2)), pp. 2044–2056, 1972.

[6] A. Rosenberg, "Listener performance in speaker verification tasks," *IEEE Trans. Audio Electroacoust.*, vol. 21, no. 3, pp. 221–225, Jun. 1973.

[7] J. Naik, "Speaker verification: A tutorial," *IEEE Communications Magazine*, vol. 28, no. 1, pp. 42–48, Jan. 1990.

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

[9] J. Openshaw, Z. Sun, and J. Mason, "A comparison of composite features under degraded speech in speaker recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 2, Apr. 1993, pp. 371–374.

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

[11] Leena Mary and B. Yegnanarayana, "Prosodic features for speaker verification," in *Proc. Interspeech - 2006*, Pittsburgh, PA, USA, Sep. 2006, pp. 917–920.

[12] M. K. Sonmez, L. Heck, M. Weintraub, and E. Shriberg, "A lognormal tied mixture model of pitch for prosody-based speaker recognition," in *Eurospeech97*, vol. 3, 1997, pp. 1391–1394.

[13] M. Carey, E. Parris, H. Lloyd-Thomas, and S. Bennett, "Robust prosodic features for speaker identification," in *Proc. Int. Conf. Spoken Language Processing*, vol. 3, Oct. 1996, pp. 1800–1803.

[14] K. Sonmez, E. Shriberg, L. Heck, and M. Weintraub, "Modeling dynamic prosodic variation for speaker verification," in *Proc. of ICSLP*, vol. 7, 1998, pp. 3189–3192.

[15] Farrus, Mireia and Wagner, Michael and Anguita, Jan and Hernando, Javier, "Robustness of prosodic features to voice imitation," in *Proc. Interspeech - 2008*, Brisbane, Australia, Sep. 2008, pp. 613–616.

[16] H. Gish and M. Schmidt, "Text-independent speaker identification," *IEEE Signal Processing Magazine*, vol. 11, no. 4, pp. 18–32, Oct. 1994.

[17] Q. Jin, T. Schultz, and A. Waibel, "Far-field speaker recognition," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 15, no. 7, pp. 2023–2032, Sep. 2007.

[18] I. M. Jason, J. Pelecanos, and S. Sridharan, "Robust speaker recognition using microphone arrays," in *Proc. 2001: A Speaker odyssey*, 2001.

[19] I. Zeljkovic, P. Haffner, B. Amento, and J. Wilpon, "GMM/SVM N-best speaker identification under mismatch channel conditions," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Apr. 2008, pp. 4129–4132.

[20] J. Gammal and R. Goubran, "Combating reverberation in speaker verification," in *Proc. IEEE Instrumentation and Measurement Technology Conference*, vol. 1, May 2005, pp. 687–690.

[21] Tiago H. Falk and Wai-Yip Chan, "Spectro-temporal features for robust far-field speaker identification," in *Proc. Interspeech - 2008*, Brisbane, Australia, Sep. 2008, pp. 634–637.

[22] C. Zieger and M. Omologo, "Combination of clean and contaminated GMM/SVM for far-field text-independent speaker verification," in *Proc. Interspeech - 2008*, Brisbane, Australia, Sep. 2008, pp. 1949–1952.

[23] J. Gonzalez-Rodriguez, J. Ortega-Garcia, C. Martin, and L. Hernandez, "Increasing robustness in GMM speaker recognition systems for noisy and reverberant speech with low complexity microphone arrays," in *Proc. Int. Conf. Spoken Language Processing*, vol. 3, Oct. 1996, pp. 1333–1336.

[24] T. Tazawa, T. Hatashima, N. Ohnishi, and N. Sugei, "A fully passive echo-canceler using a single microphone," in *Proc. IEEE Instrumentation and Measurement Technology Conference*, vol. 3, May 1994, pp. 1191–1194.

[25] A. Akula and P. L. D. Leon, "Compensation for room reverberation in speaker identification," in *Proc. European Signal Processing Conf.*, Aug. 2008.

[26] EDIROL R-09. http://www.edirol.com/.

[27] T. Ganchev, N. Fakotakis, and G. Kokkinakis, "Comparative evaluation of various MFCC implementations on the speaker verification task," in *Proc. 10th International Conference on Speech and Computer (SPECOM 2005)*, vol. 1, 2005, pp. 191–194.

[28] S. Nakagawa, K. Asakawa, and L. Wang, "Speaker recognition by combining MFCC and phase information," in *Proc. Interspeech - 2007*, Antwerp, Belgium, Aug. 2007, pp. 2005–2008.

[29] S. Stevens and J. Volkmann, "A scale for the measurement of the psychological magnitude pitch," *J. Acoust. Soc. Amer.*, vol. 8, no. 3, pp. 185–190, 1937.

[30] Wei Han, Cheong-Fat Chan, Chiu-Sing Choy and Kong-Pang Pun, "An efficient MFCC extraction method in speech recognition," in *Proc. IEEE International Symposium on Circuits and Systems*, 2006, pp. 145–148.

[31] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 28, pp. 357–366, Aug. 1980.

[32] H. Sakoe, "Two-level DP-matching–A dynamic programming-based pattern matching algorithm for connected word recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 27, no. 6, pp. 588–595, Dec. 1979.

[33] A. Afolabi, A. Williams, and O. Dotun, "Development of a text dependent speaker-identification security system," *Research Journal of Applied Sciences*, vol. 2, no. 6, pp. 677–684, 2007.

[34] M. Pandit and J. Kittler, "Feature selection for a DTW-based speaker verification system," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 2, May 1998, pp. 769–772.

[35] V. Ramasubramanian, A. Das, and V. Kumar, "Text-dependent speaker-recognition using one-pass dynamic programming algorithm," vol. 1, May 2006, pp. 901–904.

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

[37] C. R. Jankowski Jr., T. F. Quatieri and D. A. Reynolds, "Formant AM-FM for speaker identification," in *Proc. IEEE-SP International Symposium on Time-Frequency and Time-Scale Analysis*, Oct. 1994, pp. 608–611.

[38] T. Becker, M. Jessen, and C. Grigoras, "Forensic speaker verification using formant features and gaussian mixture models," in *Proc. Interspeech - 2008*, Brisbane, Australia, Sep. 2008, pp. 1505–1508.

[39] H. Seddik, A. Rahmouni, and M. Sayadi, "Text independent speaker recognition based on the attack state formants and neural network classification," in *Proc. IEEE International Conference on Industrial Technology*, vol. 3, Dec. 2004, pp. 1649–1653.

[40] B. Yegnanarayana, P. Satyanarayana Murthy, C. Avendano, and H. Hermansky, "Enhancement of reverberant speech using LP residual," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 1, May 1998, pp. 405–408.

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

[42] L. Ferrer, H. Bratt, V. R. R. Gadde, S. Kajarekar, E. Shriberg, K. Sonmez, A. Stolcke, and A. Venkataraman, "Modeling duration patterns for speaker recognition," in *Proc. EUROSPEECH*, 2003, pp. 2017–2020.

[43] E. Shriberg, L. Ferrer, S. Kajarekar, A. Venkataraman, and A. Stolcke, "Modeling prosodic feature sequences for speaker recognition," *Speech Communication. (Special Issue on Quantitative Prosody Modelling for Natural Speech Description and Generation)*, vol. 46, no. 3-4, pp. 455–472, 2005.

[44] F. Weber, L. Manganaro, B. Peskin, and E. Shriberg, "Using prosodic and lexical information for speaker identification," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2002, pp. 141–144.

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

[46] R. Ward and J. Gowdy, "An investigation of speaker verification accuracy using fundamental frequency and duration as distinguishing features," in *Proc. Twenty-First Southeastern Symposium on System Theory*, Mar. 1989, pp. 390–394.

[47] J. Lindh and A. Eriksson, "Robustness of long time measures of fundamental frequency," in *Proc. Interspeech - 2007*, Antwerp, Belgium, Aug. 2007, pp. 2025–2028.

[48] K. Iwano, T. Asami, and S. Furui, "Noise-robust speaker verification using F0 features," in *Proc. Int. Conf. Spoken Language Processing*, vol. 2, 2004, pp. 1417–1420.

[49] A. G. Adami, R. Mihaescu, D. A. Reynolds, and J. J. Godfrey, "Modeling prosodic dynamics for speaker recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 4, Hong Kong, China, 2003, pp. 788–791.

[50] B. Kotnik, H. Hoge, and Z. Kacic, "Evaluation of pitch detection algorithms in adverse conditions," in *Proc. 3rd International Conference on Speech Prosody*, Dresden, Germany, 2006, pp. 149–152.

[51] S. Prasanna and B. Yegnanarayana, "Extraction of pitch in adverse conditions," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 1, May 2004, pp. 109–112.

[52] G. Seshadri and B. Yegnanarayana, "Extraction of fundamental frequency from distant speech signals," *communicated to IEEE Transactions on Audio, Speech and Lang. Processing.*

[53] K. Murty and B. Yegnanarayana, "Epoch extraction from speech signals," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 16, no. 8, pp. 1602–1613, Nov. 2008.