Person Identification from their Hum with Inter-session Variability Compensation

by

Chirag R. Patel

(201011018)

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of

Master of Technology

in

Information and Communication Technology

to

Dhirubhai Ambani Institute of Information and Communication Technology

DA-IICT

May, 2012

Declaration

This is to certify that

(i) the thesis comprises my original work towards the degree of Master of Technology in Information and Communication Technology at DA-IICT and has not been submitted elsewhere for a degree,

(ii) due acknowledgement has been made in the text to all other material used.

Signature of Student

Certificate

This is to certify that the thesis work entitled *"Person Recognition from their Hum with Inter-session Variability Compensation"* has been carried out by Chirag R. Patel (201011018) for the degree of Master of Technology in Information and Communication Technology at this Institute under my/our supervision.

Dr. Hemant Patil

Acknowledgements

I acknowledge my guide Dr. Hemant Patil. He guided me and provided sufficient time for discussion. He continuously motivated me for this work and monitored the progress of my work. I am thankful to Prof. M. V. Joshi and Prof. L. Pillutla for their suggestion and feedback on this work. I would also thank PhD scholar Mr. Maulik C. Madhavi. He gave me support and guided throughout this work. I am very thankful to the DA-IICT B.Tech and M.Tech students who helped me during the data collection. The data collection for this thesis was very tough task, which became easier by their help. They provided me sufficient time for my recordings of all the sessions. I would also like to thank Prof. Sanjeev Gupta who allowed me to use radio room for recording humming data. I would also like to thank DA-IIT faculty members for their great and valuable teaching. I learnt so many things here. Also DA-IICT RC team helped me to get reference material during this thesis work. I would like to thank my family members. They supported and motivated me throughout my study.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AM-FM	Amplitude Modulation Frequency Modulation
ASR	Automatic Speaker Recognition
DCT	Discrete Cosine Transform
DET	Detection Error Trade-off
DTW	Dynamic Time Warping
EER	Equal Error Rate
ESA	Energy Separation Algorithm
FA	False Alarm
FFT	Fast Fourier Transform
FR	False Rejection
GMM	Gaussian Mixture Model
HMM	Hidden Markov Model
HTD	Hilbert Transform Demodulation
LPC	Linear Prediction Coefficients
LPCC	Linear Prediction Cepstral Coefficients
MFCC	Mel Frequency Cepstral Coefficients
Q-Feature	Q- Factor based Feature
ROC	Receiver Operating Characteristic
TEO	Teager Energy Operator
VQ	Vector Quantization
VTEO	Variable length Teager Energy Operator
VTMFCC	Variable length Teager Energy based MFCC

Abstract

In this thesis, design of person recognition system from their hum is discussed. The emphasis is given to the inter-session variability of the recognition system. Standard database is not available for the inter-session variability of humming-based person recognition systems. Therefore, humming database of 50 subjects is collected in two training and six testing sessions.

The MFCC (Mel Frequency Cepstral Coefficients) is the state-of-the-art feature set in the field of speech and speaker recognition systems. In this thesis, another cepstral feature *viz.*, VTMFCC (Variable length Teager energy based MFCC) is used along with MFCC. VTMFCC captures the vocal source information. Two modulation-based features, *viz.*, AM-FM and Q-features are introduced in this thesis. The performance of all of the four features in multi-session environment is evaluated using discriminately-trained polynomial classifier. Polynomial classifier uses out-of-class information while creating person- specific person model.

Inter-session variability degrades the performance of person recognition systems due to difference in training and test sessions. This variability can be classified as intrinsic variability and extrinsic variability according to its source of origin. Inter-session variability due to speaker's health, aging, emotional state, etc. is called intrinsic inter-session variability. The session variability due to environment conditions, noise, change in microphone and acoustic channel is called extrinsic inter-session variability. The inter-session variability degrades the performance of all four features, *i.e.*, MFCC, VTMFCC, AM-FM and Q-feature. The difference in % EER (Equal Error Rate) of particular test session to base test session is

used as the inter-session variability measure. The base test session is a test session which is collected with the training session.

In this thesis, two new approaches have been proposed for the compensation of inter-session variability, *viz.*, *feature-level fusion* and *model-level fusion*. These two approaches reduce the degradation in the performance of person recognition system due to inter-session variability and make the system robust.

1 Introduction

Speaker recognition means recognizing the persons from their voices with the help of machines. Voice signal production is a very complex process and involves many organs. Two persons cannot have similar voice because their vocal tract shape, larynx size and other parts of their voice production mechanism are very different. In addition to these physical differences, each speaker has his or her own way of speaking, *i.e.,* use of particular accent, pronunciation style, etc. Therefore, voice can be used as a biometric.

Human ear can identify speakers by listening them, which is called *auditory* speaker recognition. When human expert recognize persons using various descriptive features like spectrogram, waveforms transition curves, etc. then it is called *semi-automatic* speaker recognition. The speaker recognition process is called *fully automatic* when the whole recognition process is performed by the machine without the help of human [1].

Speaker recognition is a multidisciplinary biometric problem. Understanding of acoustic, signal processing and pattern recognition is necessary in this biometric. There are many other state-of-the-art biometric systems available like iris, fingerprint and retina scans. All of the above biometrics are image-based. A dedicated hardware is required to scan them. Thus, collection of above biometric is costly. In addition, above biometric techniques are not very user-friendly. Speech is primary mode and powerful form of human communication. Speech biometric is a voice signal-based biometric; only a microphone is required to input it to machine and is very user-friendly and biometric operation can be performed on someone without his or her knowledge.

Voice biometric has many applications such as granting access to a particular device, authorized area or information and forensic applications. Voice can be transmitted over telephone network so this biometric opens a new field of reliable long distance authentication such as internet banking transaction, mobile banking, online information retrieval, teleshopping, etc [2]. In forensic speaker recognition, the speakers can be considered as non-cooperative as they do not intend to be recognized. On the other hand, in speaker recognition based services and access control, the users are considered as cooperative [3].

1.1 Speaker Recognition as a Pattern Recognition Problem

Speaker recognition is a pattern recognition problem and has generally two basic steps, *viz.*, feature extraction and classification. Let us start with the introduction of pattern recognition. An example of pattern recognition is object recognition performed daily by human beings. Humans recognize object from its characteristics and assign it a category or class. For example, we have an object called fruit. Depending on its characteristics (size, shape, smell, weight, color etc.) we will classify it and assign a class (orange, apple, pineapple, etc.) to it. In pattern recognition discipline, objects are known by generic term patterns. The distinct characteristic of objects by which we classify them is described as feature. The definition of feature can be given as below:

"*Feature can be defined as a distinctive aspect, quality or characteristic which can be symbolic (color) or numeric (height).The combination of such 'n' features is called n-dimensional feature vector''* [4].

The *n*-dimensional space spanned by *n*-dimensional feature vectors is defined as feature space. Objects are now identified to points in feature space. In the present example, color and shape are used as features. Hence, our feature vector is as follows

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x_1} \\ \boldsymbol{x_2} \end{bmatrix} = \begin{bmatrix} \boldsymbol{colour} \\ \boldsymbol{shape} \end{bmatrix}. \tag{1.1}$$

The *2*-dimensional features of two objects (apple and orange) are shown in Figure 1.1. The blue circles show the features of one object (*e.g.*, apples), while the red crosses show the features of another object (*e.g.*, oranges). Let us consider that a machine is designed to separate apples from orange using the described features. The machine is first trained with known patterns (objects), whose true class (*i.e.*, orange or apple) is known. These known patterns are called training patterns or training feature vectors. Now, an unknown fruit is presented to the machine and it is supposed to recognize whether it is orange or apple. Machine then computes a distance between features of unknown fruit and features of training model and then assigns it to class of apples or oranges. This task is called pattern classification.

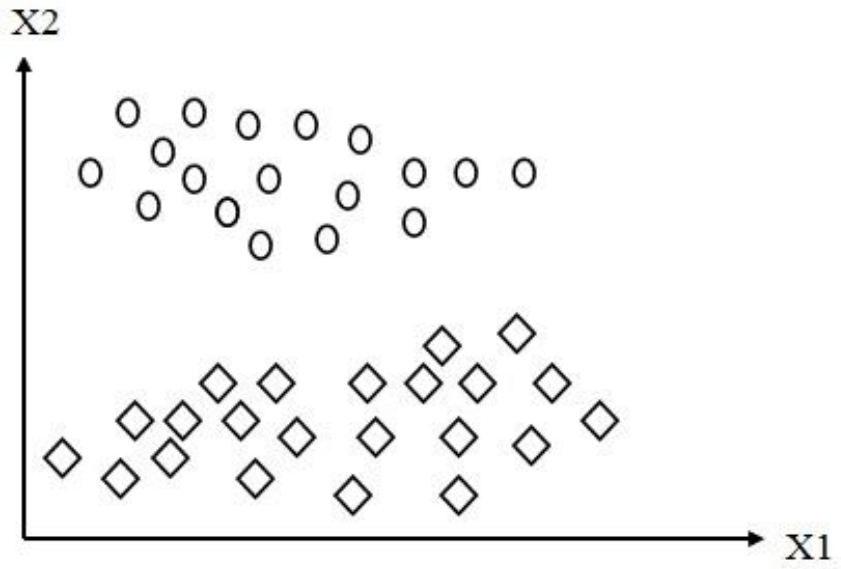


Figure 1.1 Features in *2-D* feature space [5]

The blue line shown in Figure 1.2 is a line of demarkation. The pattern recognition task of assigning an object to a class is said to be a classification task [6]. The basic issues in pattern recognition problem are as follows [7] :

- How to select the class-specific features? This is feature generation stage and is problem-dependent. They should be selected such that separation between classes is very high, so that even a simple pattern classifier can classify them.
- How many features should be used? This is feature selection stage. In general, large numbers of features are generated and best of them are selected.
- The next important task is to design a suitable classifier. The blue curve shown in Figure 1.2 is a straight line. This curve should be optimum and need not be straight. The classifier design should be done with respect to some *optimality criteria*.
- After the design of classifier, performance assessment should be performed, *i.e.,* classification error rate is calculated. This is system evaluation stage.

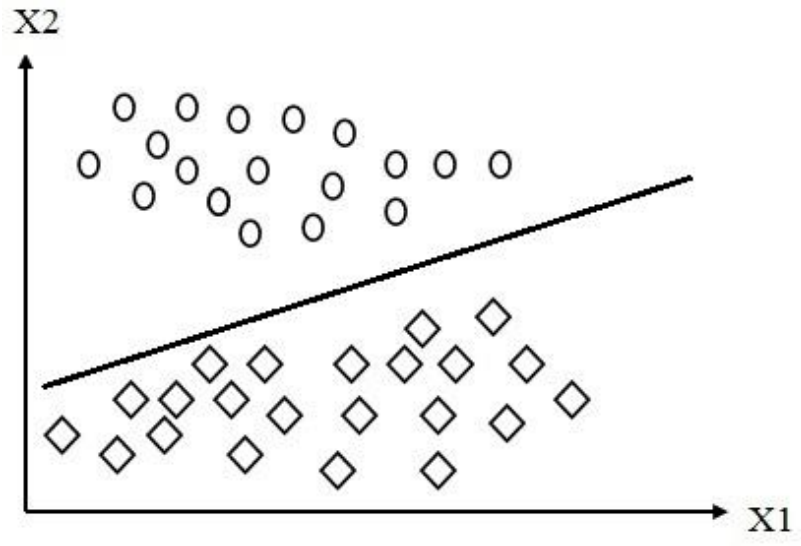


Figure 1.2 Classification task in pattern recognition [5]

Figure 1.3 shows various stages of the pattern recognition system. As shown in the figure these stages are not independent. On the contrary, they are interrelated and depending on the results, one can go back to previous stages and redesign them in order to improve overall system performance.

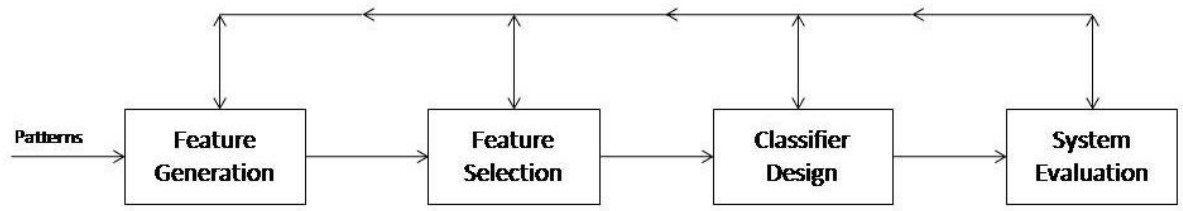


Figure 1.3 Various stages of pattern recognition system [7]

In practice, classification systems are not perfect and produce errors for several reasons [6]

- The features used are insufficient.
- The training features used to train the classifier are not sufficiently representative.
- The classifier is not accurate enough to separate classes.
- There is an intrinsic overlap between the classes that no classifier can separate it.

In automatic speaker recognition (ASR) systems, the machine attempt to classify the input feature vector and assign it to the different classes (speakers). The objectives of research in the field of ASR are the generation of features those are speaker-specific, selection of feature depending on the problem, and design of classifier that give optimum performance.

1.2 Motivation

The research on ASR system is usually performed on a corpus collected in almost ideal conditions (*e.g.,* laboratory speech, read speech, etc.). Performance of such system degrades if these systems are used on the data collected in the real-world. This degradation may be due to several factors such as noise, channel mismatch, room acoustics, aging, health of subjects, etc. These factors should be considered while designing the ASR system to make the system robust. Here, an attempt is made to consider the effect of inter-session variability on the performance of the ASR system. The inter-session variability is the performance degradation due to difference in training and testing sessions. Inter-session variability occurs due to change in the health of person over session, aging effect on speech, change in environmental conditions, etc. The objective is to study the degradation due to session variability and explore techniques to reduce this degradation.

1.3 Overview of ASR

A human cannot identify a person unless he or she knows the person from previous encounters. Similarly, ASR system cannot identify a person unless it is trained with the voice of that person previously. Therefore, an ASR system has two modes, *viz.,* enrollment and recognition. In enrollment mode, user provides a voice signal to the system to train it. In recognition mode, user provides another voice signal, which is compared with the model prepared from the enrolled signal and decision is taken.

Figure 1.4 shows the architecture of the speaker recognition. Speech signal input is recorded and is converted to a digital form. Some signal processing is performed to improve the quality of speech signal. The feature extraction is performed to both training and testing. The feature extractor module (*i.e.,* algorithm) converts the digital speech into feature vector. Features provide more stable, robust and compact representation of signal than raw input

signal. Feature extraction is basically a data reduction process that captures speaker-specific information at low data rate. In recognition mode, features are extracted from unknown speaker's speech. In pattern matching block, features of the test data is compared with the model stored in database and score is generated. Finally, a decision is taken based on the score generated by pattern matching.

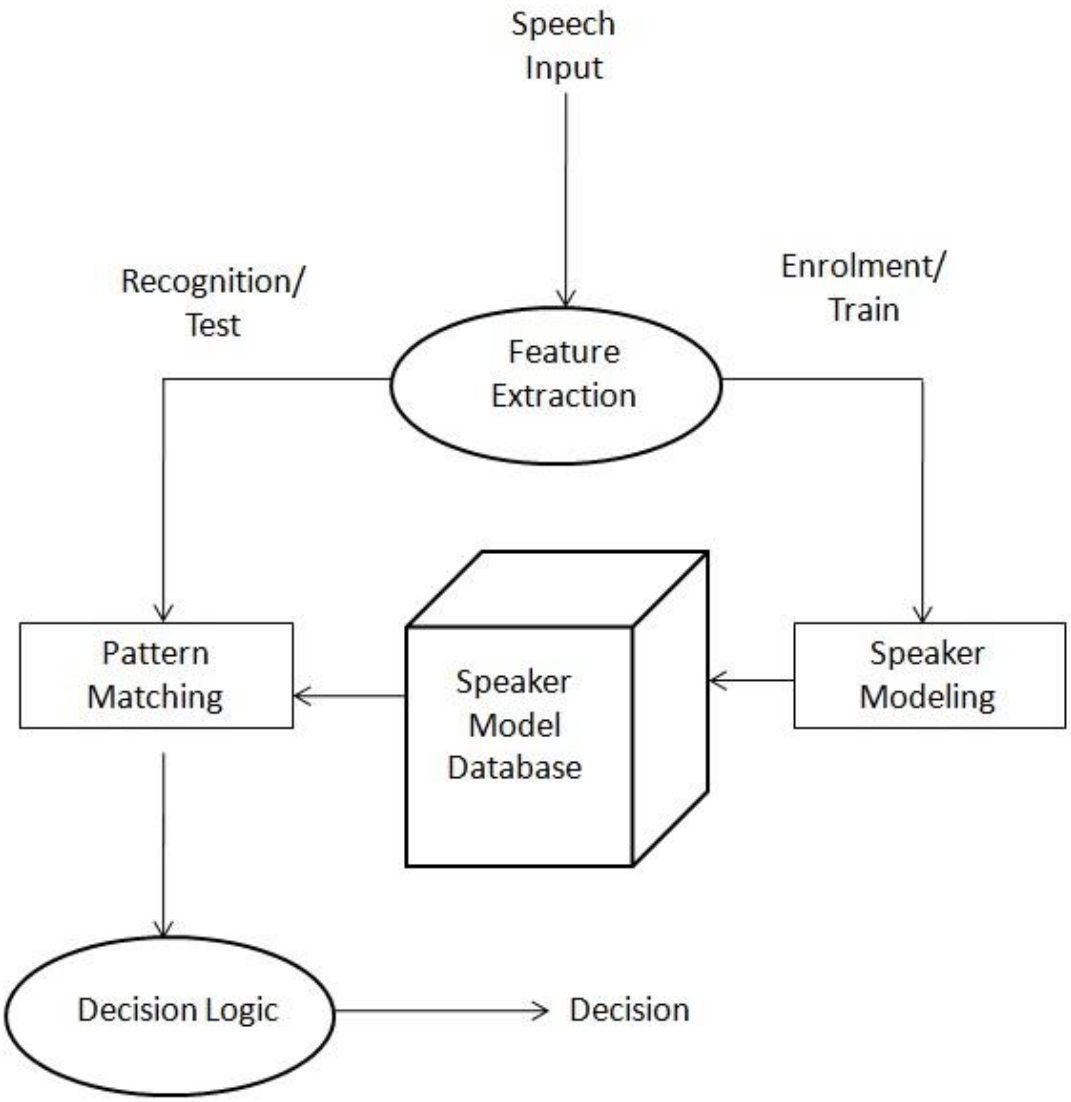


Figure 1.4 Architecture of ASR system

Speaker recognition is usually classified as *speaker identification* and *speaker verification*. Both can be further classified as *text-dependent* and *text-independent*. Speaker identification can be of *open set* or *closed set type* [8].

1.3.1 Speaker Identification *vs.* Speaker Verification

In speaker identification task, unknown speaker is compared against the database of ***N*** known speakers and the best matching speaker is returned by the machine as a decision. In verification task, it is to be decided whether the speech produced is of the claimed speaker or not. An identity claim (*e.g.*, roll number of student) is made prior to the recognition task. Therefore, the unknown speaker is compared against the model of the claimed speaker only. If the matching score is above some threshold then the speaker is accepted as a genuine speaker, otherwise it is rejected. Identification task is apparently much more difficult than verification task, as much more comparisons are involved in identification task. As number of registered speaker (population size) increases, the performance of speaker identification system decreases (rate of false decision increases) [9], [10]. On the other hand, the performance of the person verification task, is unaffected by the population size, since only claimed speaker is compared.

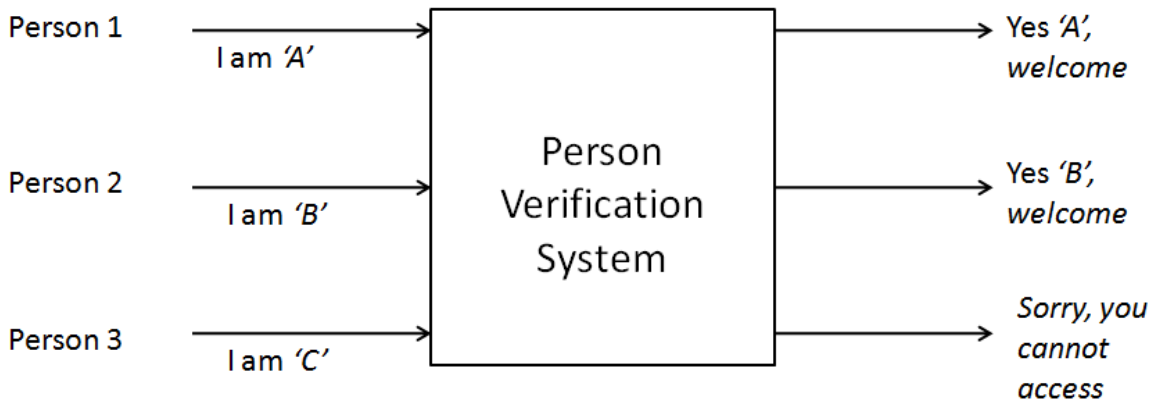


Figure 1.5 Speaker verification system

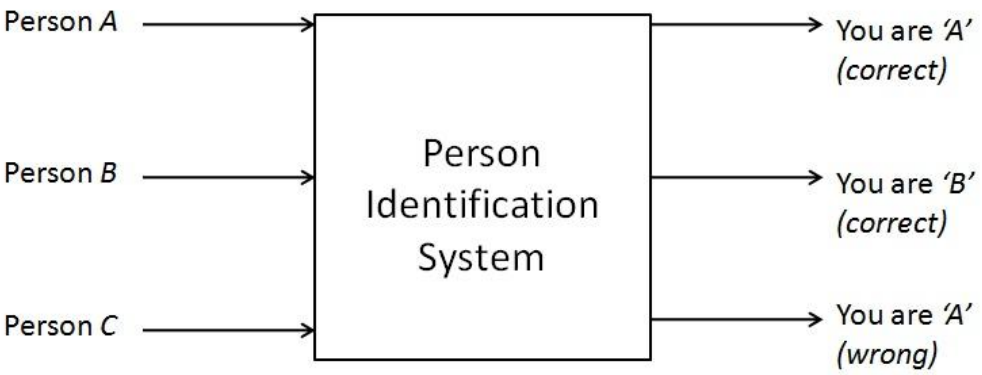


Figure 1.6 Speaker identification system

1.3.2 Text-Independent and Text-Dependent Tasks

A text-dependent system requires training and testing text to be same and known. The known text can be used in security applications to improve performance and reliability. The advantage of the text-dependent recognition is that the intruder cannot know what the phrase is. The recognition decision is the combination of two tasks, *viz.*, utterance verification (speech recognition) and speaker verification. Dynamic time warping (DTW) can be used to align the time-axis of input speaker to the time-axis of reference model [11]. The recognition system can be made such that it prompts a new keyword every time the system is used. The keyword can be displayed as characters or spoken by the system as a synthesized voice. As vocabulary is unlimited, intruder cannot know in advance what the sentence will be requested. This text-prompted recognition system thus rejects recorded speech. A text-independent speaker recognition system places no constrain on the text to be used for training and testing. The advantage of such system is that person can be recognized from any form of utterance and hence this system is user-friendly. Moreover, person can be recognized even without knowledge of him or her during normal conversation (*e.g.*, telephone banking, online

shopping etc.). Text-independent systems are more difficult to build. In this thesis, text-independent case is considered.

1.3.3 Open Set and Closed Set Identification

The speaker identification task can be further classified into open set and closed set identifications. If the unknown speaker is one of the registered speakers, the identification task is a closed set problem. If the unknown speaker can be from outside of the registered speakers, identification task is open set problem. In closed set problem, system makes a decision simply choosing best matched speaker from speaker database even if it produces poor score. In open set problem, the system has pre-defined threshold, so if score falls below the threshold level, decision is made that the unknown speaker is none of the enrolled speaker. Open set identification systems are difficult to build. In this thesis, person recognition in closed set mode is considered.

1.4 Hum as Biometric

Hum is created by keeping oral cavity closed and passing the airflow from the nasal cavity as shown in Figure 1.7. Hum is basically a nasal sound. Small intra-speaker variations and large inter-speaker variation is desired in speech biometric.

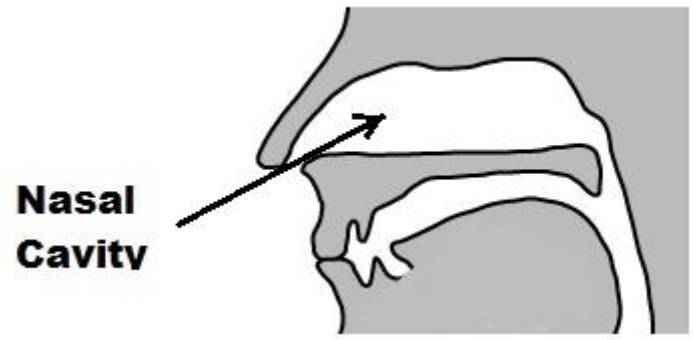


Figure 1.7 Nasal sound production [12]

Nasal sounds have velar movement that is beyond the scope of speaker, small intra-speaker variation and large inter-speaker variations that make it interesting as a biometric [13], [14]. Hum can be produced by a child and also by a deaf person, so it is more universal than speech. The hum of a person can be used to retrieve songs from multimedia database. The system can be developed such that melody tracking feature identify which song is to be retrieved and speaker-specific features verify the authenticity of person. So, only a subscribed user can access the database.

1.5 Prior Work

Hum as a biometric was first used proposed by Patil *et. al.* [15], [16], [17] and subsequently by Jin *et. al.* [9] . Query-by-Humming system was developed by Ghias *et.al.* [18]. Jin found that pitch is not conductive to person verification and identification as it is highly melody-dependent [9]. The performance of MFCC was found better than linear prediction-based features such as LPC and LPCC [16], [15]. The above studies were conducted on small population size. The study on large population was done by Madhavi. [10]. Madhavi conducted study on the population size of 170 and used Variable Length Teager Energy Operator (VTEO) based features like VTMFCC. He also exploited the significant of phase information in humming-based person identification [19].

1.6 Objective of this Thesis

The limitation of all the above studies was that they were done on single session database. It is well known fact that in voice-based biometric that if recording of training data and testing data has been done on different sessions, performance of the system degrades. An attempt is made to address inter-session variability problem in this thesis.

1.7 Organization of Thesis

The thesis is organized as described in this paragraph. Chapter 2 describes the different features used in this thesis. It describes the feature extraction scheme for MFCC, VTMFCC, AM-FM and Q-features. Chapter 2 also gives some introduction to classifiers and performance evaluation measures. Chapter 3 describes data collection and corpus design for this thesis. Chapter 4 describes the performance of collected data on the features described in chapter 2. Inter-session variability of collected data is also shown. The effect of number of filters in filter-bank, length of feature vector and effect of analysis window is studied. Chapter 5 proposes the methods for compensation of inter-session variability. The two proposed methods called feature-level fusion and model-level fusion is described. Chapter 6 concludes the thesis and provides direction for future research.

1.8 Chapter Summary

In this chapter, introduction of speaker recognition problem is discussed. The classification of ASR system according to various factors is discussed. The problem statement and the motivation behind choosing present study are discussed. Prior work in humming based recognition system is also discussed in brief.

2 Features and Classifiers

2.1 Introduction

This chapter discusses different types of features and classifier techniques used by speaker recognition systems. The speech signal carries the information about text spoken and about the speaker. It is difficult to separate out linguistic information about speech and the speaker related information from given speech signal. Therefore, most of the features are common to both speech and speaker recognition systems. Speech signal provides different level of information as shown in Figure 2.1. Humans use this different level of information to understand the speech. This different level of information and features derived from this information is described in following Section. The pre-processing system needed before feature extraction is also described in this chapter. Then different features used in this thesis are described. A polynomial classifier is used for this problem which is discussed in detail in this chapter. The criteria for performance evaluation of speaker recognition system are described in the last section.

2.2 Types of Features

The speech spectrum reflects the geometry of speech production system, so the features derived from the spectrum have been proven to be most effective. Different levels of information are conveyed through the speech as shown in Figure 2.1. The high level features are derived from learned traits. The high level features are less likely to be affected by channel mismatch or noise [20], [21]. Learned traits such as semantics, dictions, pronunciations, etc. are difficult to extract, but offer a great potential. The fusion between features derived from learned and physical traits can make the system more robust.

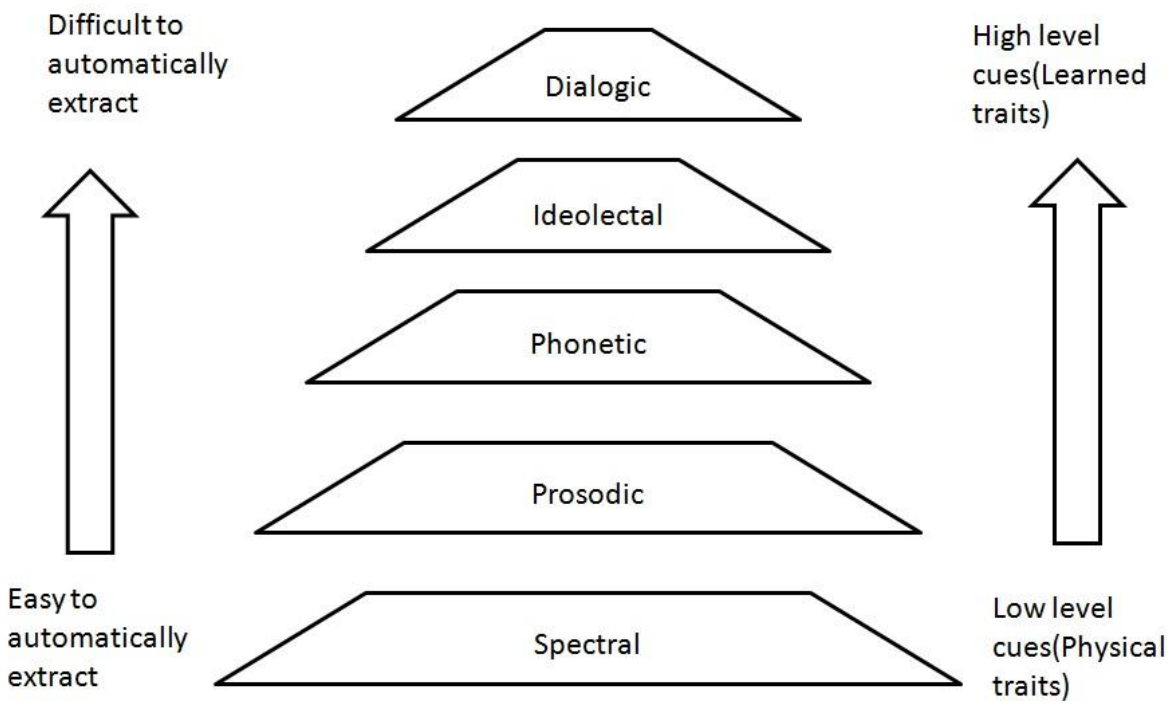


Figure 2.1 Different level of information for speaker recognition [20], [22]

2.2.1 Spectral

The anatomical structure of vocal tract is very easy to extract automatically. Different speakers have different spectra for the same sound. The features are extracted by temporal methods such as Linear Prediction Coefficients (LPC) or frequency-domain methods such as Mel Frequency Cepstral Coefficients (MFCC). The very nice property of spectral features is the logarithmic scales, which mimic the functional properties of human ear. The speech signal is generated in such a way that ear can recognize it easily. Therefore, the analysis that follows the functional property of human ear yields better results.

2.2.2 Prosodic

The measurement of stress, accent and intonation is called *prosody*. The features derived from prosody of speech are called prosodic features. The prosodic features can be

derived from the measurements of speech, energy and duration information. The prosodic features are less affected by the transmission channel mismatch as compared to spectral features. Therefore, it can be used as a complement to spectral features in situations where microphone or transmission channels can change. Thus, the use of prosodic features can improve the *robustness* of the system.

2.2.3 Phonetics

It is possible to characterize speaking pattern and speaker-specific pronunciations using phone patterns. It is well known that the same phoneme can be pronounced in a different way without changing the meaning of the whole word. This variability in phoneme pronunciation can be used as a feature. This feature captures dialectal characteristics of the speaker which includes geographical and cultural traits.

2.2.4 Idiolectal

The speaker can be recognized by the way of usage of words in the language. It is well known fact that some persons use specific words repeatedly. An advantage of this method is that it not only considers the vocabulary specific to speaker but also, context and short-time dependency between words, which is more difficult to imitate.

2.2.5 Dialogic

When two or more speakers are in dialogue, we would like to segment part corresponding to each speaker. Conversational patterns are important for determining when speaker change has occurred and from grouping speech from the same speaker.

2.3 Criterion for Feature Selection

The following points should be kept in mind before selection of features for the speaker recognition problem [23], [24].

- Features should have less intra-speaker variability and high inter-speaker variability.
- Features should be difficult to mimic.
- The health of the speaker and long-term change in voice should not affect the features.
- Features should frequently occur in speech signal.
- Features should be robust against noise and channel mismatch.

It is very unlikely that one feature fulfills all of the above ideal characteristics. However, we have many features available, which fulfil different sets of criterion discussed above. Thus, by combining one or more features, we can make speaker identification system robust.

2.4 Pre-processing

The digital speech signal is first processed and then features are extracted. This pre-processing includes framing, windowing and pre-emphasis operations.

2.4.1 Framing

Speech production system is a time-varying system due to continuous articulatory movements. Therefore, speech signal must be analyzed in short segments, called frames. This is due to the fact that speech signal is assumed to be *stationary* for a small period of time. The selection of particular frame length is very crucial, due to trade-off between time and

frequency resolutions. Longer frames give good frequency resolution and poor time resolution. On the other hand, shorter frames give poor frequency resolution, but good time resolution. The frame should be long enough to give good frequency resolution. On the other hand, it should be short enough so that it can capture local cepstral properties. Typically, a frame of 10 to 30 *ms* is used with overlap of 25-50 % of frame length. Overlapping is done to avoid loss of information.

2.4.2 Windowing

Windowing is done to reduce the effect of spectral artifacts that results from the framing process. Windowing operation in time-domain is defined as the point-wise multiplication of frame and the window function. A good window function should have narrow main lobe and low side lobes in their transfer function. It should taper down smoothly at the edges of frames to remove the effect of discontinuity. Window functions commonly used in analysis of speech signal are Hamming window and Hanning window. Hamming window function is defined as [25]:

$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right) \tag{2.1}$$

Hanning window function is defined as [25] :

$$w(n) = 0.5 - 0.5cos\left(\frac{2\pi n}{N-1}\right) \tag{2.2}$$

In this thesis, Hamming window is used as a window function.

2.4.3 Pre-emphasis

Pre-emphasis is done to amplify or boost high frequency regions of speech signal. Voice signal has steep roll-off in the high frequency region. Therefore, pre-emphasis is used to balance this roll-off. For voiced sounds, the glottal source has an approximately -12 dB/octave slope. However, when acoustic energy radiates from lips, the spectrum gets +6dB/octave boost. As a net result, when speech signal is recorded with a microphone, it gets -6dB/octave slope compared to the original spectrum of vocal tract [26]. To eliminate this effect and to prevent the lower frequency components to dominate the signal, pre-emphasis is performed before feature extraction.

In short, pre-emphasis is simply high pass filtering operation. Transfer function of a typical pre-emphasis filter is given by

$$H(z) = 1 - \alpha z^{-1}\,, \quad 0.9 \leq \alpha \leq 1.0, \tag{2.3}$$

where α controls the slope of the filter. The impulse response of the above pre-emphasis filter is given by

$$h[n] = \{1, -\alpha\}. \tag{2.4}$$

The filter is a first-order differentiator, *i.e.*,

$$y[n] = s[n] - \alpha \cdot s[n-1], \tag{2.5}$$

where $s[n]$ is input signal and $y[n]$ is output signal. In this thesis, the value of α is taken as 0.97.

2.5 Cepstral Features

There are various features that have been developed in speech signal processing domain by various researchers. Cepstral features are the most successful features in speech and speaker recognition tasks.

2.5.1 Mel Frequency Cepstral Coefficients (MFCC)

The speech signal is first pre-processed by framing, windowing and pre-emphasis. The signal is then converted into frequency-domain by taking FFT. The magnitude spectrum of the signal is then warped to Mel scale filterbank. The Mel scale is based on human auditory system. It is known that low frequency spectrum of speech signal is more important for human to recognize person than high frequency components. Mel scale gives more resolution to the lower frequencies and low resolution to higher frequencies. By doing this, more importance is given to the low frequency spectrum in sub-band processing. After warping the magnitude spectrum to the Mel scale, sub-band energy is calculated followed by the logarithmic operation. Sub-band energy (the energy of response of each filter of the filter-bank) is calculated. Finally, DCT (Discrete Cosine Transform) operation is performed to get the MFCC feature vectors. The number of coefficients in MFCC feature vector is usually between 12 to 24. Some studies drop the zeroth coefficient because it is the average log-energy of the frame. The block diagram is shown in Figure 2.2

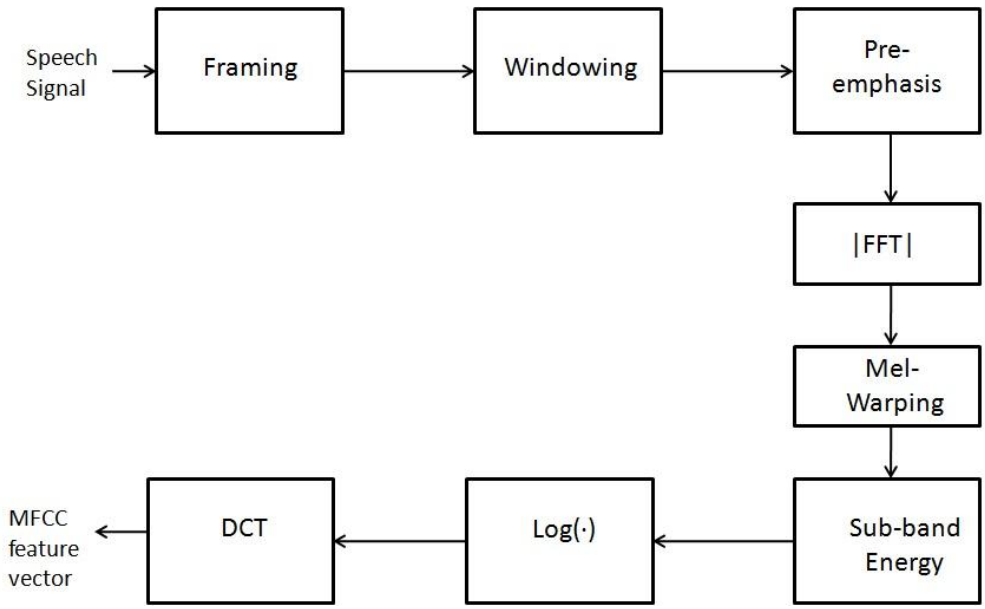


Figure 2.2 Block diagram of MFCC feature extraction scheme [27]

2.5.2 Variable length Teager Energy based MFCC (VTMFCC)

VTMFCC is acronym of the Variable length Teager energy based Mel Frequency Cepstral Coefficients. It is a cepstral feature like MFCC, but here energy is calculated in time-domain. Energy calculated in time-domain is not an absolute energy but VTEO energy [17]. TEO and VTEO are used to calculate energy of the signal.

TEO (Teager Energy Operator) is a running estimate of signal energy. A unit 10 Hz signal is said to have the same energy as the unit 1000 Hz signal, but the energy required to produce both the signal is different. The energy required to produce 1000 Hz signal is much larger than the energy required for generating 10 Hz signal. The energy required to generate the signal depend on amplitude and frequency of the signal. The block diagram of VTMFCC feature extraction scheme is shown in Figure 2.3.

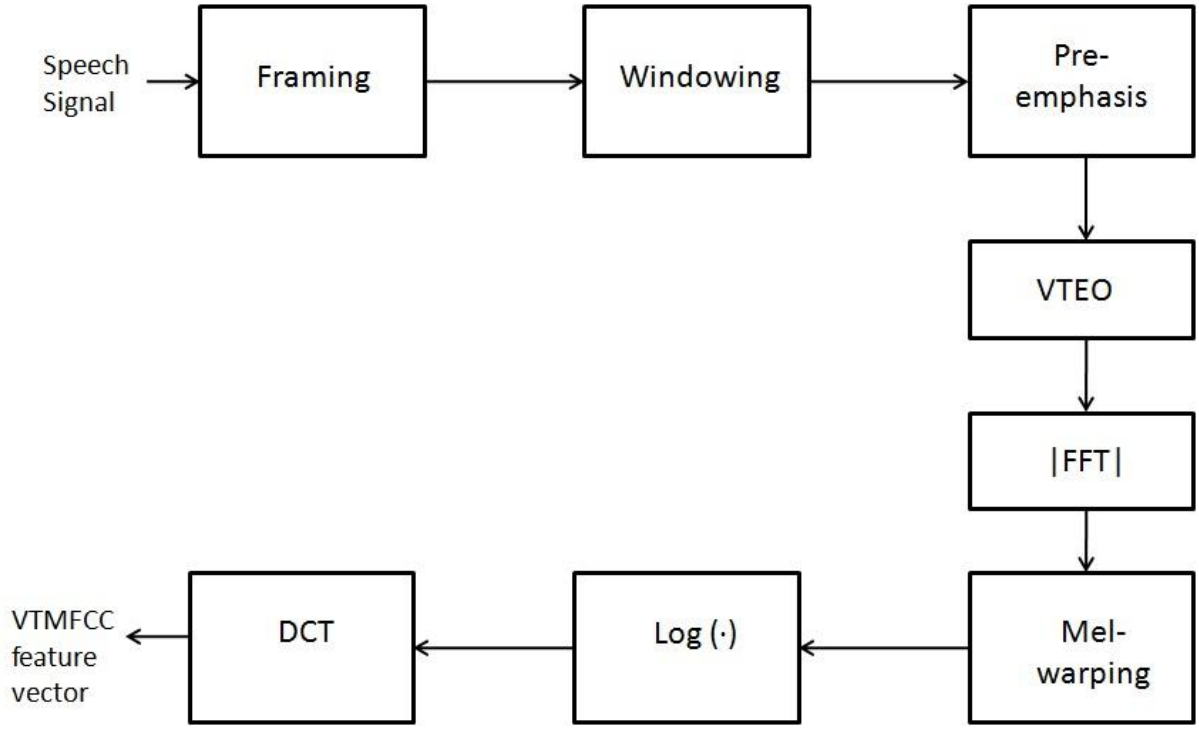


Figure 2.3 VTMFCC feature extraction scheme [17]

Real-time estimation of this required energy for signal generation can be done by TEO [28]. Only three consecutive samples are required to estimate the signal energy. In addition, TEO has unique noise suppression characteristics [28]. However, when the amplitude difference between two consecutive samples is very small, TEO gives zero output which may not be true in physical sense. To address this problem, VTEO (Variable length Teager Energy Operator) was proposed in [29]. VTEO is a generalization of concept of TEO. In general, VTEO with dependency index i (*i.e.*, DI= i) is expected to give the running estimate of energy by considering present, past i^{th} and future i^{th} samples.

First of all, the speech signal is pre-processed as shown in Section 2.4. The VTEO is then calculated on this pre-processed frame. Then, FFT is taken to calculate magnitude spectrum. Mel warping is performed on this spectrum followed by log and DCT computation to get VTMFCC feature vector.

2.6 Modulation Features

Cepstral features are the most successful features used in automatic speech and speaker recognition systems. These features are computed from magnitude spectrum of each frame leaving the phase information. The magnitude spectrum employed in features like MFCC and LPCC (Linear Prediction Cepstral Coefficients) is highly sensitive to changes in speaking conditions, such as speaking style and channel variations. This approach is accepted because it is believed that phase carries little information about speech content and speaker's identity. However, it has been proved that phase information is important in human perception of speech [30]. The cepstral features do not capture changes within the duration of single frame. The delta cepstral features capture spectral changes between the frames, but do not capture changes within the frame.

One approach to overcome this shortcoming is to use the features that characterize intra-frame spectral variation. One such feature is based on AM-FM model of speech. The AM-FM modeling technique has been applied to the speech signal in the area such as formant tracking, speech recognition and speaker recognition. It was first proposed by Potamianos *et. al*. for format tracking [31]. Macro Grimaldi *et. al.* indicated that characterization of different instantaneous frequency plays an important role in capturing the identity of speaker [32].

Speech signal is passed through the cochlear filter (band-pass filter of human ear) in human auditory system. Slow varying AM and FM component can be extracted from the output of this cochlear filter. A speech signal is modeled as real signal that has instantaneous amplitude $A(t)$ and instantaneous phase $\phi(t)$, *i.e.*,

$$x(t) = A(t)\cos[\phi(t)], \tag{2.6}$$

where the AM is represented by the magnitude $A(t)$ and is known as envelop of $x(t)$. The FM is represented by the derivative of phase $\phi(t)$.

- **AM-FM Model**

AM-FM model technique is widely used in the context of radio transmission. It can be effectively used for modeling of speech production system. Vocal tract resonances can change rapidly both in frequency and in amplitude even within a single pitch period. This is due to rapidly varying and separated airflow in the vocal tract [33]. The effective air masses in vocal tract cavities and effective cross-sectional areas of the airflow vary rapidly, causing modulation of air pressure and volume velocity. Thus, actual speech signal $s(t)$, can be thought of a sum of *n* resonances as

$$s(t) = \sum_{j=1}^{N} R_j(t), \tag{2.7}$$

where $R(t)$ is a single speech resonance which can be represented as an AM-FM signal,

$$R(t) = a(t)\cos\left[2\pi\left(f_c t + \int_0^t q(\tau)d\tau\right) + \theta\right], \tag{2.8}$$

where f_c is the centre value of resonance frequency, $q(t)$ is frequency modulating signal and $a(t)$ is amplitude modulating signal. The instantaneous frequency of signal is defined as,

$$f_i(t) = f_c + q(t). \tag{2.9}$$

There are two popular methods for estimating instantaneous amplitude and frequency, *viz.*, Energy Separation Algorithm (ESA) and Hilbert Transform Demodulation algorithm (HTD). In the present work, we have used Hilbert Transform Demodulation (HTD) algorithm to calculate instantaneous amplitude and instantaneous frequency, which is discussed below.

- **Hilbert Transform Demodulation (HTD)** [33]

In order to find instantaneous frequency for a real-valued signal, an analytic signal is first constructed. For a given signal $s(t)$, the analytic signal $s_a(t)$ can be computed as,

$$s_a(t) = s(t) + j\hat{s}(t), \tag{2.10}$$

where $\hat{s}(t)$ is the Hilbert transform of $s(t)$. Hilbert transform operation shifts the phase of all frequency components of $s(t)$ by 90°. We can write $s_a(t)$ as follows

$$s_a(t) = |s_a(t)|e^{j\phi(t)}, \tag{2.11}$$

where $a_i(t) = |s_a(t)|$ is called instantaneous amplitude of the signal, and

$$\phi(t) = \angle s_a(t) = tan^{-1}\frac{\hat{s}(t)}{s(t)}, \tag{2.12}$$

is called *instantaneous phase*. The phase is in time-domain. Instantaneous frequency $f_i(t)$ is derived from unwrapped instantaneous phase as follow

$$f_i(t) = \frac{1}{2\pi}\frac{d\phi_u(t)}{dt}, \tag{2.13}$$

where ϕ_u is unwrapped phase. The instantaneous frequency estimation is one of the effective methods to detect and track frequency changes of mono component signal. Speech signal is to be separated in its components using proper bandpass filtering. Therefore, speech signal is passed through filterbank and $a_i(t)$ and $f_i(t)$ is calculated for each filter output.

- **Short time frequency estimation**

After the estimation of instantaneous frequency, the short-time frequency is estimated. The estimate is weighted moment of $f_i(t)$ using squared amplitude $a_i^2(t)$ as its weight. The weighted mean provide more accurate frequency estimate and it is more robust for low energy and noisy band [31], *i.e.*,

$$F_{iw} = \frac{\int_{t_0}^{t_0+\tau}[f_i(t)a_i^2(t)]dt}{\int_{t_0}^{t_0+\tau}[a_i^2(t)]dt}. \tag{2.14}$$

2.6.1 AM-FM Feature [34]

First, the speech signal is pre-processed as shown in Section 2.4. The pre-processed frame is then applied to the Gabor filterbank. The Gabor filterbank is similar to the triangular filterbank used in the MFCC. The only difference is, instead of using triangular filter, Gabor filter is used in Gabor filterbank. Gabor filters are *smooth* in both time-domain and frequency-domain, therefore Gabor filters are chosen here. The Gabor filterbank consist of non-uniformly spaced filters, whose bandwidth varies according to Mel scale. Here, more importance is given to the low frequency region than high frequency region (as in MFCC).

The output of each Gabor filter is then applied to Hilbert Transform Demodulation (HTD) algorithm to get instantaneous frequency and instantaneous amplitude. The mean amplitude weighted instantaneous frequency estimate is obtained from instantaneous frequency and instantaneous amplitude. We get vector of length equal to number of filters in filterbank per frame. Finally, DCT is applied to get the feature vector. The block diagram for AM-FM feature extraction is shown in Figure 2.4.

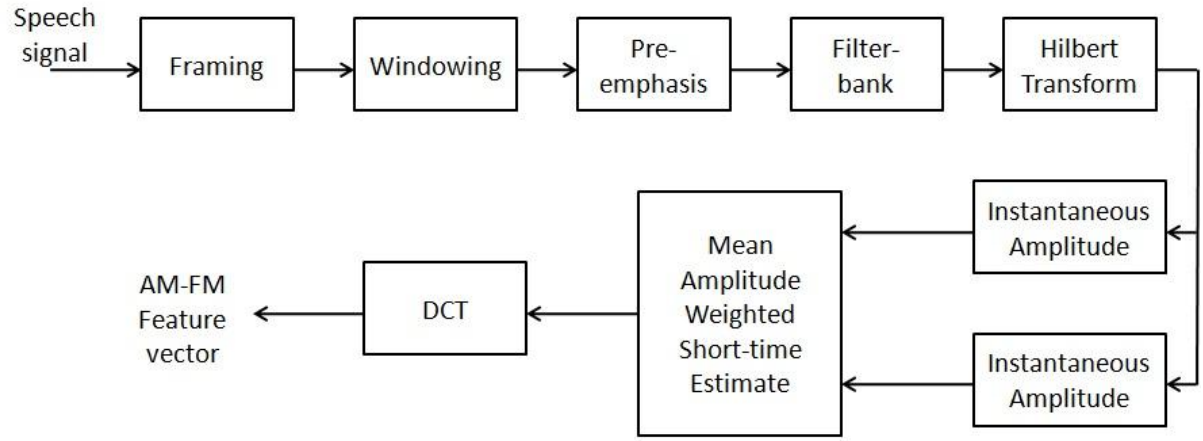


Figure 2.4 AM-FM feature extraction technique [34]

2.6.2 Q-feature [34]

The feature is based on quality factor 'Q' of each of the resonating filters in the filterbank. It is found that obtaining feature vector in terms of value 'Q' capture speaker-specific information. 'Q' of resonator is defined as the ratio of resonating frequency to its bandwidth. We have calculated the mean-amplitude weighted instantaneous frequency estimate in AM-FM feature in above section. Now, we need to calculate the bandwidth estimate. The mean-amplitude weighted instantaneous bandwidth is calculated as per the following formula:

$$[B_{iw}]^2 = \frac{\int_{t_0}^{t_0+\tau}[\left(\frac{\dot{a}_i(t)}{2\pi}\right)^2 + \left(f_i(t) - F_{iw}(t)\right)^2 \cdot a_i^2(t)\,]dt)}{\int_{t_o}^{t_0+\tau}[a_i^2(t)]dt}, \quad (2.15)$$

where the additional term $\left(\frac{\dot{a}_i(t)}{2\pi}\right)^2$ in $[B_{iw}]^2$ accounts for amplitude modulation contribution of the bandwidth [31] , $a_i(t)$ is the instantaneous amplitude, B_{iw} is the mean-amplitude weighted instantaneous bandwidth and F_{iw} is the mean-amplitude weighted instantaneous frequency. The ratio of F_{iw} and B_{iw} is called 'Q' factor of the filter of the filterbank, *i.e.*,

$$Q_i = \frac{F_{iw}}{B_{iw}}. \tag{2.16}$$

After obtaining the value of Q for each filter, DCT is taken to obtain Q-feature vector. The block diagram for AM-FM feature extraction is shown in Figure 2.5.

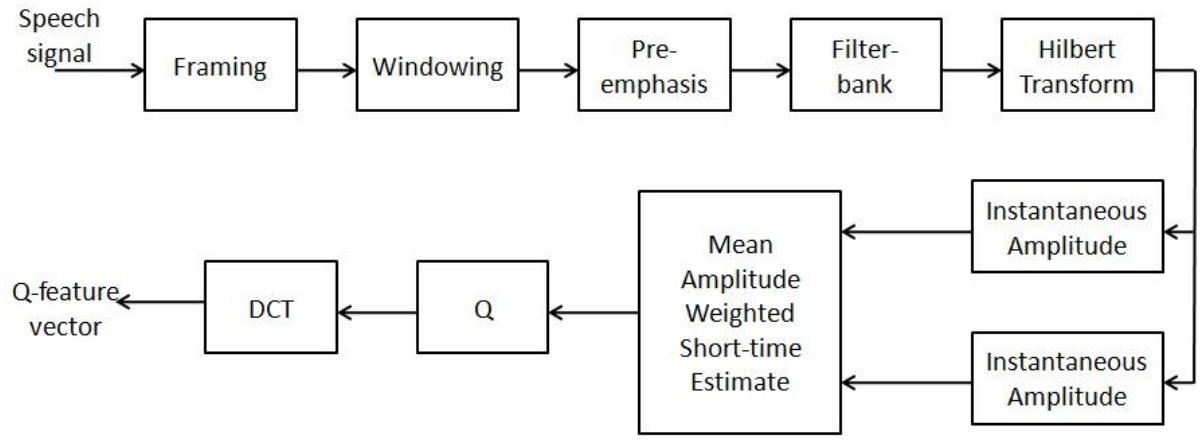


Figure 2.5 Q-feature extraction technique [34]

2.7 Classification

The classification process consists of two steps, *viz.*, modeling and matching. In modeling process, person is enrolled by creating a model for identification system. Matching process is the process of generating scores through which we measure similarity between model and test segment.

There are many popular state-of-the-art modeling techniques available like vector quantization (VQ), Gaussian Mixture Model (GMM), Hidden Markov Model (HMM), etc. In this thesis, polynomial classifier technique is used for classification task.

2.7.1 Polynomial Classifier

Polynomial classifier was first proposed by Campbell *et. al.* [35]. It is an approximation to the Bayes classifier. Polynomial classifier uses out of the class data through discriminative-training to optimize classifier performance. The basic structure of single speaker polynomial classifier is shown in Figure 2.6. Here $\boldsymbol{P}(\boldsymbol{X})$ is the polynomial expansion of test feature vector and $\boldsymbol{W}_{\boldsymbol{spk}}^{\boldsymbol{t}}$ is the speaker model. For example, the polynomial expansion of a two-dimensional feature vector $\boldsymbol{X} = [x_1\ x_2]^t$ and degree two is given by,

$$\boldsymbol{P}(\boldsymbol{X}) = [1 \quad x_1 \quad x_2 \quad x_1^2 \quad x_1x_2 \quad x_2^2]^t \tag{2.17}$$

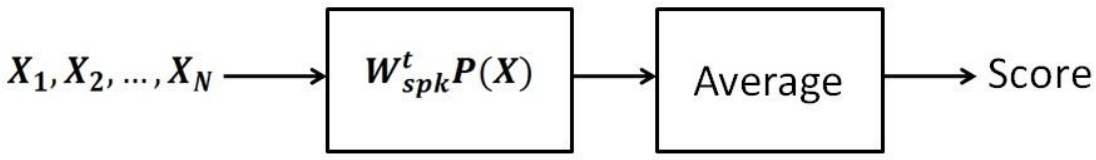


Figure 2.6 Polynomial Classifier

For each input feature vector x_i, the score is produced by taking the inner product of $\boldsymbol{W}_{\boldsymbol{spk}}^{\boldsymbol{t}}$ and $\boldsymbol{P}(\boldsymbol{X}_{\boldsymbol{i}})$. The final score is the average along the time axis and is given by,

$$s = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{W_{spk}^{t}} \cdot \boldsymbol{P}(\boldsymbol{X_i}). \tag{2.18}$$

As discussed in Section 1.3.1, for verification purpose, the score is compared with the threshold and decision is made to accept or reject. For identification task, operation described in Figure 2.6 is performed for all speaker models and the highest scoring speaker is selected. It is clear that scoring operation is nothing but inner product operation between $\boldsymbol{P}(\boldsymbol{X})$ and $\boldsymbol{W_{spk}}$. Polynomial expansion of feature vectors maps the features to the high-dimensional space. The feature occupancy reduces in the high dimensional feature space. Therefore, the distance computation in high dimensional space yields better results.

For good separation between the classes, discriminative-training with mean-squared error (MSE) criteria is used. For speaker's test data output of '1' is expected, on the other hand, output of '0' is expected for imposter data. For simplicity, consider a verification problem. Let $\boldsymbol{W_{spk}}$ be the optimum speaker model, w the class label, and $y(w)$ the ideal output, *i.e.*, $y(spk) = 1$ and $y(imp)$=0. The problem is to find $\boldsymbol{W_{spk}}$ such that it satisfy mean-squared criterion, *i.e.*,

$$\boldsymbol{W_{spk}} = \underset{\mathrm{w}}{\mathrm{argmi\ n}}\ \boldsymbol{E}\{\left(\boldsymbol{W^{t}P}(\boldsymbol{X}) - y(w)\right)^{\boldsymbol{2}}\}, \tag{2.19}$$

where $\boldsymbol{E}$ denotes the expectation operator over x and w. The above equation can be approximated as,

$$\boldsymbol{W_{spk}} = \underset{\mathrm{w}}{\mathrm{argmin}}\ \left[\sum_{i=1}^{N_{spk}} |\boldsymbol{W^{t}P}(\boldsymbol{X_i}) - 1|^2 + \sum_{i=1}^{N_{imp}} |\boldsymbol{W^{t}P}(\boldsymbol{Y_i})|^2\right]. \tag{2.20}$$

Here, $\boldsymbol{X_1, X_2, \ldots, X_{N_{spk}}}$ are speaker's training data and $\boldsymbol{Y_1, Y_2, \ldots, Y_{N_{imp}}}$ are the imposter's training data. The training algorithm can also be written in the matrix form. Let,

$$\boldsymbol{M_{spk}} = \begin{bmatrix} \boldsymbol{P(X_1)^t} \\ \boldsymbol{P(X_2)^t} \\ \vdots \\ \boldsymbol{P\left(X_{N_{\text{spk}}}\right)^t} \end{bmatrix} \tag{2.21}$$

Also, let the similar matrix for the imposter $\boldsymbol{M_{imp}}$. Define,

$$\boldsymbol{M} = \begin{bmatrix} \boldsymbol{M_{spk}} \\ \boldsymbol{M_{imp}} \end{bmatrix} \tag{2.22}$$

The optimization problem becomes the linear approximation problem,

$$\boldsymbol{W_{spk}} = \underset{\mathbf{w}}{\mathbf{argmin}} \, \|\mathbf{MW} - \mathbf{O}\|_2, \tag{2.23}$$

where $\mathbf{O}$ consist of N_{spk} ones followed by N_imp zeros. Solving the above equation, we get,

$$\boldsymbol{W_{spk}} = (\boldsymbol{M^t M})^{-1} \boldsymbol{M^t O}. \tag{2.24}$$

Other details of training algorithm are given in Campbell *et. al.* [35].

2.8 Performance Evaluation

The final stage of any speaker recognition system design is its performance assessment. In this thesis, % identification rate and % Equal Error Rate (EER) are used as performance assessment criterions.

2.8.1 % Identification Rate

The % Identification rate is defined as follow:

$$\%\, ID = \frac{Number\ \ of\ correctly\ \ identified\ \ persons}{Total\ \ number\ \ of\ persons} \times 100 \tag{2.25}$$

2.8.2 % EER rate

Two types of errors occur in verification tasks. They are as follows.

- False Alarm / False Acceptance (FA) Rate. It means impostor is identified as true (genuine person).
- Miss detection/False Rejection (FR) Rate. It means true person is not detected.

By varying the operating point (threshold), different FA and FR rates can be attained. To obtain % Equal Error Rate (EER), threshold is selected such that FR and FA rates become equal.

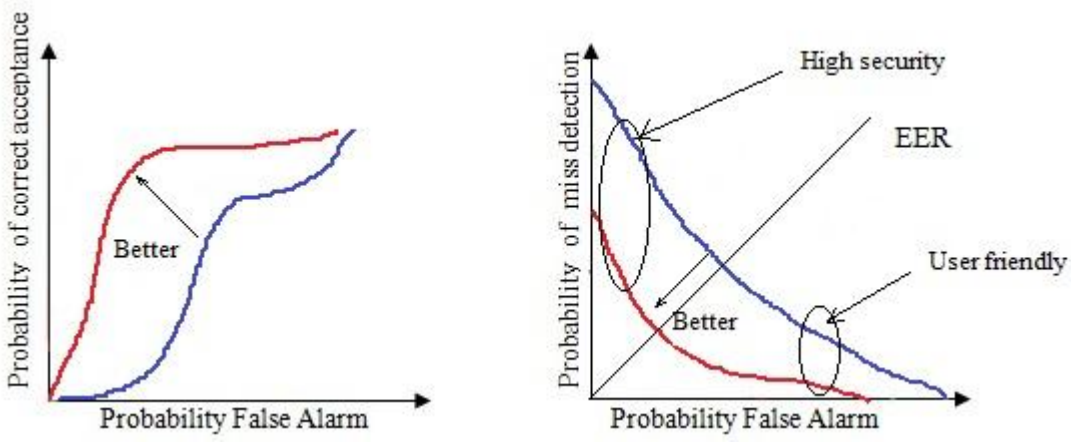


Figure 2.7 Comparison between ROC and DET curves (a) Receiver Operating Characteristic (ROC) curve and (b) Detection Error Trade-off (DET) curve [10]

Receiver Operating Characteristics (ROC) plots correct acceptance probability against false acceptance probabilities as shown in Figure 2.7 (a) [36]. Performance improves as the curve shifts towards top left corner. ROC curve gives us the information about security of the system. It does not provide the information about user-friendliness of the system as it does not consider the miss probability.

A Detection Error Trade-off (DET) curves plots probability of miss detection against the probability of false acceptance as shown in Figure 2.7 (b). There is a trade-off between miss detection and false acceptance. Reducing one type of error will increases the other type of error. When these two errors become equal, it is known as % EER.

2.9 Chapter Summary

In this chapter, different levels of information of speech signal are discussed. The criterions for selection of features are explained. The preprocessing steps before feature extraction are discussed. The feature extraction schemes for MFCC, VTMFCC, AM-FM and Q-features are described. The polynomial classifier and performance metrics of ASR systems are explained. In the next chapter, corpus design procedure will be described.

3 Corpus Design

Many standard corpora are available for the study of speaker recognition systems. However, as the speaker recognition from hum is a recent application, no standard corpus is available for this study. Therefore, corpus is being designed at DA-IICT for last few years. Mr. Prakhar and Mr. Robin built corpus of 53 speakers. Mr. Krishna designed a corpus of another 30 speakers. Mr. Maulik collected samples of additional 97 speakers. However, all these above corpus were designed for single session study. Hence, no inter-session corpus was available for this study. The author of this thesis collected the samples of 50 speakers in 2 training and 6 testing sessions.

3.1 Ideal Characteristics of Corpus

The following issues should be taken care while designing corpus in order to make the corpus ideal.

- The samples should be collected from enough number of subjects in order to make the experiment *statistically* meaningful.
- The duration should be considerably long in order to get reliable speaker-specific information from it.
- Subjects should be selected from wide range of ages so that the study becomes age unbiased.
- Corpus should be gender unbiased, *i.e.*, equal number of male and female subjects should be there in the corpus.
- Samples should be collected in practical environment.
- Samples should be collected in multiple sessions.

– Data should be collected from different types of microphones and acoustic channels.

Figure 3.1 Author (right) collecting humming sound from a subject (left)

3.2 Data Collection

15 popular Hindi movie songs are chosen for this corpus design. Subjects were asked to hum for these songs. 10 songs are used for training purpose and 5 songs are used for testing purpose. The humming duration for training song is of 2 minutes and testing song is of 1 minute. The corpus is *text-independent* as the songs used in testing are different than the songs used in training. The training is done in two sessions and testing is done in 6 sessions. Training session 1 and testing session 1 is recorded on the same session. The training session 2 and testing session 6 is recorded on the same session (*i.e.*, training session 2 and testing session 6 are intra-session data). The data is recorded over the period of 1 month. The first four testing session are recorded over the gap of one to two days. The fifth testing session is

recorded over the gap of seven to ten days and the sixth session is recorded over the gap of fifteen to twenty days. The author collecting the humming sample from a subject is shown in Figure 3.1. The description about corpus design procedure is given in Table 3.1

Table 3.1 Details of humming corpus design

Item	Details
Number of persons (subjects)	50 (49 male and 1 female)
Number of training sessions	2
Number of testing sessions	6
Text type	Humming for 15 popular song
Sampling rate	22050 Hz
Sampling format	Mono channel, 16-bit resolution
Microphone	HP dual earphone with SCHURE mic
Acoustic environment	DA-IICT hostel and Lab

The database was collected at DA-IICT hostel rooms and Lab, so babble noise and other noise is there in database. The database was collected from the speakers without any payment. Hence, their mode of excitation is not altered and subjects who were ready to volunteer were considered. Two laptops Acer and Sony Vaio were used for recording with inbuilt noise suppression characteristic. The files are stored in '.wav' format using Audacity 1.3 Beta software (a publically available software).

3.3 Experiences During Data Collection

The task of collecting the data from 50 speakers in 2 training and 6 testing sessions is very difficult and interesting. The corpus involves 2 minutes of humming for training purpose and 1 minute of humming for testing purpose. The studio room of DA-IICT provides excellent environment for recording. The recording was very clean as there was pin drop silent environment and standard recording machines. But, I dropped the idea of collecting data in studio room as it was very difficult to record almost 400 sessions there. Then I started data collection again with laptop. This time I decided to go to the rooms of the subjects instead of calling them at particular location. I got very good success as the things were going very well. The level of co-operation was high compared to last time as the subjects did not have to come at studio room. I collected data from hostel and lab. I completed the task of data collection in around one and half month period.

3.4 Chapter Summary

In this chapter, ideal characteristics of corpus are explained. Corpus design procedure for humming data is described. Next chapter will describes the performance of all four features, *viz.*, MFCC, VTMFCC, AM-FM and Q-feature in single-session and multi-session environment is evaluated.

4 Experimental Results

4.1 Introduction

This chapter discusses the performance evaluation of person recognition system using the features described in Section 2.5 and Section 2.6 for the single session and multi session humming data. % EER and % identification rate is taken as the performance metric. First, we will see how the distribution of speaker varies with the session variations.

4.2 Single Session Experiment

In this section, the performance of the MFCC, VTMFCC, AM-FM and Q-features is evaluated for single session data. In this database, training and testing data is collected on the same session. The performance is evaluated in terms of the % EER and identification rate. The detail about experimental setup is given in Table 4.1.

The results are obtained using polynomial classifier of order 2 and order 3. The DET curves of all the four features are shown in Figure 4.1 .

The '*' and '*o*' in the DET- curve indicates the % Equal Error Rate (% EER), where the % probability of false alarm (FA) and % probability of miss detection are equal. MFCC have lowest EER rate. The evaluation in terms of performance metric % identification rate and % EER is shown in Table 4.2 and Table 4.3. From Figure 4.1, Table 4.2 and Table 4.3, it is obvious that MFCC performs best for single session data amongst the four features discussed.

Table 4.1 Experimental setup for single session experiment

Item	Details
No. of speakers	50 (49 male and 1 female)
Features used	MFCC, VTMFCC, AM-FM, Q-feature
Feature dimension	12-D
Filters in filterbank	24
Analysis window	23.25 *ms* or 512 samples
Classifier	Polynomial classifier order 2
Train segments	30 s, 60 s, 90 s (3 training segments)
Test segments	1 s, 2 s, 3 s, ...,30 s (30 testing segments)
Genuine trials	30 (per test segment) × 3 (per train segment) × 50 (number of speakers)= 4500
Imposter trials	50 ×50×30×3 – 4500=220500

DET curve shifts downward as polynomial classifier of order-3 is used as shown in Figure 4.1. Polynomial classifier of higher order gives better results as the feature occupancy reduces in higher dimensional feature space, however it increases computational complexity considerably [10]. Therefore, further experiments are limited to polynomial classifier of order 2 only.

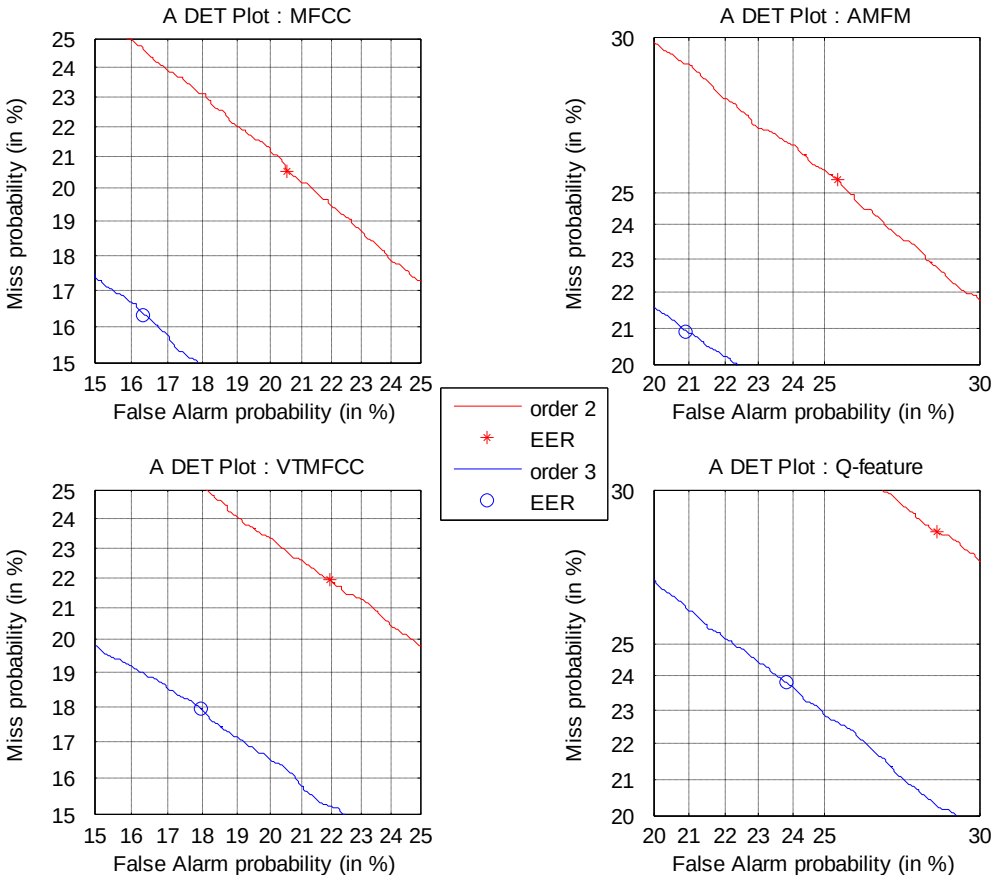


Figure 4.1 DET curves of various features

Table 4.2 % EER and % Identification rate of different feature using polynomial classifier of order 2

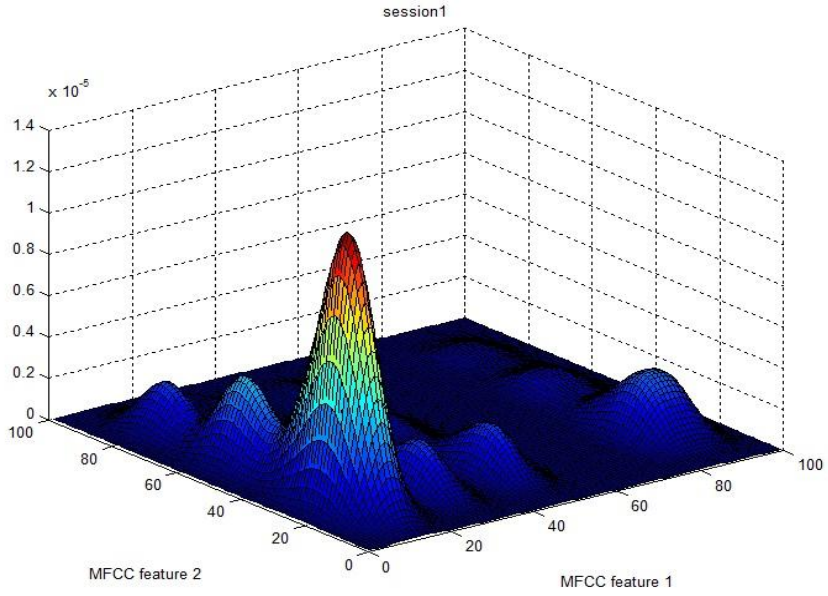
Features	% EER	% ID
MFCC	20.57	59.57
VTMFCC	21.91	56.20
AM-FM	25.40	45.13
Q-feature	28.58	31.42

Table 4.3 % EER and % identification rate of different features using polynomial classifier of order 3

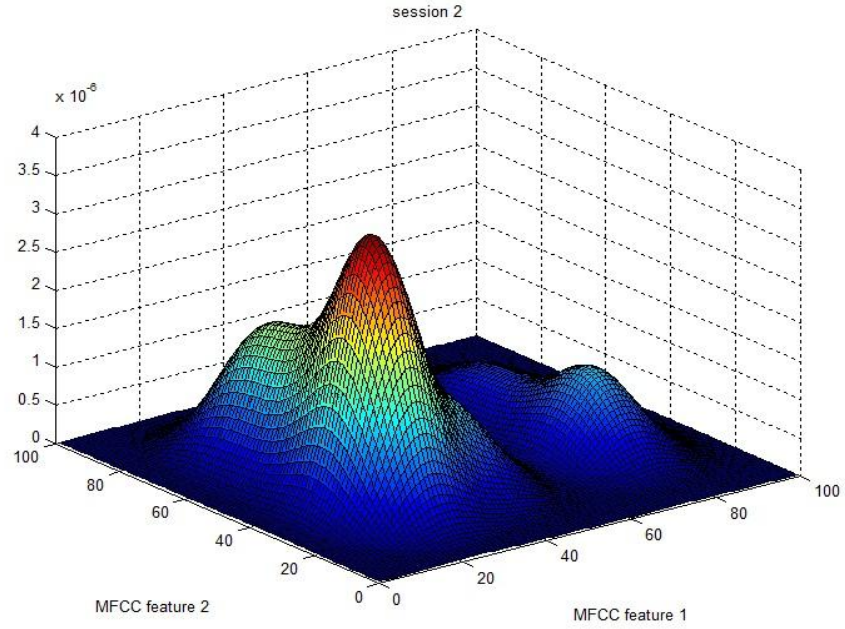
Features	% EER	% ID
MFCC	16.31	69.69
VTMFCC	17.93	67.11
AM-FM	20.91	55.0
Q-feature	23.82	43.33

4.3 Change in distribution of MFCC

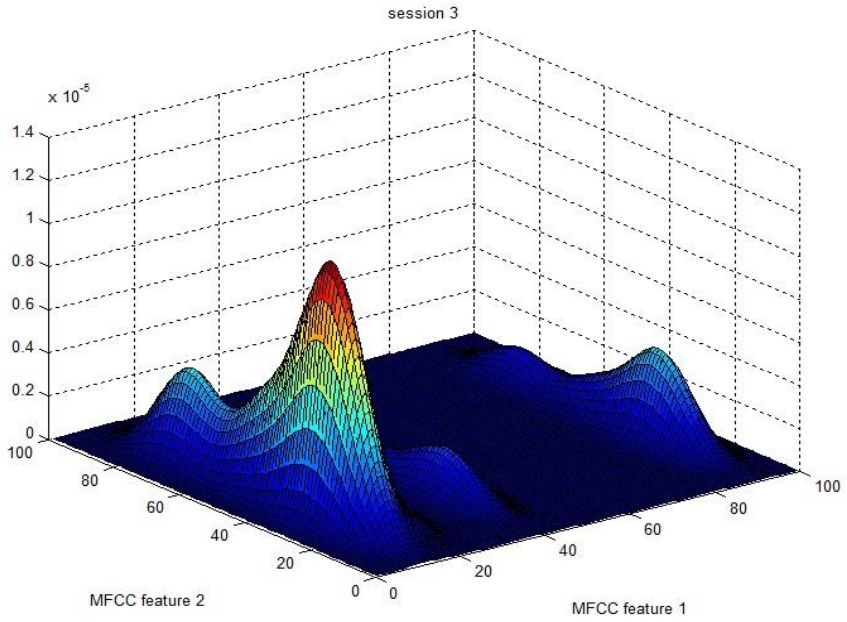
In this section, the joint *pdf* of the first two coefficients of MFCC features have been plotted for different sessions to observe change due to inter-session variability in the same speakers. The joint *pdf* plots have been shown for all six test session of same person. The algorithm and MATLAB code for obtaining joint *pdf* are described in Appendix A.



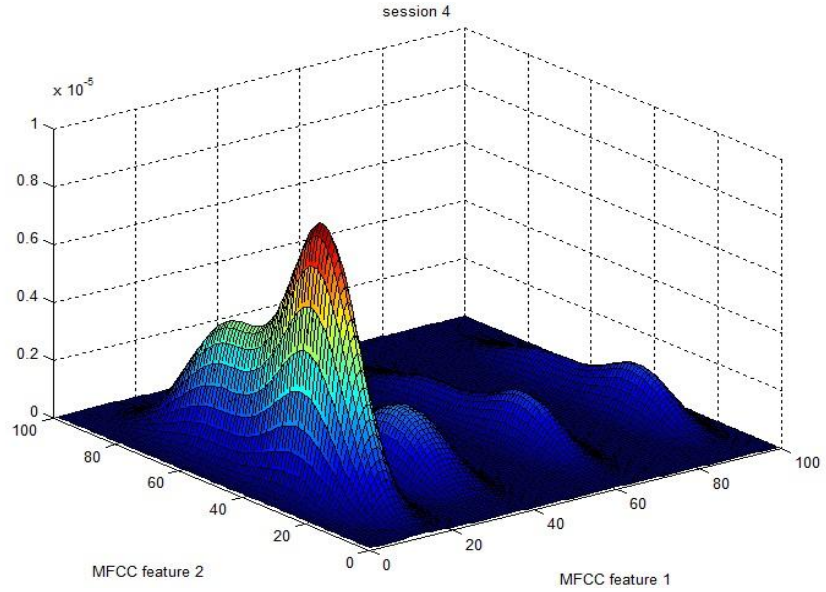
(a) Joint density function for session 1



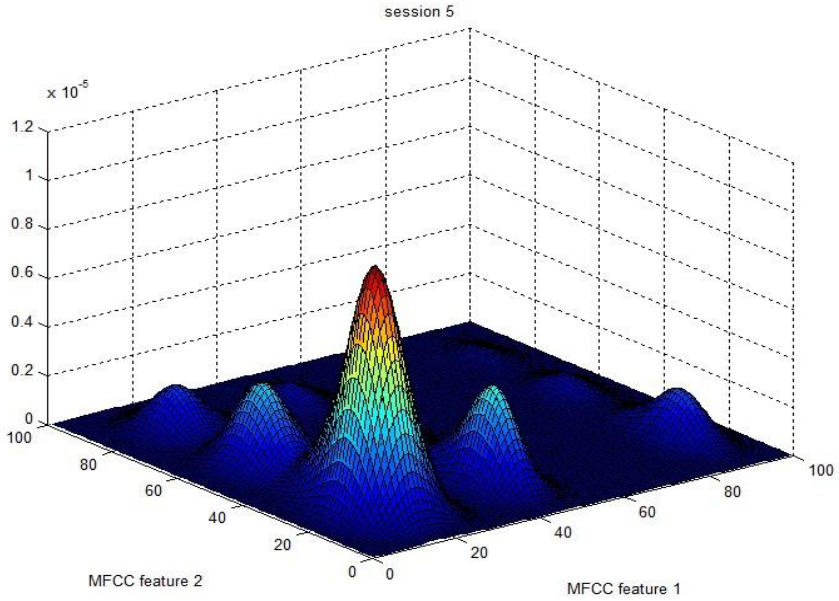
(b) Joint density for session 2



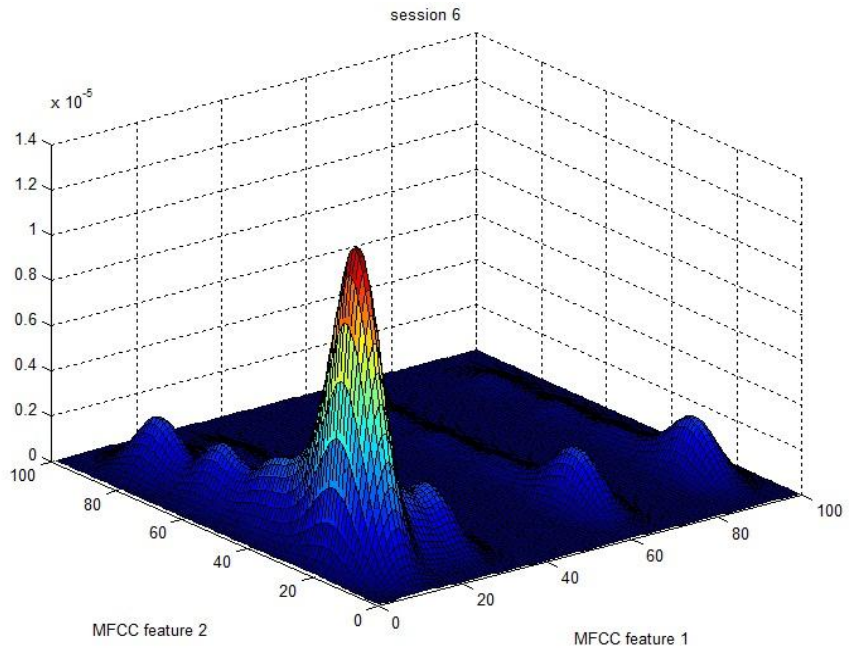
(c) Joint density for session 3



(d) Joint density for session 4



(e) Joint density for session 5



(f) Joint density for session 6

Figure 4.2 Joint density function of first two coefficient of MFCC for 6 sessions

From the above plots, it is clear that the joint *pdf* of all the sessions are different, *i.e.*, considerable change in feature vectors with respect to sessions as we move from session 1 to session 6. These initial results indicate that the person recognition performance is expected to get affected by inter-session variability.

4.4 Performance on Multi-Session Data

The next important task is to see how above mentioned features perform on multi-session data. In other words, robustness of the recognition system against degradation due to inter-session variability is to be evaluated. In this experiment, the training session 1 (the songs recorded on the first day for the training) is used for the training of the system. The six testing sessions are used to test the system against intersession variability. The DET curves for all features are shown in Figure 4.3. It shows that the DET curve line shifts upward for

the sessions different than the first session. The first session has lower % EER as training and testing are done in the same session for that session. It is clear that the inter-session variability degrades the performance of the system.

The difference in the % EER is taken as the measure of the inter-session variability. The difference of the current session to the first session is used to calculate degradation in current session. In other words, increase in the difference of % EER is considered equivalent to increase in degradation.

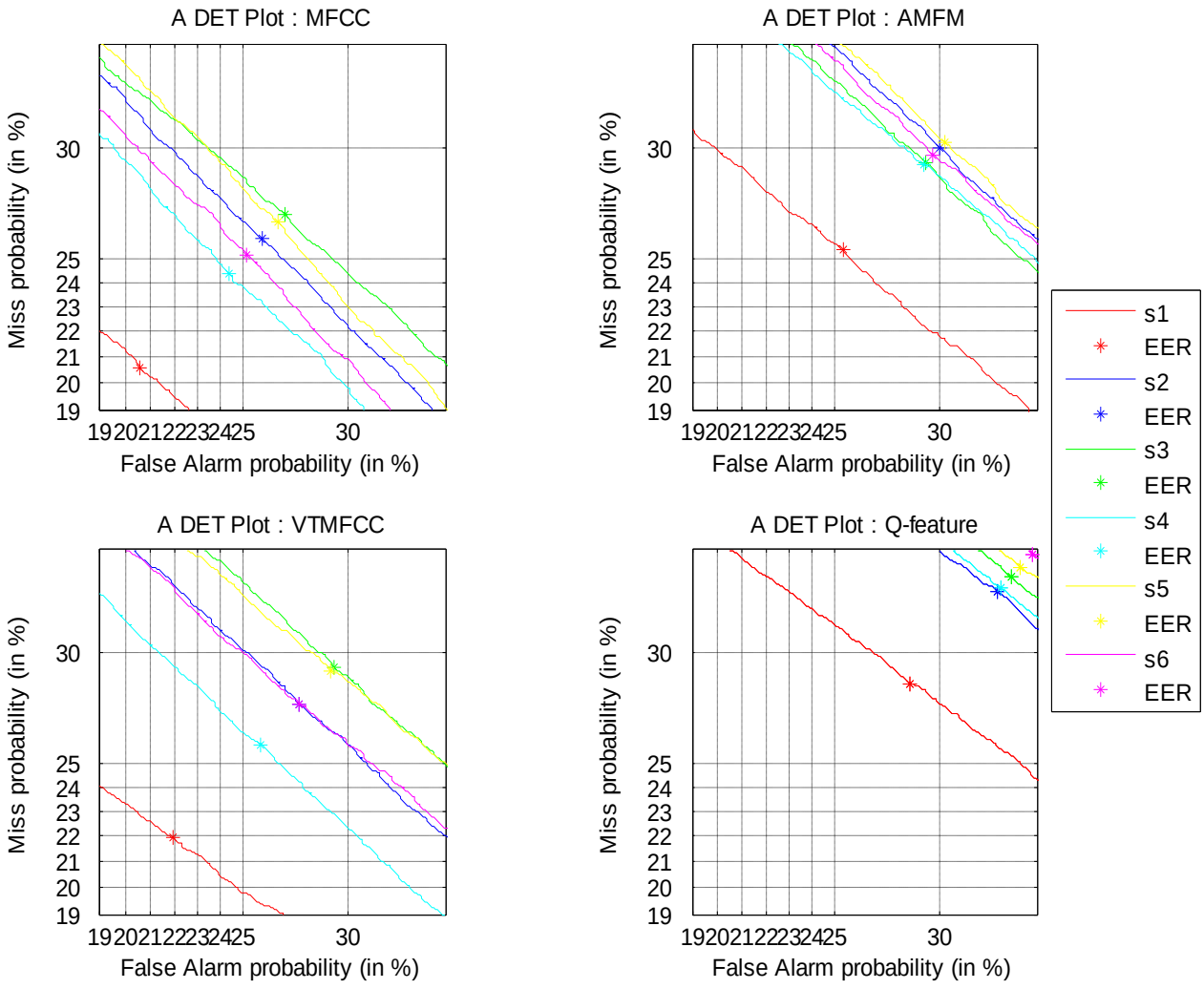


Figure 4.3 DET curves of different feature for multi-session data

The degradation due to inter-session variability is shown in the Figure 4.4. The Figure 4.4 (a) shows the how the % EER of different features varies according to sessions. The Figure 4.4 (b) shows how the % identification rate varies with respect to sessions. It is clear that the % EER increases for the test sessions other than the test session recorded with training session. The % identification rate decreases considerably for the test sessions other than the test session recorded with training session.

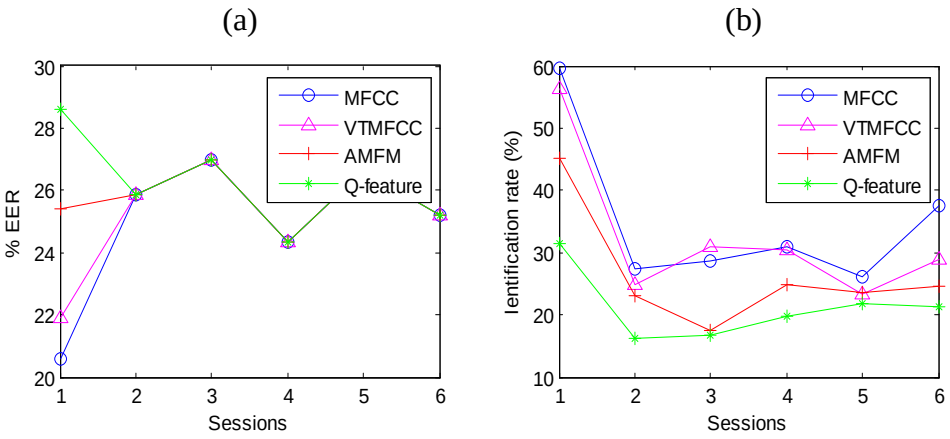


Figure 4.4 degradation due to sessions (a) % EER and (b) % Identification Rate

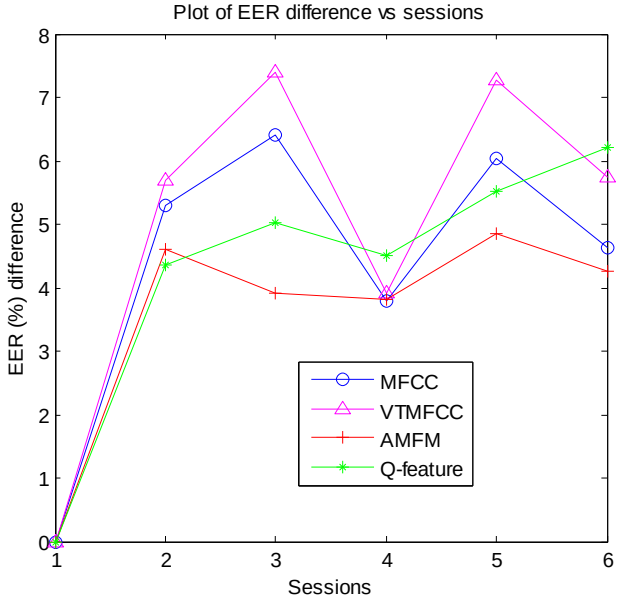


Figure 4.5 Change in %EER *vs.* sessions

As shown in Figure 4.5, the change in % EER due to inter-session variability is highest for the VTMFCC feature set, followed by MFCC. The AM-FM feature set has the lowest change among the four feature sets. It is shown that VTMFCC is more sensitive than MFCC to the inter-session variability, because their base session (testing session recorded with training session) % EER are almost equal.

4.5 Effect of Number of Filters used in Filterbank

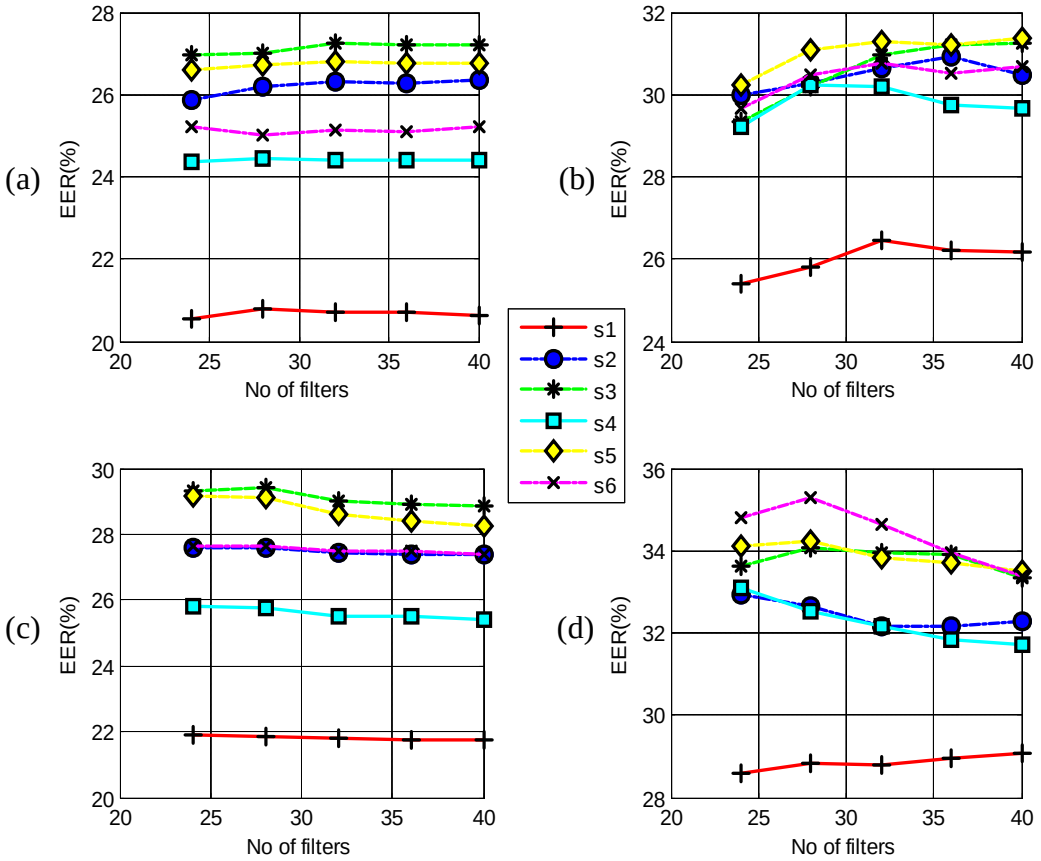


Figure 4.6 Effect of number of filters in filterbank on performance of (a) MFCC, (b) AM-FM, (c) VTMFCC and (d) Q-feature

The experimental setup in the Table 4.1 mentions the number of filters to be 24. In this experiment, the number of filters in filterbanks (in both triangular filterbank used in

MFCC, VTMFCC and Gabor filterbank in AM-FM, Q-feature) is varied to observe its effect on performance of person recognition system.

As shown in Figure 4.6, the number of filters does not have any significant effect on cepstral features and the performance of AM-FM feature slightly degrades with the increase in the number of filters.

4.6 Effect of Feature Dimension

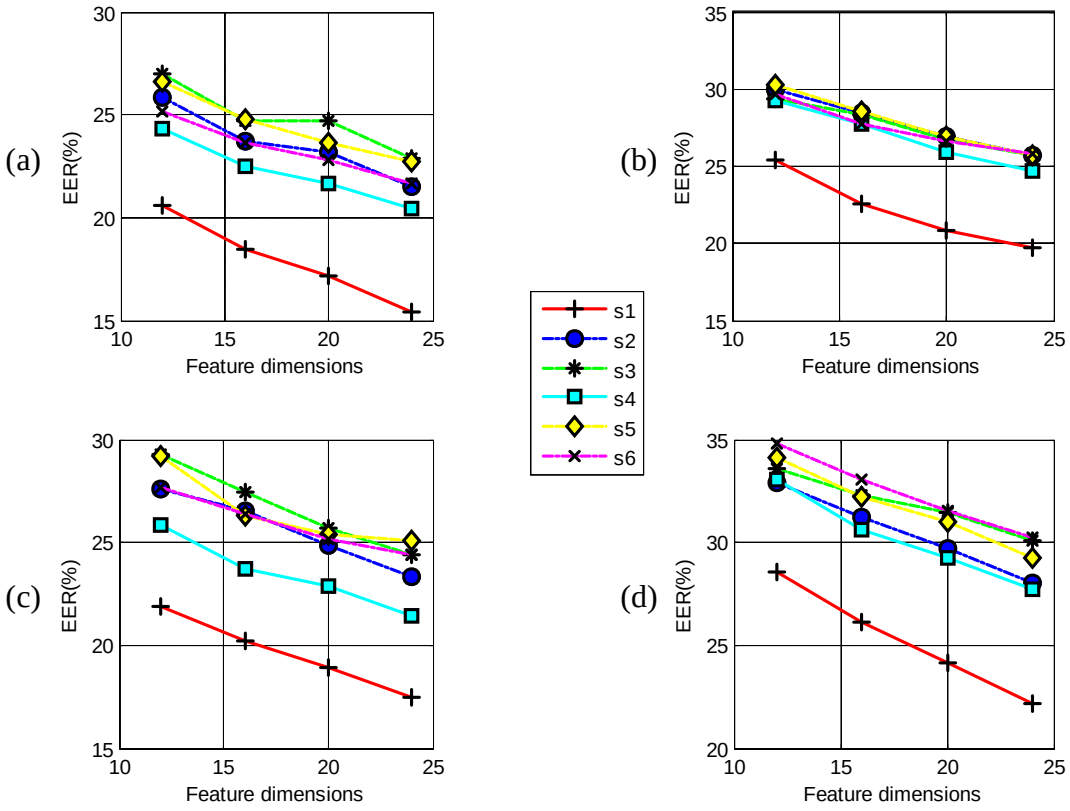


Figure 4.7 Effect of feature dimension on performance of (a) MFCC, (b) AM-FM (c) VTMFCC and (d) Q-feature

The experimental set-up in Table 4.1 shows the feature dimension (number of coefficients in the feature vector) to be *12*. In this experiment, feature dimension is varied to observe its effect on performance. All other parameters are kept as stated in Table 4.1. Figure 4.7 shows that as feature dimension increases, the performance of the system is

improved for all of the four features. This is due to the fact that feature occupancy is reduced in higher-dimensional feature space; therefore they can be better classified by the same classifier. The only problem is that, it increases the complexity of the classification algorithm. Therefore, the number of coefficients in feature vector is kept to be *12* for further experiments, even though the higher length of feature vector yields better performance.

4.7 Effect of length of analysis window

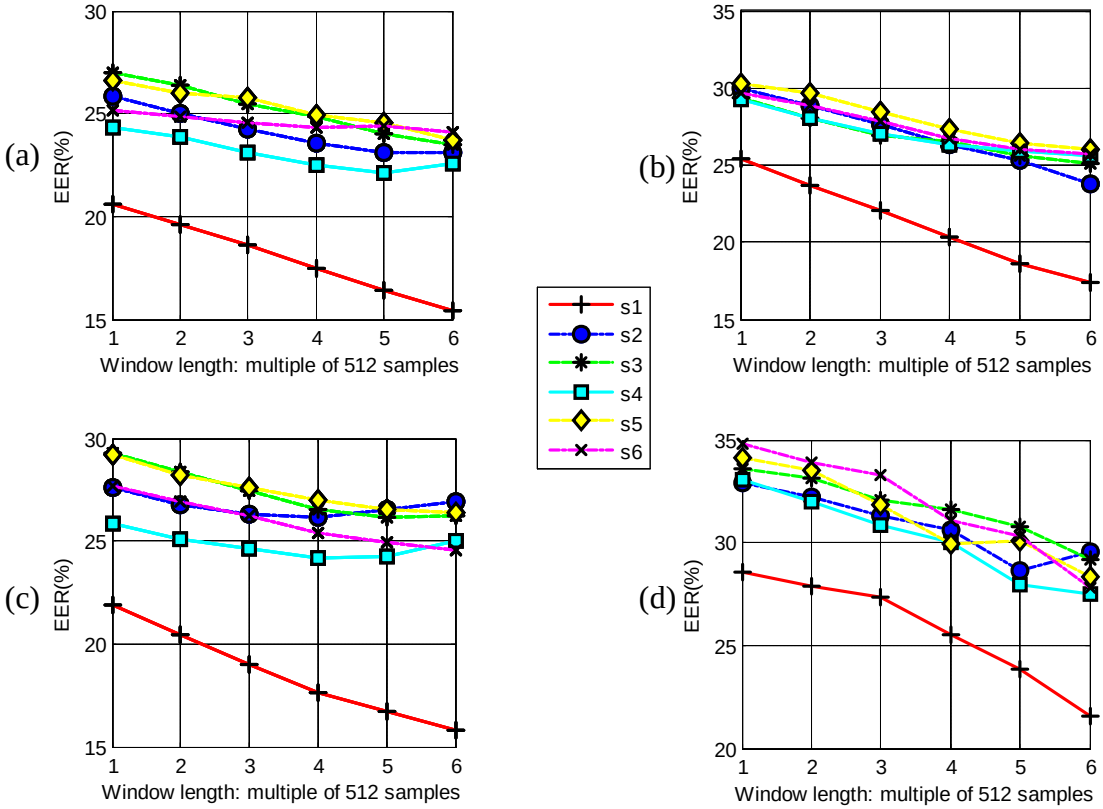


Figure 4.8 Effect of length of analysis window on performance of (a) MFCC, (b) AM-FM, (c) VTMFCC and (d) Q-feature

The experimental set-up in Table 4.1 shows the length of analysis window to be 23.25 *ms* or 512 samples. In this experiment, the length of analysis window is varied to observe its effect on performance. All other parameters are kept as mentioned in Table 4.1.

Figure 4.8 shows the length of analysis windows on the performance. The integer on the X-axis is the multiplier of 512 samples or 23.2 *ms.* The details of this are given in Table 4.4. This is done to avoid spreading of graph on X-axis.

Table 4.4 The relation between multiplier and length of analysis window

multiplier	Number of Samples	Frame duration in Time (*ms*)
1	512	23.2
2	1024	46.4
3	2048	92.9
4	4096	185.8
5	8192	371.5
6	16384	743.0

The performance of the system improves as the length of analysis window increases. The performance of the first session improves almost linearly with the increase of analysis window. The performance of the sessions other than first session improves up to the length of 8192 samples or 371.5 *ms* and then seems to be remaining stable or slightly degrade.

Typically, the length of analysis window in speech signal processing experiment is taken 10 to 30 *ms*, as the speech signal is believed to remain stationary for typically 10 to 30 *ms*. Thus, if the length of analysis window is taken higher than that, the performance typically degrades. Initially by the same assumption, the length of analysis window is taken as approximately 23 *ms* for the experiments on humming signal. However, this experiment shows that the performance of the system improves for the length of analysis window of up to 743 *ms*. This means there should be higher *stationarity* in humming signal than speech signal. This may be due to the fact that hum is nasal sound and nasal cavity remains almost fixed during the production of hum sound. In addition, longer analysis window implies less number

of features for the same length of humming data. The less number of features reduces the computational complexity and the recognition system becomes faster.

4.8 Chapter Summary

In this chapter, the performances of the four features discussed in Chapter 2 are compared. The degradation due to intersession variability is compared. The performance criterion for intersession variability is defined. The performance of system against the variation of number of filters in the filterbank, length of feature vector and length of analysis window is discussed. In next chapter, we propose novel techniques to alleviate degradation due to intersession variability for the person recognition system.

5 Inter-session Variability Compensation

5.1 Introduction

In Chapter 4, the degradation in performance due to intersession variability is discussed. % EER of test sessions other than that of recorded with training session is higher compared to base session. The four features discussed in the Chapter 4 do not provide much immunity to the intersession variability. Therefore, a new technique is required that can provide some robustness in performance of proposed person recognition system.

Table 5.1 Experimental set up

Item	Details
Number of persons	50 (49 male and 1 female)
Features used	MFCC, VTMFCC, AM-FM, Q-feature
Dimension of feature vector	12
Number of filters in filterbank	24
Analysis window duration	185.8 *ms* or 4096 samples
Classifier structure	Polynomial classifier order 2
Train segments duration	30s, 60s, 90s (3 training segments)
Test segments duration	1s, 2s, 3s, ...,30s (30 testing segments)
Number of Genuine trials	30 (per test segment) × 3 (per train segment) × 50 (number of speakers)= 4500
Number of Imposter trials	50 ×50×30×3 – 4500=220500

The two techniques, *viz., feature-level fusion* and *model-level fusion* have been proposed to address this problem. The training session number 6 is used along with training

session number 1 in the experiments of this chapter. The experimental setup is described in Table 5.1.

5.2 Feature-Level Fusion

The person recognition system performs best when the training condition matches with the testing condition as seen in Section 4.4. However, practically training and testing take place in different condition (*i.e.,* sessions). The acoustic environment, noise level, speaker emotions, etc. will be changed, if testing is done in another session. The objective is to train the system with multiple sessions of data so that at least one training condition will match with the testing condition. By this way, some robustness can be added to the system.

In *feature-level fusion* approach, features of two training data are fused. The fusion is simple; average of training feature vectors from two sessions is taken. This averaged feature vector is then used to train the polynomial classifier model.

The plots of feature-level fusion for MFCC, VTMFCC, AM-FM and Q-feature are shown in Figure 5.1, Figure 5.2, Figure 5.3, and Figure 5.4, respectively. The DET curves of these features without fusion are also plotted for comparison. The graphs of difference in % EER *vs.* sessions of these features are also plotted.

The EER of MFCC, VTMFCC, AM-FM and Q-feature with and without fusion are shown for all sessions, in Table 5.2, Table 5.3, Table 5.4, and Table 5.5

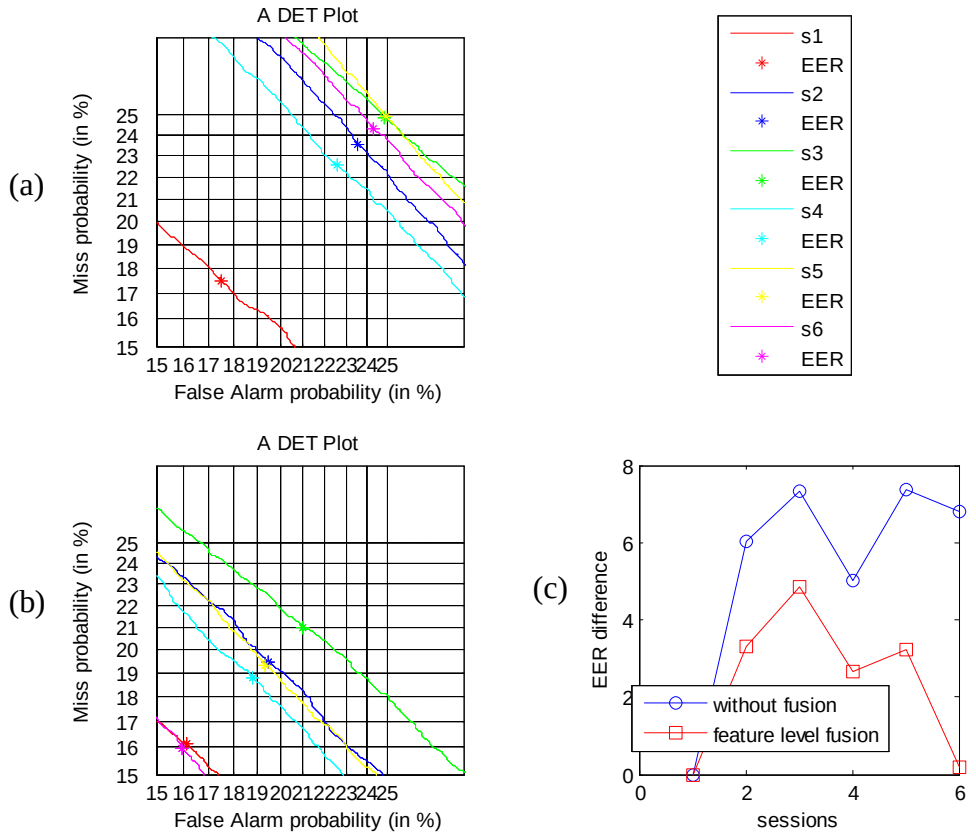


Figure 5.1 Feature-level fusion of MFCC (a) DET plot of MFCC feature without fusion, (b) DET plot of MFCC feature with fusion, and (c) % EER difference *vs.* sessions in MFCC

Table 5.2 Change in % EER for MFCC with feature-level fusion of training sessions 1 and 2

Index of Test Sessions	**MFCC**	**MFCC with feature-level fusion**
S1	17.51	16.14
S2	23.55	19.46
S3	24.84	21.00
S4	22.53	18.81
S5	24.90	19.35
S6	24.30	15.97
Max diff	**7.389**	**4.86**

Training session number 1 is used for the experiment without fusion as shown in 2nd column of the Table 5.2. Training session numbers 1 and 2 is used for the experiment with feature-level fusion as shown in 3rd column of Table 5.2.

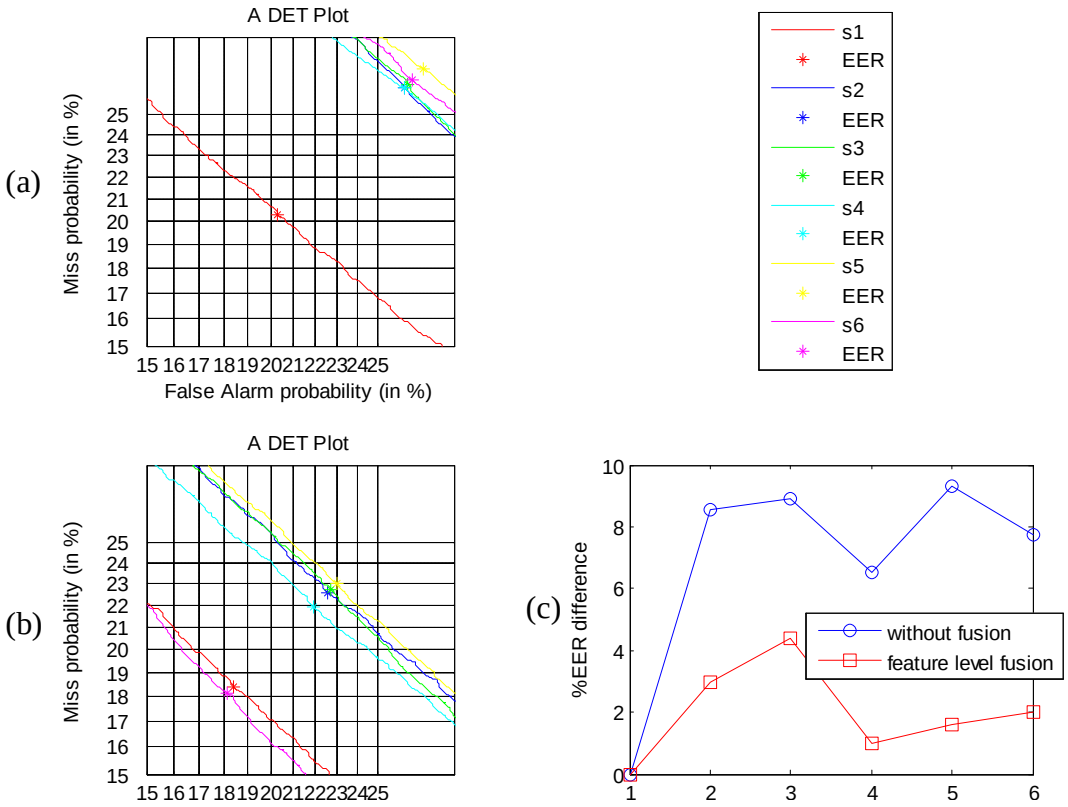


Figure 5.2 Feature-level fusion of VTMFCC (a) DET plot of VTMFCC feature without fusion, (b) DET plot of VTMFCC feature with fusion, and (c) % EER difference *vs.* sessions in VTMFCC

Table 5.3 Change in % EER for VTMFCC with feature-level fusion of training sessions 1 and 2

Index of Test Sessions	**VTMFCC**	**VTMFCC with feature-level fusion**
S1	17.64	17.75
S2	26.18	20.71
S3	26.55	22.16
S4	24.16	18.76
S5	26.95	19.33
S6	25.38	15.75
Max Diff	**9.31**	**4.40**

Training session number 1 is used for the experiment without fusion as shown in 2nd column of the Table 5.3. Training session numbers 1 and 2 is used for the experiment with feature-level fusion as shown in 3rd column of Table 5.3.

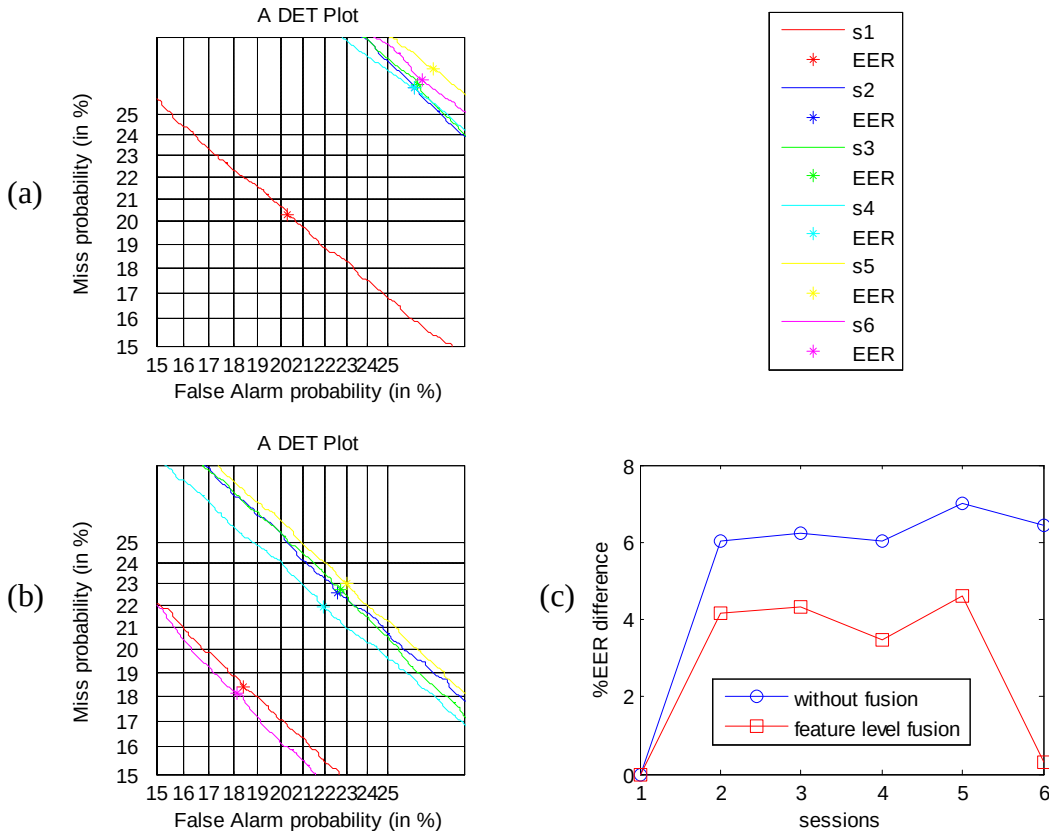


Figure 5.3 Feature-level fusion of AM-FM (a) DET plot of AM-FM feature without fusion, (b) DET plot of AM-FM feature with fusion, and (c) % EER difference *vs.* sessions in AM-FM

Table 5.4 Change in % EER for AM-FM with feature-level fusion of training sessions 1 and 2

Index of Test Sessions	**AM-FM**	**AM-FM with feature-level fusion**
S1	20.30	18.4
S2	26.34	22.56
S3	26.53	22.73
S4	26.31	21.89
S5	27.31	23.0
S6	26.72	18.11
Max diff	**7.01**	**4.6**

Training session number 1 is used for the experiment without fusion as shown in 2nd column of the Table 5.4. Training session numbers 1 and 2 is used for the experiment with feature-level fusion as shown in 3rd column of Table 5.4.

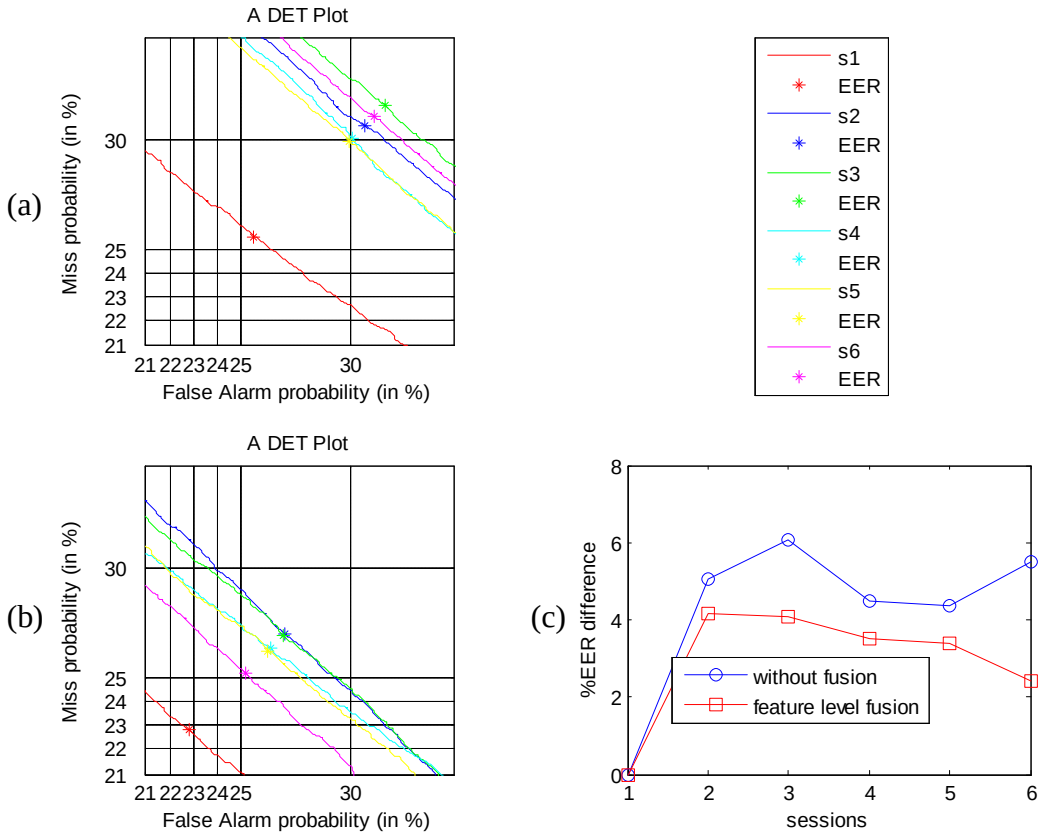


Figure 5.4 Feature-level fusion of Q-feature (a) DET plot of Q-feature without fusion, (b) DET plot of Q-feature with fusion, and (c) % EER difference *vs.* sessions in Q-feature

Table 5.5 Change in %EER for Q-feature with feature-level fusion of training sessions 1 and 2

Index of Test Sessions	**Q-feature**	**Q-feature with feature-level fusion**
S1	25.35	22.76
S2	30.62	26.91
S3	31.64	26.83
S4	30.03	26.26
S5	29.91	26.17
S6	31.07	25.17
Max Diff	**6.08**	**4.14**

Training session number 1 is used for the experiment without fusion as shown in 2nd column of the Table 5.5. Training session numbers 1 and 2 is used for the experiment with feature-level fusion as shown in 3rd column of Table 5.5.

The above plots indicate that feature-level fusion shifts the DET curve downward for all the features. Feature-level fusion techniques, reduces the difference between the % EER of particular session to its base session. Hence, It provides robustness against inter-session variability.

5.3 Model-Level Fusion

In this experiment, the $\boldsymbol{W_{spk}}$ model is created for two training data as described in Section 2.7.1. Then, the average of these two $\boldsymbol{W_{spk}}$ is taken. Let us say this average of two $\boldsymbol{W_{spk}}$ be $\boldsymbol{W_{avg}}$. Now this $\boldsymbol{W_{avg}}$ is considered as our training model for the experiment. The plots of model-level fusion for MFCC, VTMFCC, AM-FM and Q-feature are shown in Figure 5.5, Figure 5.6, Figure 5.7, and Figure 5.8, respectively. The DET curves of these features without fusion are also plotted for comparison. The graphs of difference in % EER *vs.* sessions are also plotted. The % EER of MFCC, VTMFCC, AM-FM and Q-feature with and without model-level fusion are shown for all sessions in Figure 5.5, Figure 5.6, Figure 5.7, and Figure 5.8.

.

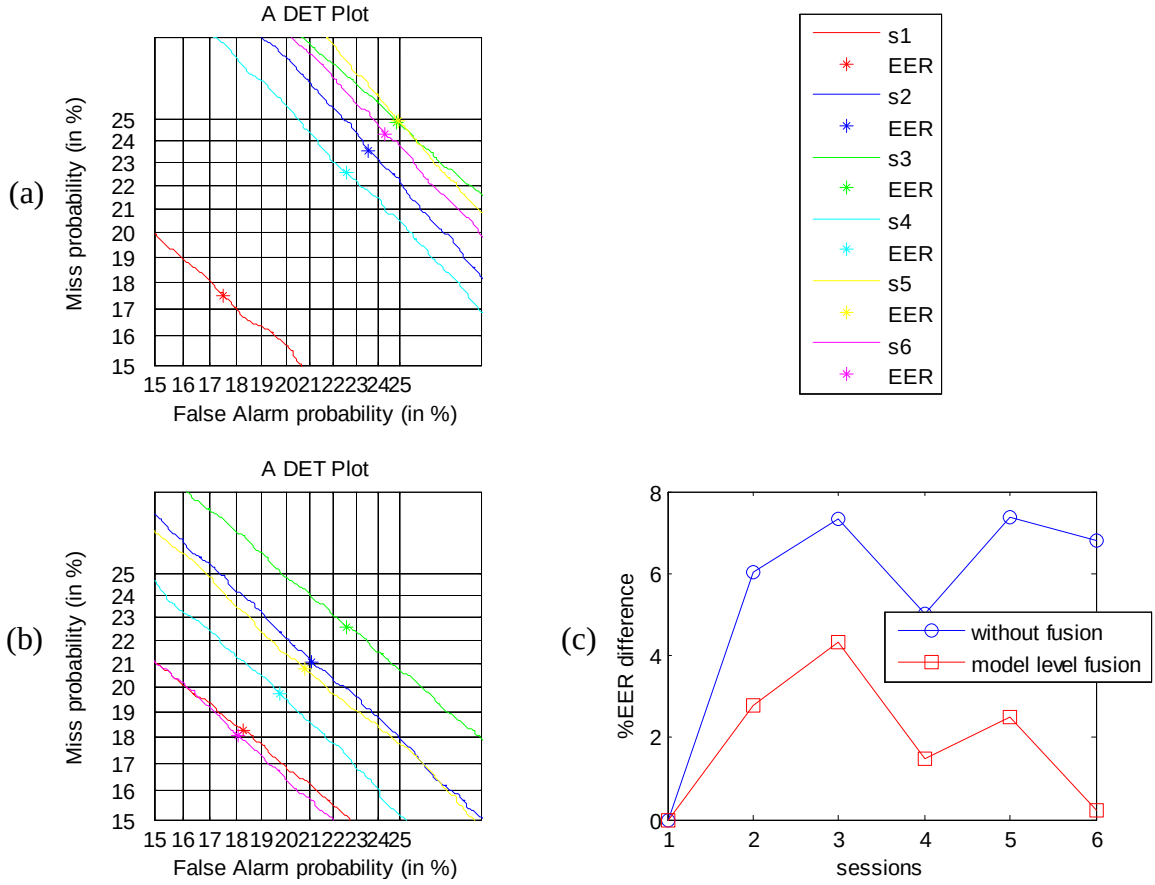


Figure 5.5 Model-level fusion of MFCC (a) DET plot of MFCC feature without fusion, (b) DET plot of MFCC feature with fusion, and (c) % EER difference *vs.* sessions in MFCC

Table 5.6 Change in % EER for MFCC with model-level fusion of training sessions 1 and 2

Index of Test Sessions	**MFCC**	**MFCC with model-level fusion**
S1	17.51	18.26
S2	23.55	21.06
S3	24.84	22.58
S4	22.53	19.75
S5	24.90	20.75
S6	24.30	18.04
Max diff	**7.389**	**4.31**

Training session number 1 is used for the experiment without fusion as shown in 2nd column of the Table 5.6. Training session numbers 1 and 2 is used for the experiment with model-level fusion as shown in 3rd column of Table 5.6.

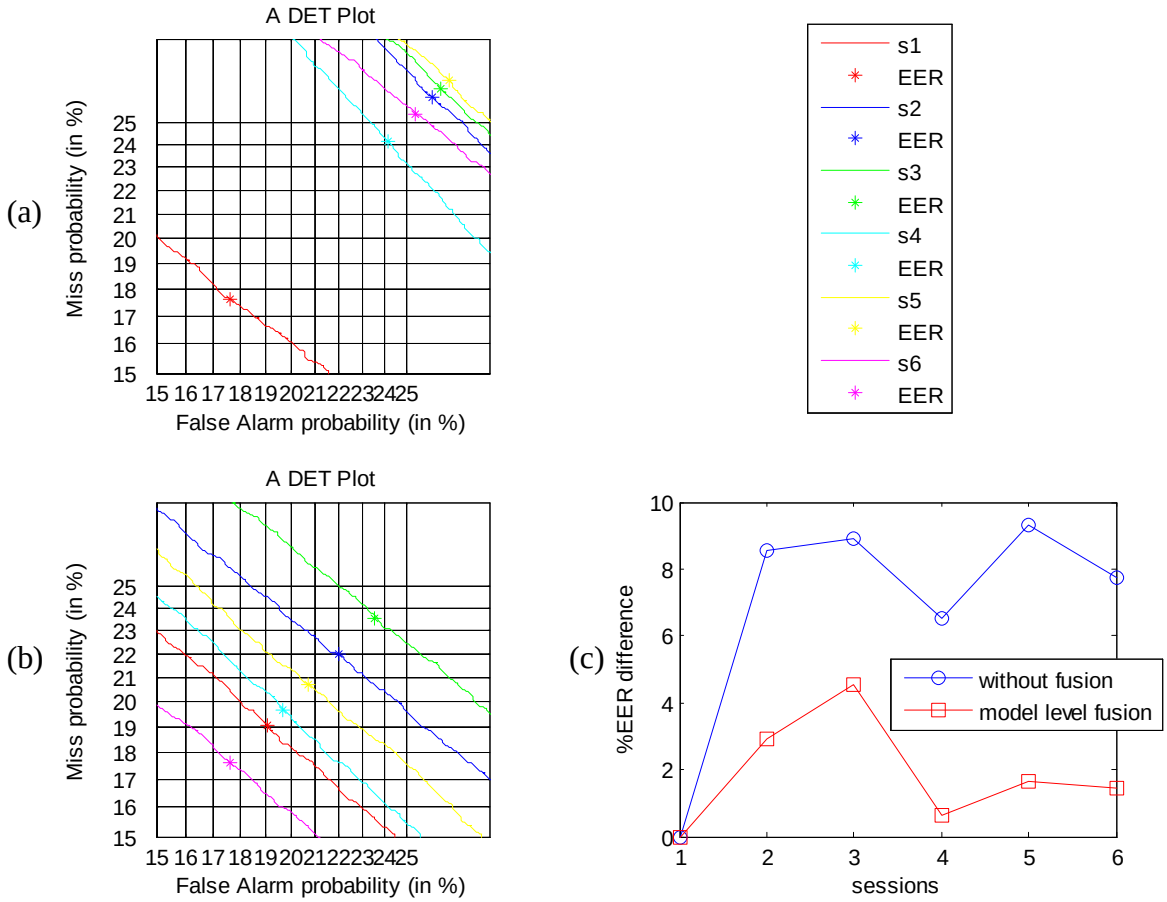


Figure 5.6 Model-level fusion of VTMFCC (a) DET plot of VTMFCC feature without fusion, (b) DET plot of VTMFCC feature with fusion, and (c) % EER difference *vs.* sessions in VTMFCC

Table 5.7 Change in % EER for VTMFCC with model-level fusion of training sessions 1 and 2

Index of Test Sessions	**VTMFCC**	**VTMFCC with model-level fusion**
S1	17.64	19.04
S2	26.18	21.96
S3	26.55	23.56
S4	24.16	19.68
S5	26.95	20.71
S6	25.38	17.6
Max Diff	**9.31**	**4.52**

Training session number 1 is used for the experiment without fusion as shown in 2nd column of the Table 5.7. Training session numbers 1 and 2 is used for the experiment with model-level fusion as shown in 3rd column of Table 5.7.

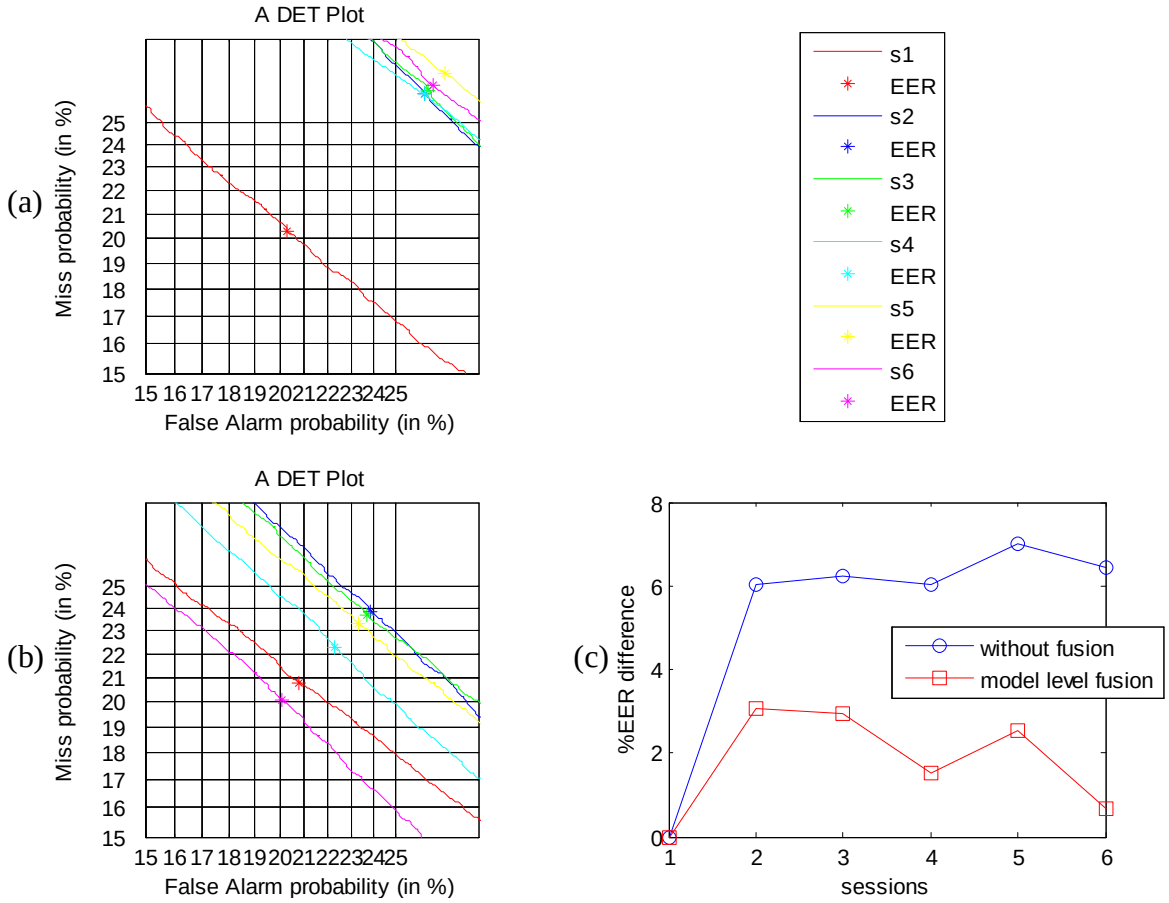


Figure 5.7 Model-level fusion of AM-FM (a) DET plot of AM-FM feature without fusion, (b) DET plot of AM-FM feature with fusion, and (c) % EER difference *vs.* sessions in AM-FM

Table 5.8 Change in % EER for AM-FM with model-level fusion of training sessions 1 and 2

Index of Test Sessions	**AM-FM**	**AM-FM with model-level fusion**
S1	20.30	20.77
S2	26.34	23.85
S3	26.53	23.71
S4	26.31	22.28
S5	27.31	23.31
S6	26.72	20.08
Max diff	**7.31**	**3.07**

Training session number 1 is used for the experiment without fusion as shown in 2nd column of the Table 5.8. Training session numbers 1 and 2 is used for the experiment with model-level fusion as shown in 3rd column of Table 5.8.

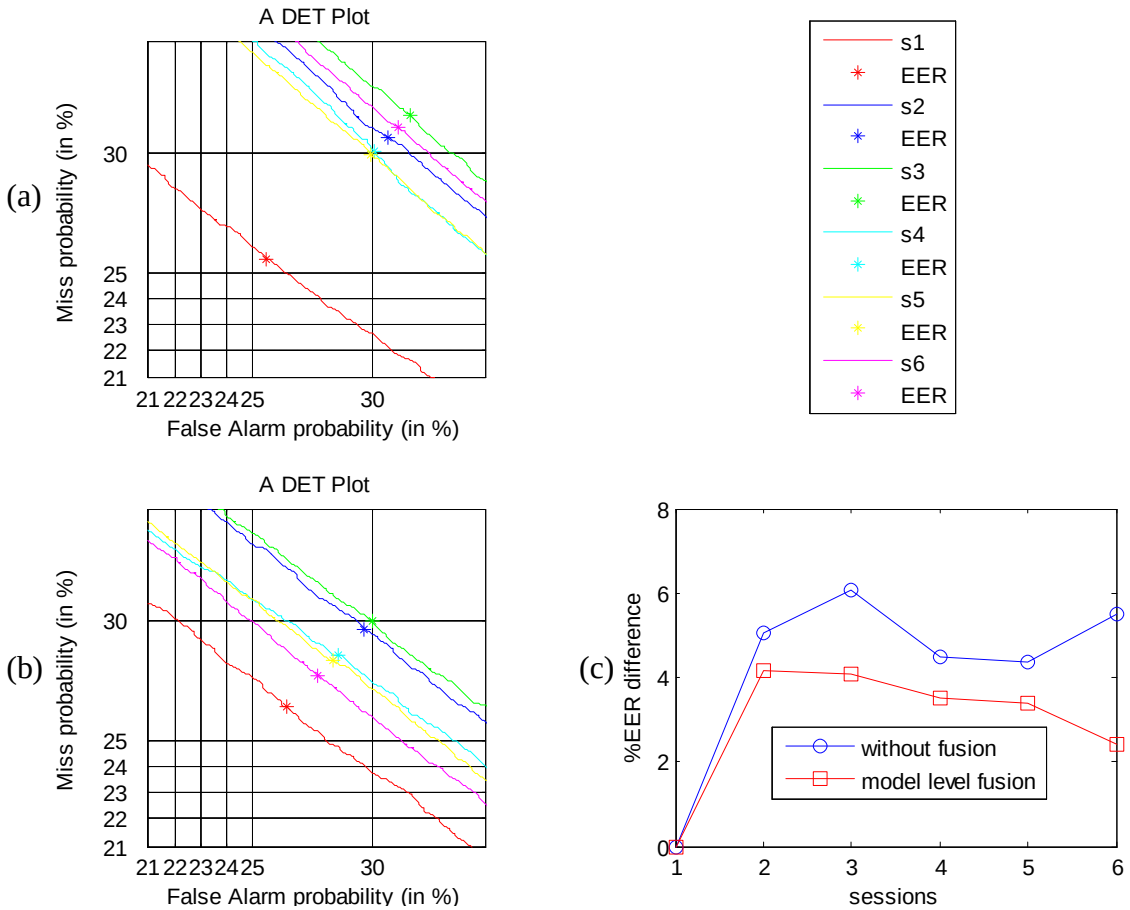


Figure 5.8 Mode-level fusion of Q-feature (a) DET plot of Q-feature without fusion, (b) DET plot of Q-feature with fusion, and (c) % EER difference *vs.* sessions in Q-feature

Table 5.9 Change in % EER for Q-feature with model-level fusion of training sessions 1 and 2

Index of Test Sessions	**Q-feature**	**Q-feature with model-level fusion**
S1	25.35	26.33
S2	30.62	29.63
S3	31.64	29.93
S4	30.03	28.48
S5	29.91	28.25
S6	31.07	27.66
Max Diff	**6.08**	**3.6**

Training session number 1 is used for the experiment without fusion as shown in 2nd column of the Table 5.9. Training session numbers 1 and 2 is used for the experiment with model-level fusion as shown in 3rd column of Table 5.9.

The above plots indicate that model-level fusion shifts the DET curve downward for all the features. Model-level fusion technique reduces the difference between the % EER of particular session to its base session. Hence, it provides robustness against inter-session variability.

5.5 Chapter Summary

The two new approaches; feature-level fusion and model-level fusion are proposed. The experiments on these two approaches are performed. These two approaches are evaluated by DET curves and difference in % EER from its base session (*i.e.* first session). The two approaches successfully provide robustness to the inter-session variability. Feature-level fusion provides better results than the model-level fusion as the overall % EER is lower in the feature-level fusion.

6 Conclusions and Future Directions

6.1 Conclusions

This thesis evaluates the inter-session variability of the person recognition system from hum. The four features, *viz*., MFCC, VTMFCC, AM-FM and Q-features are used in this thesis. This thesis proposes two new approaches to compensate for the degradation due to inter-session variability.

Following are the observation from the results obtained in this thesis:

- Varying the number of filters in the filterbank does not provide improvement in performance for single session, nor robustness against intersession variability.
- As the feature dimension (number of coefficients in feature vector) increases, the performance improves (as the feature occupancy reduces in the high dimensional feature space).
- Increase in the length of the analysis window, improves performance considerably. The performance improves for up to the length of 371.5 *ms*. This is contrary to the common belief that the length of analysis window should be of the order of 10 to 30 *ms*. This may lead to fact that humming signal should have more stationarity than speech signal.
- The two proposed techniques, feature-level fusion and model-level fusion reduces the intersession variability. Training should be done in maximum possible sessions and then these two techniques should be applied to get better results.

6.2 Future Directions

This thesis leads to the following future scope for research:

- This thesis gets better results for longer analysis window. However, the stationarity of the humming signal needs to be further verified statistically.
- The corpus collected in this thesis is gender biased. The corpus should be made gender unbiased by using equal number of male and female subjects.
- The corpus is collected using single microphone. However, this is not the case in practice. So corpus should be collected using different microphones and different acoustic channels like telephone, cellular channel, etc.
- The corpus is collected in college hostel room, so it has similar (not same) type of acoustic environment for all sessions. Instead, data should be collected from different acoustic environments like canteen, market, class room, etc. for more realistic scenarios.
- The database is collected within the span of one month, so it does not include the aging effect of speech. The database should be collected for the very large time span for actual estimation of intersession variability.

Appendix A

MATLAB Code for Joint *pdf*

```
clc;
clear all;
close all;
for i=1:6      %  Loop for 6 sessions
  load(['D:\final thesis\MFCC\24f_512w_12c\test\mfcc_h1_test'
  num2str(i)'.mat']);  % load the features into program

  x=Z(1,:);            % Assign first feature to X
  y=Z(2,:);            % Assign second feature to Y

  s=hist(x,10);        % Plots histogram
  ss=ksdensity(s);     %computes a probability density estimate of the
                       %   sample in the vector x.
  t=hist(y,10);
  tt=ksdensity(t);

  c=kron(ss,tt');       % generate outer product
  figure;

  surf(c);
  xlabel(' MFCC feature 1');
  ylabel('MFCC feature 2');
  title(['session ' num2str(i)]);

end
```

Appendix B
Details of Songs used for Humming

Training

Songs	Movies
Chand sifaris	Fanna
Tu hi meri shab hain	gangster
Mere Sapno ki Rani	Aradhana
Behati hawa sa tha voh	3 idiots
Iktara	Wake up sid
Aye ajababi	Dil se
Mere haath main	Fanna
Teri meri prem kahani	Bodyguard
Luka chuppi	Rang de basanti
Taal se taal	Taal

Testing

Songs	Movies
Give me some sunshine	3 idiots
Allah ke bande	Waisa bhi hota hain part-2
Chalte chalte mere ye geet	Chalte chalte
Tu meri adhuri payas	Gajini
Bhage re man	Chameli

References

[1] G. T. Dimitrov, "Speaker Recognition," Wire Communications Laboratory, Department of Computer and Electrical Engineering, University of Patras, Greece, Ph.D Thesis, November 2005.

[2] Z. Saquib, N. Salam, R. Nair, N. Pandey, and A. Joshi, "A survey on automatic speaker recognition systems," in *Signal Processing and Multimedia : Communications in Computer and Information Science*. Berlin: Springer, 2010, pp. 134-145.

[3] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech Communication*, vol. 52, no. 1, pp. 12-40, January 2010.

[4] www.byclb.com.

[Online]. http://www.byclb.com/TR/Tutorials/neural_networks/ch1_1.htm

[5] www.applied-mathematics.net.

[Online]. http://www.applied-mathematics.net/classification/C45.html

[6] J.P. Marques de sa, *Pattern Recognition : Concepts, Methods and Applications*. Berlin, Germany: Springer, 2001.

[7] S. Theodoridis and K. Koutroumbas, *Pattern Recognition*. San Diago, California, USA: Academic Press, 1998.

[8] D.A. Reynolds, "Automatic speaker recognition: Current approaches and future trends," in *Third Workshop on Automatic Identification Advanced Technologies*, Tarrytown, New York, USA, March14-15, 2002.

[9] M. Jin, J. Kim, and C.D. Yoo, "Humming-based human verification and identification," in *IEEE Int. Conf. on Acoust. Speech and Signal Process., ICASSP 2009.*, April 2009, pp. 1453-1456.

[10] M. Madhavi, "Person recognition from their hum," Dhirubhai Ambani Institute of Information and Communication Technology (DA-IICT), Gandhinagar, M.Tech Thesis 2011.

[11] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Transactions on Acoust. Speech and Signal Process.*, vol. 26, no. 1, pp. 43- 49, February 1978.

[12] BIN-06 pronounciation II class blog. [Online]. http://bin-06.blogspot.in/2011/02/week-7-nasals.html

[13] K.Amino and T. Arai, "Perceptual speaker identification using monosyllabic stimuli - effects of the nucleus vowels and speaker characteristics contained in nasals," in *INTERSPEECH'08*, 2008, pp. 1917-1920.

[14] K. Amino and T. Arai, "Speaker-dependent characteristics of the nasals," *Forensic Science International*, vol. 185, no. 1–3, pp. 21-28, March 2009.

[15] H. Patil, R. Jain, and P. Jain, "Identification of speakers from their hum," *Lecture*

Notes in Artificial Intelligence, LNAI, vol. 5426, pp. 461-468, 2008.

[16] H. Patil, P.Jain, and R. Jain, "A novel approach to identification of speakers from their hum," in *Seventh Int. Conf. on Advances in Pattern Recognition, ICAPR,* 2009, pp. 167-170.

[17] H. Patil and K. Parhi, "Novel variable length teager energy based features for person recognition from their hum," in *Int. Conf. on Acoust. Speech and Signal Process., ICASSP'10*, Dallas, Texas, USA, 2010, pp. 4526-4529.

[18] A. Ghias, D. Logan, and B. Smith, "Query-by-humming- musical information retrieval in an audio database," in *ACM multimedia*, San Francisco, 1995, pp. 216-221.

[19] H. Patil and M.Madhavi, "Significance of magnitude and phase information via VTEO for humming based biometrics," in *5th IAPR/IEEE Int. Conf. on Biometrics (ICB)*, New Delhi, 2012.

[20] M. Faundez-Zanuy and E. Monte-Moreno, "State-of-the-art in speaker recognition," *IEEE Aerospace and Electronic Systems Magazine*, vol. 20, no. 5, pp. 7-12, May 2005.

[21] D.A. Reynolds, "An overview of automatic speaker recognition technology," in *Int. Conf. Acoust. Speech and Signal Process., ICASSP'02*, 2002, pp. 4072-4075.

[22] D. A. Reynolds et al., "The SuperSID project: Exploiting high-level information for high-accuracy speaker recognition," in *Int. Conf. Acoust. Speech, and Signal*

Process., ICASSP'03, Hong Kong, China, 2003, pp. 784-787.

[23] M. Sambur, "Selection of acoustic features for speaker identification," *IEEE Trans. on Acoust. Speech and Signal Process.*, vol. 23, no. 2, pp. 176-182, April 1975.

[24] B. Atal, "Automatic recognition of speakers from their voices," *Proceedings of the IEEE*, vol. 64, no. 4, pp. 460-475, April 1976.

[25] A. Oppenheim and R. Schafer, *Discrete-Time Signal Processing*.: Prentice-Hall, 1989.

[26] T. Quatieri, *Discrete Time Speech Signal Processing, Principles and Practices*.: Pearson Education, 2004.

[27] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously," *IEEE Trans. Acoust. Speech and Signal Process.*, vol. ASSP-28, no. 4, pp. 357-366, August 1980.

[28] J. Kaiser, "On a simple algorithm to calculate the energuy of a signal," in *International Conf. on Acoust. Speech and Signal Process., ICASSP*, 1990, pp. 381-384.

[29] V. Tomer and H. Patil, "On the developement of Variable length Teager Energy operator (VTEO)," in *INTERSPEECH'08*, Brisbane, Australia, 2008, pp. 1056-1059.

[30] K. Paliwal and L. Arsteris, "Usefulness of phase spectrum in human speech perception," in *EUROSPEECH-03*, Geneva, 2003, pp. 2117-2120.

[31] A. Potamianos and P. Maragos, "Speech formant frequencies and bandwidth

tracking using multiband energy demodulation," *Journal of Accoustical Society of America*, vol. 99, no. 6, pp. 3795-3806, 1996.

[32] M. Grimaldi and F. Cummins, "Speaker identification using instantaneous frequencies," *IEEE Trans. on Audio, Speech and Signal Process.*, vol. 16, no. 6, pp. 1097-1111, 2008.

[33] P. Maragos, J. Kaiser, and T. Quatieri, "Energy seperation in signal modulation with application to speech analysis," *IEEE Trans. on Signal Process.*, vol. 41, no. 10, pp. 3024-3051, 1993.

[34] M. Deshpande, "Text-independent speaker identification: new spectral feature extraction techniques," Swami Ramanad Teerth Marathwada University, Nanded, Ph.D Thesis 2010.

[35] W. Campbell, K. Assaleh, and C. Broun, "Speaker recognition with polynomial classifiers," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 4, may 2002.

[36] A. F. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The DET curve in assesment of detection task performance," in *EUROSPEECH'97*, Rhodes, Greece, 1997, pp. 1899-1903.

[37] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861-874, June 2006.