STUDY OF SPEECH ANALYSIS PARAMETERS FOR SPEAKER RECOGNITION

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Master of Technology

by

Gidda Reddy Gangula
(Roll No. 04307046)

under the supervision of

Prof. P. C. Pandey

Department of Electrical Engineering
Indian Institute of Technology, Bombay
July 2006

Gidda Reddy Gangula / Prof. P. C. Pandey (supervisor): "Study of speech analysis parameters for speaker recognition", *M.Tech. dissertation*, Department of Electrical Engineering, Indian Institute of Technology, Bombay, July 2006.

ABSTRACT

The objective of this project is to study speech analysis parameters to improve the performance of speaker recognition systems. Towards this end, peak amplitudes of the speech signal and Harmonic plus Noise Model (HNM) parameters are investigated. The investigated HNM parameters include maximum voiced frequency (F_m), pitch (F_0), and relative noise band energy (α). Distribution of these parameters is studied, using statistical moments and correlation coefficients. HNM parameters have shown good variation across the speakers and may be useful for speaker recognition. Speaker recognition experiments were conducted on HNM parameters using VQ algorithm, with Mahalanobis distance measures. The experiments have shown that the performance of speaker recognition based on HNM parameters is comparable to that of well established Mel Frequency Cepstral Coefficients (MFCC). Further, the performance of the speaker recognition is improved by using MFCC and the three HNM parameters together.

ACKNOWLEDGEMENTS

I would like to express my deep sense of gratitude to Prof. P. C. Pandey for his invaluable help, guidance, support, and encouragement throughout the course of the project. I am thankful to Prof. Preeti Rao for her valuable suggestions during my presentations.

I am thankful to my lab-mate Parveen K. Lehana for his help during the discussions and particularly in HNM coding. I would also like to thank P. Pradeep Kumar of Digital Audio Processing Lab for his valuable suggestions. I am also thankful to Pushkar P. Patwardhan from the above lab for his valuable help in providing me a speech database. I would like to express my thanks to my colleagues Priyanko Mitra and A. R. Jayan for many interesting and helpful discussions. I would also like to acknowledge the support of my friends whose recordings have been used in my study. I am also grateful to my lab-mates J. N. Sarvaiya, L. Venkatachalam, and Vinod K. Pandey for their memorable and enjoyable company.

Gidda Reddy Gangula

July 2006

CONTENTS

LIST OF SYMBOLS

Symbol	Explanation
F_m	Maximum voiced frequency
F_0	Fundamental frequency or pitch
α	Relative noise band energy
M_1	Mean value
M_2	RMS deviation about mean
M_3	Skewness
M_4	Kurtosis
ΔF_m	Jitter in maximum voiced frequency
ΔF_0	Jitter in pitch
ΔA_0	Shimmer in intensity
$\overline{x}$	Mean value
σ_x	RMS deviation about mean
r	Correlation coefficient
F	F-ratio value
E	Energy of the speech frame
r_1	First order reflection coefficient
e_p	LPC residual signal
$\boldsymbol{\mu}$	Mean vector
$\mathbf{C}$	Covariance matrix
F_2	Second formant frequency
F_3	Third formant frequency
F_4	Fourth formant frequency
$a_k(t)$	Time varying amplitude of the k^{th} harmonic
$\theta_k(t)$	Time varying phase of the k^{th} harmonic
G	Gain scaling factor
a_k	LPC filter coefficients

$H(z)$	Transfer function of the vocal tract
$\varepsilon(n)$	LPC error signal
c_n	Cepstrum obtained from LPC coefficients
B	Bandwidth of the speech signal
$X(\omega)$	Speech spectrum
$V(\omega)$	Frequency response of vocal tract filter
$E(\omega)$	Excitation spectrum
$c(n)$	Cepstrum obtained using FFT
F_M	Mel frequency
$C_{\text{Mel}}(n,m)$	MFCC for the speech frame at time n with order m
$s'(t)$	Synthesized harmonic part of the speech
$n(t)$	Synthesized noise part of the speech
$h(\tau,t)$	Time-varying normalized all-pole filter
$w(t)$	Energy envelope function

LIST OF ABBREVIATIONS

Abbreviation	Explanation
AR	Autoregressive
ASI	Automatic speaker identification
ASV	Automatic speaker verification
D.F.	Degrees of freedom
dB	Decibel
DCT	Discrete cosine transform
DTW	Dynamic time warping
EER	Equal error rate
FFT	Fast Fourier transform
GCI	Glottal closure instant
GMM	Gaussian mixture model
HMM	Hidden Markov model
HNM	Harmonic plus noise model
IFFT	Inverse fast Fourier transform
LPC	Linear predictive coding
LPCC	Linear prediction cepstral coefficients
LSP	Line spectrum pair
MFCC	Mel frequency cepstral coefficients
MSA	Mean square among speakers
MSW	Mean square within speakers
PLP	Perceptual linear prediction
TTS	Text to speech
VQ	Vector quantization

LIST OF FIGURES

LIST OF TABLES

Chapter 1

INTRODUCTION

1.1 Project overview

Speaker recognition is the process of identifying a speaker from the speaker information contained in the speech utterance. It is a pattern recognition problem and it is divided into two stages: training stage and testing stage [1], [2], [3]. Training stage consists of extraction of speaker dependent features from the training utterance and creation of a model representing the speaker, from the extracted features. Testing stage consists of comparison of speaker dependent features, extracted from the test utterance, with the existing models of the speakers in the database to recognize the target speaker. Hence, extraction of speaker dependent features plays an important role in speaker recognition. Speaker dependent features are the parameters that contain specific speaker information useful for speaker recognition.

The specific speaker information in speech is due to the differences in physiological and behavioral aspects of the speech production system of the speakers [2], [4], [5]. Mel frequency cepstral coefficients (MFCC) and linear prediction cepstral coefficients (LPCC) are the well established speaker dependent features for speaker recognition [6]. The performance of the speaker recognition systems based on these features is still not satisfactory. The speaker recognition systems employing MFCC or LPCC features do not perform well in text-independent environment with limited training data [6]. Further, the performance of these systems is generally affected by background noise, transmission medium characteristics, etc [1]. In order to improve the performance of the speaker recognition systems for real world applications, additional parameters that can separate speaker dependent information from the linguistic information should be investigated.

After extracting the speaker dependent features, the next step in speaker recognition process is to create a model representing the speaker. Minimum distance

classifier (or long term averaging), vector quantization (VQ), hidden Markov model (HMM), and Gaussian mixture model (GMM) are the commonly used techniques for speaker modeling [1], [7]. Among these techniques, VQ method performs better than minimum distance classifier technique, while HMM and GMM techniques perform well (better than VQ method) only when large amount of data are available for training [7], [8].

1.2 Project objective

The objective of this project is to investigate various sets of parameters and to establish the most suitable set of parameters for improving the performance of speaker recognition systems. The parameters to be investigated should contain speaker information separated from the linguistic information. Towards this end, investigations are carried out on the peak amplitudes of the speech signal and the parameters of the Harmonic plus Noise Model (HNM). The investigated parameters are studied using statistical methods, namely, histogram analysis, statistical moments, and correlation, in order to compare their intra-speaker and inter-speaker variability. Finally, speaker recognition experiments are performed on these parameters using VQ algorithm for speaker modeling.

1.3 Dissertation outline

The next chapter describes the speaker recognition overview, which includes different types of speaker recognition systems and their applications, extraction of well established features for speaker recognition, and different speaker modeling techniques. Chapter 3 describes the investigated parameters for speaker recognition: peak amplitudes of the speech signal and HNM parameters (maximum voiced frequency (F_m), pitch (F_0), and relative noise band energy (α)). In Chapter 4, the investigated parameters are studied through different statistical methods, namely, histogram analysis, statistical moments, and correlation, to determine their possible use in speaker recognition. Chapter 5 describes the VQ algorithm for speaker modeling and discusses the results of speaker recognition experiments conducted on HNM parameters and MFCC features. In Chapter 6, a summary of the work carried out and suggestions for the future work of the project are presented. Appendix A provides the results of multi-sample histogram analyses on the investigated parameters (peak amplitudes and HNM parameters).

Chapter 2

SPEAKER RECOGNITION

Speaker recognition consists of two stages namely training stage and testing stage [1], [2], [3]. During the training stage, the features relevant to the speaker, also called speaker dependent features, are to be extracted from the training utterance and a model of the speaker is to be created. During the testing stage (i.e. recognition stage), the features extracted from the test utterance are to be compared with the existing models of the speakers in the database to recognize the speaker. Extraction of speaker dependent features plays an important role in speaker recognition. Well established speaker dependent features for speaker recognition are mel frequency cepstral coefficients (MFCC), linear prediction cepstral coefficients (LPCC), etc [6]. After extracting the speaker dependent features, a model representing the speaker is to be created. Most commonly used techniques to create a model of the speaker are minimum distance classifier, vector quantization (VQ), hidden Markov model (HMM), and Gaussian mixture model (GMM) [1], [7]. This chapter gives an overview of speaker recognition systems and their applications, extraction of established speaker dependent features, and most commonly used techniques for speaker modeling.

2.1 Speaker recognition systems

A block diagram of a speaker recognition system is shown in Fig. 2.1, containing two stages; training stage and testing stage [1], [2], [3]. The tasks in the training stage are:

a. Extraction of speaker dependent features from the training utterance.
b. Creation of a model representing the speaker from the extracted features.
c. Storing the speaker models in the database.

Similarly, the tasks in the testing stage are:

a. Extraction of speaker dependent features from the test utterance.

b. Comparison of the extracted features with the existing speaker models in the database by calculating the match score using different techniques such as distance measure, maximum likelihood, etc.

c. Finally, making a decision based on the matching score obtained in the previous step.

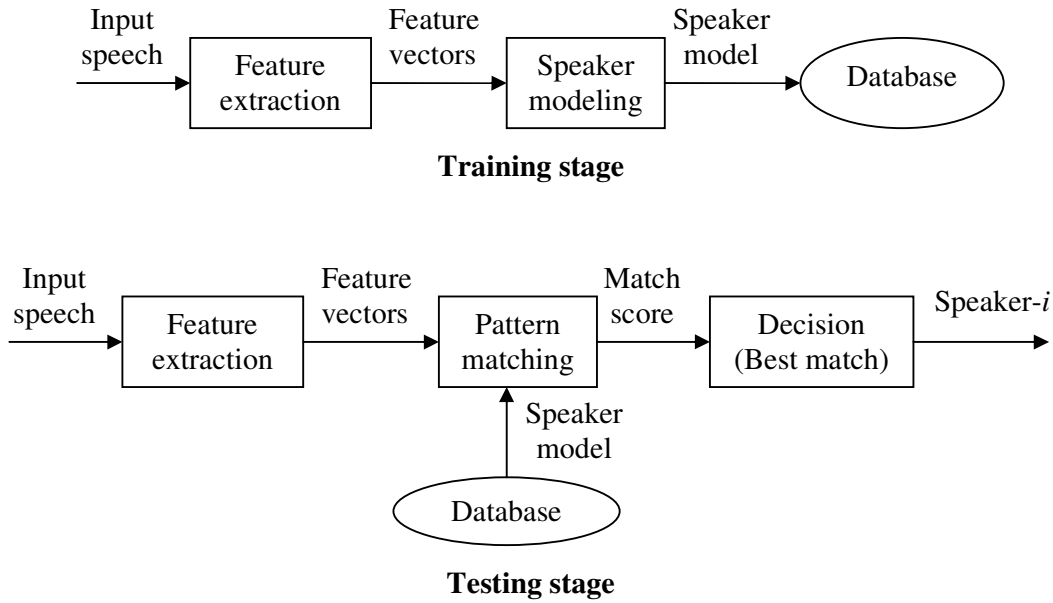

Fig 2.1 Block diagram of speaker recognition system, from [1]

Commonly used features for speaker recognition are formants, MFCC, LPCC, etc. The techniques used for speaker modeling are VQ method, HMM, GMM, etc. The details of extraction of speaker dependent features and the techniques to create a model of the speaker are described in Sections 2.2 and 2.3 respectively.

2.1.1 Types of speaker recognition systems

The speaker recognition systems are divided into two categories based on their task [1], [9]:

- Automatic speaker verification (ASV)
- Automatic speaker identification (ASI)

In case of automatic speaker verification, the task is to verify whether the person claiming is the correct speaker or not. In this process, the person initially claims an identity in the database. Then, the system compares the features of the test utterance with the claimed

speaker model. Finally, the match score will be compared with a threshold and the system will accept or reject the claim depending on the comparison result.

In case of automatic speaker identification, the task is to recognize one speaker from many speakers. Hence, the features of the test speaker will be compared with all the speaker models in the database and the one which gives the best match will be identified as the target speaker. Speaker identification can be either closed set or open set identification. In closed set speaker identification, system determines to which member of the database of speakers the voice belongs to. In open set speaker identification, the system has also to decide whether the person identified actually belongs to the existing speaker database or not. Open set identification is a closed set identification with an additional verification. In open set identification, if the verification fails, the test speaker will be declared as an unknown speaker.

Based on the sentences used in training and testing stages, the speaker recognition systems are divided into three categories [9]:

- Text dependent speaker recognition systems
- Text independent speaker recognition systems
- Text prompted speaker recognition systems

In text dependent speaker recognition, voice recordings (utterances) of the same sentence will be used in both training and testing phases. Text dependent recognition is easier to implement and more reliable than text independent recognition [10]. Also, text dependent speaker recognition systems require fewer training sentences. The performance of text dependent systems is highly correlated with the vocabulary that is chosen [10]. These systems are used where speakers are cooperative such as in security application for accessing computer system and for control access to a physical facility.

In text independent speaker recognition, the test sentence will be different from the training sentence. Text independent speaker recognition systems use long-term statistical data. These systems require more duration of training data (approximately 30 s) to ensure that a wide variety of sounds are spoken, providing more information to build a voice model for the speaker [8], [10], [11]. Text independent systems are suitable for applications where speakers are uncooperative such as in forensic and law and order applications. In both text dependent and text independent speaker recognition systems, there is a possibility of capturing of the voice by imposters. This problem can be solved in text prompted speaker recognition systems.

In text prompted speaker recognition, the utterance to be produced by the speaker is not predetermined. In each access to the system, the system will prompt the speaker to say a particular sentence or a word. In this case, the system will initially verify the sentence uttered and then will verify the speaker. If the sentence spoken by the person does not match with the one prompted by the system, the system will reject the speaker.

2.1.2 Applications of speaker recognition systems

Speaker recognition systems have many applications [1], [2], [3], for example,

1. Security: Voice controlled access to secure facilities such as access to buildings, bank accounts etc.
2. Remote authentication: Financial transaction/ banking where authorization is done by telephone voice.
3. Forensics: In law enforcement, speaker recognition systems can be used to identify the suspect from the voice.
4. Multi-speaker environment: Identification of a particular person's voice in the cases where many speakers are present such as teleconferencing, panel discussions, etc.
5. Gender recognition can be used to improve the performance of the speech recognizer.
6. Speaker recognition approach can be used to estimate a person's age approximately.

2.1.3 Evaluation of speaker recognition systems

The performance of a speaker recognition system depends on many factors such as background noise, transmission channel characteristics, etc [1]. The performance of a speaker recognition system can be evaluated from the errors produced by the system in different tests. These errors can be controlled by a threshold value used in the testing phase [1], [9], [12]. The threshold value plays an important role in speaker recognition. In both ASI and ASV systems, the match score (assumed distance measure) will be compared to a threshold below which the speaker will be accepted. Therefore, the estimation of the optimal threshold is an important task in the case of a good performance system. If the decision threshold is too low, too many speakers will be rejected as imposters, such an error is referred to as false rejection [9], [12]. Such a threshold will screen imposters very well, but at the cost of high rejection rate. If the threshold is too

high, too many imposters will be accepted, such an error is called as false acceptance or false alarm [9], [12]. Such a threshold will accept valid speakers with little difficulty, but at the cost of high imposter acceptance rate. Low thresholds are generally preferred because false acceptances are usually more expensive. The performance of speaker recognition systems is often measured in terms of equal error rate (EER), corresponding to the decision threshold in which the false rejection rate is equal to the false acceptance rate [12].

2.2 Features for speaker recognition

As discussed earlier, extraction of speaker dependent features from the speech plays an important role in speaker recognition. Speaker dependent features are the parameters containing the information related to a particular speaker. For the application of speaker recognition, these features should vary as much as possible across the speakers and must be consistent within the speaker [1], [2], [4], [13].

In general, speech contains the information related to both the speaker and the linguistic message. Hence, the performance of a speaker recognition system depends on how well the speaker information (speaker dependent features) is separated from the linguistic information [4].

Due to the differences in physiological and behavioral aspects of the speech production system across speakers [2], [4], [5], speech contains specific characteristics that are related to a particular speaker. The physiological features representing the speaker in speech are vocal tract shape, pitch and nasality [2], [4]. The differences in the shape of the vocal tract will cause the differences in the characteristic resonances (also called formants) of the spectrum of the speech signals. Similarly, the differences in the pitch are due to the variations in the size of the vocal chords and the differences in the nasalized speech are due to the variations in the size of the velum. Hence, some of the speaker dependent features are indicated by the formant frequencies and pitch.

It is common in speaker recognition systems to make use of the features derived only from the vocal tract, because vocal tract shape is the one very difficult to be imitated by a second person. The formant frequencies of the speech are directly related to the unique shape of the vocal tract and supplies important information about the speaker's identity. In a study by Sambur [4], it is shown that F2 in nasals, F2, F3, and F4 in vowels, and mean F0 are useful for speaker recognition. Sambur carried out the investigation by first determining an initial set of acoustic parameters which, on the basis of theoretical

considerations and past experimental details [4], can indicate the unique properties of a speaker's vocal apparatus. Then, the investigated features are evaluated by using probability of error criterion. This criterion is a method of feature evaluation in which the relative performance of each feature is estimated to determine the ordered list of feature effectiveness. There is a difficulty in calculation of the formant frequencies accurately. Several other sets of parameters which indirectly represent the formants, namely LPC coefficients, cepstral parameters, MFCC, LPCC, PLP coefficients, delta cepstrum, etc, have been investigated [1], [6], [13], [14]. Among these, MFCC and LPCC are the well established parameters for speaker recognition. An overview of LPC analysis, cepstral analysis and mel frequency cepstral analysis is described in the following subsections.

2.2.1 Linear predictive coding analysis

Linear predictive coding (LPC) analysis is a powerful tool used to separate vocal tract filter from the glottal excitation [2], [4], [15]. The basic assumption of LPC analysis is the representation of the transfer function of the vocal tract $H(z)$ by an all-pole model as:

$$H(z) = \frac{G}{1 - \sum_{k=1}^{P} a_k z^{-k}} \tag{2.1}$$

where P represents the order of the filter, G is a gain scaling factor, and a_k, for $k = 1, 2, ..., P$ are the LPC filter coefficients. It has been empirically found that $P = 2(B+1)$, where B is the speech signal bandwidth in kHz, is adequate. From the above equation, the discrete time response $x(n)$ to an excitation signal $e(n)$ is given by,

$$x(n) = Ge(n) + \sum_{k=1}^{P} a_k x(n-k) \tag{2.2}$$

The coefficients for the second term of the above expression are generally computed to give an approximation to the original sequence, which will yield a spectrum for $H(z)$ that is an approximation to the original speech spectrum. Thus, the speech signal is predicted by a weighted sum of its previous values, given by,

$$x'(n) = \sum_{k=1}^{P} a_k x(n-k) \tag{2.3}$$

The difference between the predicted value and the actual value is referred to as the error signal (also called residual error) given by $\varepsilon(n) = x(n) - x'(n)$. The coefficients $\{a_k\}$ are chosen to minimize the mean squared error. The resulting error signal can be viewed as an approximation to the excitation function.

The LPC coefficients $\{a_k\}$ contain vocal tract filter related speaker information and are useful for speaker recognition [2], [4]. These coefficients are meaningful only as a set in describing the vocal tract filter and can not be individually smoothened across analysis frames. Hence, they are often transformed into other parameters, namely, reflection coefficients, log area ratios, cepstral coefficients, etc. The details of these transformations are given in [6], [13], [14], [15]. Among these, LP derived cepstral parameters are the most commonly used features for speaker recognition [6], [13], [14]. The cepstrum parameters, $c(n)$, can be calculated from the LPC coefficients using the following formula [15]:

$$\left.\begin{aligned} c_1 &= a_1 \\ c_n &= a_n + \sum_{k=1}^{n-1}\left(1 - \frac{k}{n}\right)a_k c_{n-k} \qquad 1 \le n \le p \end{aligned}\right\} \tag{2.4}$$

The main advantage of all-pole model of vocal tract filter is the efficient computations involved in estimating the coefficients. However, this model is suitable only for non-nasalized sounds, with glottal excitation. For sounds involving nasal pathway and sound segments with frication as excitation source, the vocal tract filter contains both poles and zeros. During these segments, an all-pole model with higher order reduces the errors [2], [4].

2.2.2 Cepstral analysis

The basic idea behind the cepstral analysis is the de-convolution of the excitation and the vocal tract response [12], [14]. Using source-filter model of speech production, the spectrum of the speech signal, $X(\omega)$, can be represented as the product of the excitation spectrum, $E(\omega)$, and the vocal tract filter spectrum, $V(\omega)$:

$$X(\omega) = E(\omega)V(\omega) \tag{2.5}$$

Taking the logarithm of the magnitude on both sides of the above equation, we get

$$\log|X(\omega)| = \log|E(\omega)| + \log|V(\omega)| \tag{2.6}$$

From the above equation, it is clear that the logarithmic spectrum is separated as two parts, namely, the log spectral components that vary rapidly with ω (high-time components; first term in the right side of the above equation) and the log spectral components that vary slowly with ω (low-time components; second term in the right side of the above equation). This process is called de-convolution. The cepstrum is given by taking the inverse Fourier transform of the log magnitude spectrum [12], [16].

$$c(n) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \log\left|X(\omega)\right| e^{i\omega n} d\omega \tag{2.7}$$

where $c(n)$ is the n-th cepstral coefficient. Thus, the contribution of the excitation and the vocal tract filter can be separated in cepstral domain. Both components can be inverted to generate the original spectral magnitudes. However, separation of these two components is very difficult, due to the possibility of overlapping of the two components in the cepstrum. Fig. 2.2 describes the cepstral analysis method.

For speaker recognition, LPC method is more suitable than cepstral analysis method because LPC scheme is directly related to the formant frequencies of the spectrum. In addition, cepstral analysis scheme involves more computations than that of the LPC analysis. The advantage of cepstral analysis is that it does not refer to any model as happened in the case of LPC analysis. However, cepstral analysis is practically used for F_0 (pitch) or formant estimation especially in speech recognition [9]. For speaker recognition systems, a modified form of cepstral analysis called mel frequency cepstral analysis is useful as described in the following subsection.

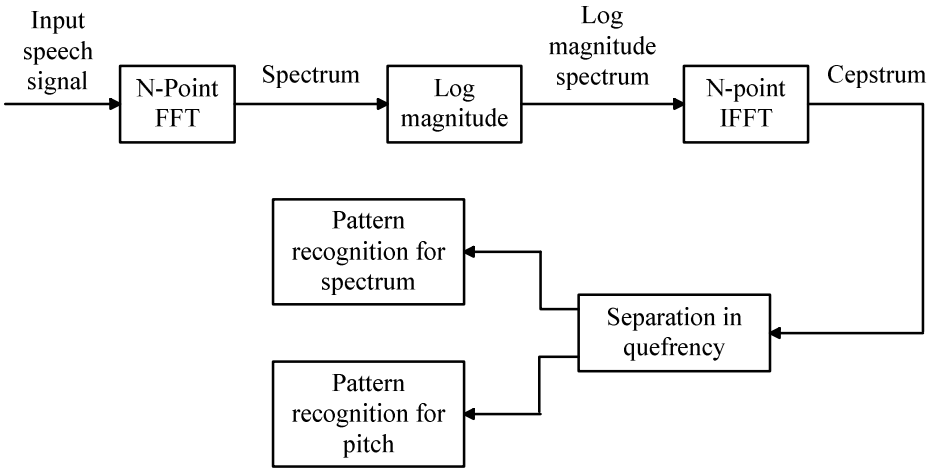

Fig 2.2 Description of cepstral analysis, from [9]

2.2.3 Mel Frequency Cepstral Coefficients (MFCC)

Mel frequency scale is a mapping of the real pitch to the pitch perceived by the listeners [5], [12]. Most of the researchers modeled the data by a linear fit for the frequency below 1 kHz and by a logarithmic fit for higher frequency. Frequency f in Hz can be related to mel frequency F_M as [12]:

$$F_M = 2595\log\left(1+\frac{f}{700}\right) \tag{2.8}$$

The auditory system perceives information based on the energy in a band of frequencies rather than at a single frequency [5]. A filter bank analysis, as shown in Fig. 2.3, with the filters having higher center frequency have a wider bandwidth as compared to a filter having a lower center frequency, is used to approximate the response of the auditory system.

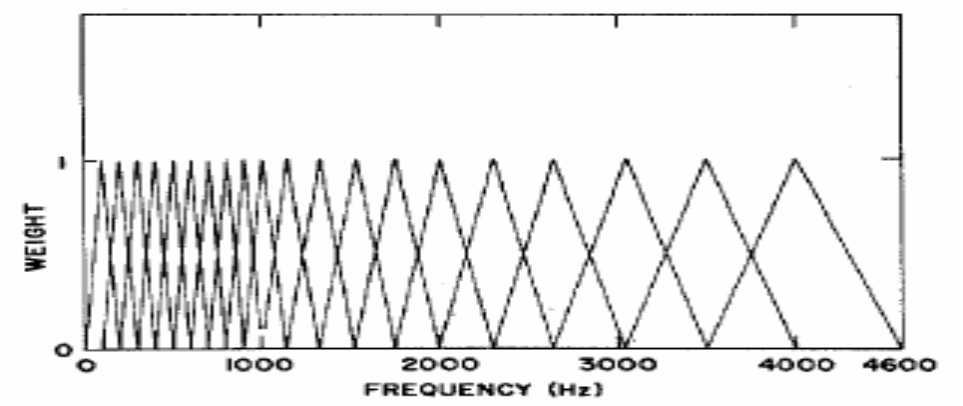

Fig 2.3 Mel scale filter bank, from [12]

The mel filter bank weights the speech spectrum $X(n,\omega_k)$, and the energy output of each filter in the filter bank is calculated by using the following equation.

$$E_{Mel}(n,l) = \frac{1}{A_l}\sum_{k=L_l}^{U_l} \left|V_l(\omega_k)\,X(n,\omega_k)\right|^2 \tag{2.9}$$

where $E_{\text{Mel}}(n,l)$ represents energy for speech frame at time n and for the l^{th} mel scale filter, $V_l(\omega_k)$ represents frequency response of l^{th} mel scale filter, $X(n,\omega_k)$ represents the magnitude response of the speech signal, L_l and U_l represent the lower and upper limits respectively of each filter in the filter bank, and A_l is the band normalizing factor given as

$$A_l = \sum_{k=L_l}^{U_l} \left|V_l(\omega_k)\right|^2 . \tag{2.10}$$

The mel cepstrum is computed as the discrete cosine transform (DCT) of the log mel energy spectrum [12].

$$c_{Mel}(n,m) = \frac{1}{R}\sum_{l=0}^{R-1} \log_{10}(E_{Mel}(n,l))\cos\left(\frac{2\pi}{R}(l+1/2)m\right) \tag{2.11}$$

where R represents the number of filters in the mel filter bank and m represents the number of mel cepstral coefficients calculated for each frame.

A comparison of the performances of various speaker dependent features is described in [1], [6], [13]. For speaker recognition, linear prediction cepstral coefficients perform better than MFCC in many cases (for e.g., small analysis order) [13]. Some other

commonly used features for speaker recognition are perceptual linear prediction (PLP) coefficients, line spectrum pair (LSP) frequencies, and delta cepstrum. The details of these techniques are given in [1], [6]. The LSP and PLP features gave better performance than cepstrum [6]. Among all the features LPCC and MFCC features gave overall better performance [1].

The effectiveness of a parameter for speaker recognition can be evaluated by a measure called the F-ratio [4], which compares inter-speaker and intra-speaker variances. Another way to evaluate the utility of features for speaker recognition is the probability of error criterion [4]. Also, the effectiveness of the features for speaker recognition can be compared by a method called add-on procedure [12].

2.3 Speaker modeling

After extracting the speaker dependent features, the task of the speaker recognition system is to create a model of the speaker. These models can be classified as stochastic and template [1]. In template model, the pattern matching is deterministic while in stochastic model, the pattern matching is probabilistic, leading to probabilistic measures. These two models are described in brief here.

2.3.1 Template models

Minimum distance classifier, VQ, and DTW techniques are the commonly used template models for speaker recognition [1], [7]. Minimum distance classifier (or long term averaging) is the simplest approach to speaker modeling that involves the calculation of average and covariance of the feature vectors over multiple analysis frames of the training data, called mean and covariance vectors [17]. These mean and covariance vectors will represent the model for a particular speaker. The mean and covariance of N feature vectors $\mathbf{x_i}$ are given by:

- The mean vector:

$$\boldsymbol{\mu} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x_i} \tag{2.12}$$

- The covariance vector:

$$\mathbf{C} = \frac{1}{N-1}\sum_{i=1}^{N}[(\mathbf{x_i}-\boldsymbol{\mu})^T(\mathbf{x_i}-\boldsymbol{\mu})] \tag{2.13}$$

In recognition, the distance between the average test and training vectors will be calculated and the one with the smallest distance from that of the test speaker will be identified as the target speaker. Most commonly used distance measure is the Euclidean distance [1]. The Euclidean distance between two vectors **x** and **y** is given by:

$$d(\mathbf{x},\mathbf{y}) = \sqrt{\sum_{k=1}^{N} |x_k - y_k|^2} = \sqrt{(\mathbf{x}-\mathbf{y})(\mathbf{x}-\mathbf{y})^T} \tag{2.14}$$

The weighted Euclidean distance is defined as:

$$d_W(\mathbf{x},\mathbf{y}) = \sqrt{(\mathbf{x}-\mathbf{y})\,\mathbf{W}\,(\mathbf{x}-\mathbf{y})^T} \tag{2.15}$$

Here **W** is a weighting matrix. If **W** is identity matrix the distance is Euclidean; if **W** is inverse covariance matrix of **x**, then it is Mahalanobis distance [1]. The Mahalanobis distance gives less weight to the components having more variance.

Long term averaging is useful for text-independent recognition [17], where large amounts of data (> 20 s) are required to construct speaker models. A problem with this method is that it does not distinguish between acoustic speech classes (like vowels, fricatives, consonants, etc), i.e., it uses an average of feature vectors per speaker computed over all sound classes. This shortcoming can be solved using vector quantization.

Vector quantization (VQ) [8], [10], [11], [13] is the most commonly used method for speaker modeling. In this method, the feature vectors will be divided into clusters using K-means clustering algorithm. Then, centroids (called mean vectors) will be calculated for each cluster and each centroid in the clustering represents an acoustic class, but without identification or labeling. These centroids are also called as code vectors and a collection of such code vectors is called codebook. The codebook will be represented as a speaker model.

VQ method performs better than minimum distance classifiers [1], [11]. But, the clustering procedure averages out temporal information, leading to loss of speaker dependent temporal information. And also the codebook size affects the performance of the system. Large codebook characterizes the speaker's voice better, thus reducing the recognition error, but at the cost of computational expense. More details of this technique are presented in Chapter 5.

The above two methods belong to the template model category. There is another method that comes under template model called Dynamic Time Warping (DTW). It is a

text dependent model used to compensate for speaking rate variability. The details of this model are given in [1], [7], [14].

2.3.2 Stochastic models

Stochastic model of speakers involves modeling by probability distribution rather than average features, and the recognition decisions are based on probabilities or likelihoods rather than distances to average features. Stochastic model for speaker recognition offers more flexibility and results in more theoretically meaningful probabilistic likelihood score [1]. The pattern-matching problem is formulated as measuring the likelihood of an observation, given the speaker model. Hidden Markov Model (HMM) and Gaussian Mixture Model (GMM) will come under stochastic models [7], [8], [10], [18]. In HMM, the speakers are modeled by the information relating to hidden state probabilities and the state transition probabilities. In GMM, the speakers are modeled by different Gaussian probability density functions represented with Gaussian mixtures that contain mean, variance, and probabilities (more exactly weights) corresponding to each Gaussian probability density function. These two models are discussed in detail in [8], [18].

In speaker identification, the scoring is based on minimum distance for VQ and maximum likelihood for HMM and GMM. In speaker verification the minimum distance should be less than threshold for VQ and maximum likelihood should be more than threshold for GMM and HMM. A comparison of the speaker modeling techniques discussed above with various speaker dependent features is given in [1]. For text dependent recognition, DTW outperforms both VQ and HMM [7], [19]. While HMM is as robust as VQ method, and when amount of available data are limited, VQ method is more robust than HMM [7], [10]. Also, VQ is numerically stable and fast, but recognition performance is low compared to GMM [8]. When trained with small amount of data, GMM can get numerically unstable [6], [8].

2.4 Problems with the existing speaker recognition systems

A complete speaker recognition system consists of extraction of speaker dependent features, creation of speaker models from the extracted features, and comparison of test features with existing models. The performance of the speaker recognition system depends mainly on the speaker dependent features and then on the speaker modeling techniques used. The recognition systems with MFCC and LPCC features have shown good performance using VQ, GMM and HMM modeling techniques [1], [6], [13]. The

existing speaker recognition systems perform well under laboratory conditions. But, they do not provide satisfactory results in real world situations. The main reason for this is that the speech features used may not be providing sufficient separation of speaker information from linguistic information. Further, the estimation of the speaker dependent features may get affected by background noise, transmission medium characteristics, microphone variability, etc [1], [20].

Hence, for developing a recognition system that will perform well even under adverse conditions, there is a need to investigate new speaker dependent features containing more speaker information and less variation with background noise, channel characteristics, etc. For this purpose, speaker dependent features should have the following properties:

- They should be easily measurable and occur frequently in normal speech.
- They should have a large variation across the speakers and low variation in the speech of a single speaker.
- They should be minimally affected by a moderate level of background noise or specific transmission channel characteristics.
- They should not be affected by the health of the speaker.
- They should not be modifiable by conscious effort of the speaker.

Thus, the objective of this project is to investigate some of the speaker dependent features for speaker recognition applications.

Chapter 3

PARAMETERS INVESTIGATED FOR SPEAKER RECOGNITION

The well established parameters for speaker recognition are mel frequency cepstral coefficients (MFCC) and linear prediction cepstral coefficients (LPCC). These parameters are primarily associated with vocal tract. The difficulty with these parameters is in separating speaker information from linguistic information. Hence, more parameters are to be investigated for speaker recognition. For this purpose, two sets of parameters have been investigated: peak amplitudes of the speech signal and some of the parameters of Harmonic plus Noise Model (HNM) analysis [21], [22], which are related to both the vocal tract and vocal chords.

3.1 Peak amplitudes of the speech signal

Because of the variations in size and shape of the vocal tract from speaker to speaker, the resonance frequency and formant bandwidth of the speech signal vary. Narrow resonance bandwidth will cause more ringing of the speech signal and wider bandwidth will cause less ringing, with a glottal pulse duration. A high formant frequency will emphasize higher harmonic of the pitch fundamental. Thus, the variations in resonance frequencies and formant bandwidths will cause the variations in ringing and decay patterns and hence the peak amplitudes of the speech signal. So, the peak amplitudes of the speech signal are associated with vocal tract that will vary due to the variations in size and shape of the vocal tract.

Positive peak amplitude is defined as a positive sample value, which is greater than the preceding and the following samples. Negative valley is defined as a negative sample, which is less than the preceding and the following samples. Peak amplitude is either a positive peak or a negative valley.

As the peaks and positive peaks are associated with vocal tract, they may be useful for speaker recognition. For application to speaker recognition, the pattern of variations in peaks and positive peaks should be studied across speakers. Some of the

techniques to study certain variation in a parameter are histogram analysis, statistical moments, and correlation. These techniques are discussed in Chapter 4.

3.2 Harmonic plus Noise Model (HNM)

Harmonic plus Noise Model (HNM) was applied by Stylianou for concatenative text-to-speech (TTS) synthesis [21], [23], [24]. HNM is a parametric model of speech and it contains various parameters that are used during the synthesis of speech. The parameters of this model are relatively less susceptible to additive background noise and hence may be suitable for speaker recognition application.

HNM assumes the speech signal to be composed of two parts, namely, harmonic part and noise part [21], [22]. The harmonic part accounts for the quasi-periodic components of the speech signal while the noise part accounts for the non-periodic components of the speech signal such as fricatives or aspiration noise, period-to-period variations of the glottal excitation, etc. These two components are separated in the frequency domain by a time varying parameter, referred to as maximum voiced frequency, F_{m}. The lower band of the spectrum, below the maximum voiced frequency, is assumed to be solely represented by the harmonics while the upper band, above the maximum voiced frequency, is represented by a modulated noise component. The synthesized speech signal $\hat{s}(t)$ is given by:

$$\hat{s}(t) = s'(t) + n(t) \tag{3.1}$$

where $s'(t)$ and $n(t)$ represent synthesized harmonic and noise parts. The harmonic part (below F_{m}) is represented by harmonically related sine waves with slowly varying amplitudes and frequencies [22]:

$$s'(t) = \mathrm{Re} \sum_{k=1}^{K(t)} a_k(t) \exp(j[\int_0^t k\omega_0(\sigma)d\sigma + \theta_k]) \tag{3.2}$$

where $\omega_0(t)$ is the fundamental frequency, $a_k(t)$ and $\theta_k(t)$ are the amplitude and phase of k^{th} harmonic, and $K(t)$ is the number of harmonics included in the harmonic part.

The upper band which contains the noise part is modeled by an autoregressive (AR) model, by filtering a white Gaussian noise $g(t)$ by a time varying normalized all-pole filter $h(\tau,t)$ and multiplying the result by an energy envelope function $w(t)$:

$$n(t) = w(t)[h(\tau,t) * g(t)] \tag{3.3}$$

3.3 Estimation of HNM parameters

The analysis using HNM is carried on frame-by-frame basis [21]. The analysis scheme of HNM is shown in Fig. 3.1. In this figure, speech signal is applied to the voicing detector, which declares the frames either voiced or unvoiced. As the analysis and synthesis in HNM is pitch-synchronous, the glottal closure instants (GCIs) need to be estimated precisely. For each voiced frame, the maximum voiced frequency (F_m) is calculated. The analysis frame is taken as twice the local pitch period. This voiced frame of the speech is analyzed at each GCI for calculating amplitudes and phases of all the pitch harmonics up to F_m.

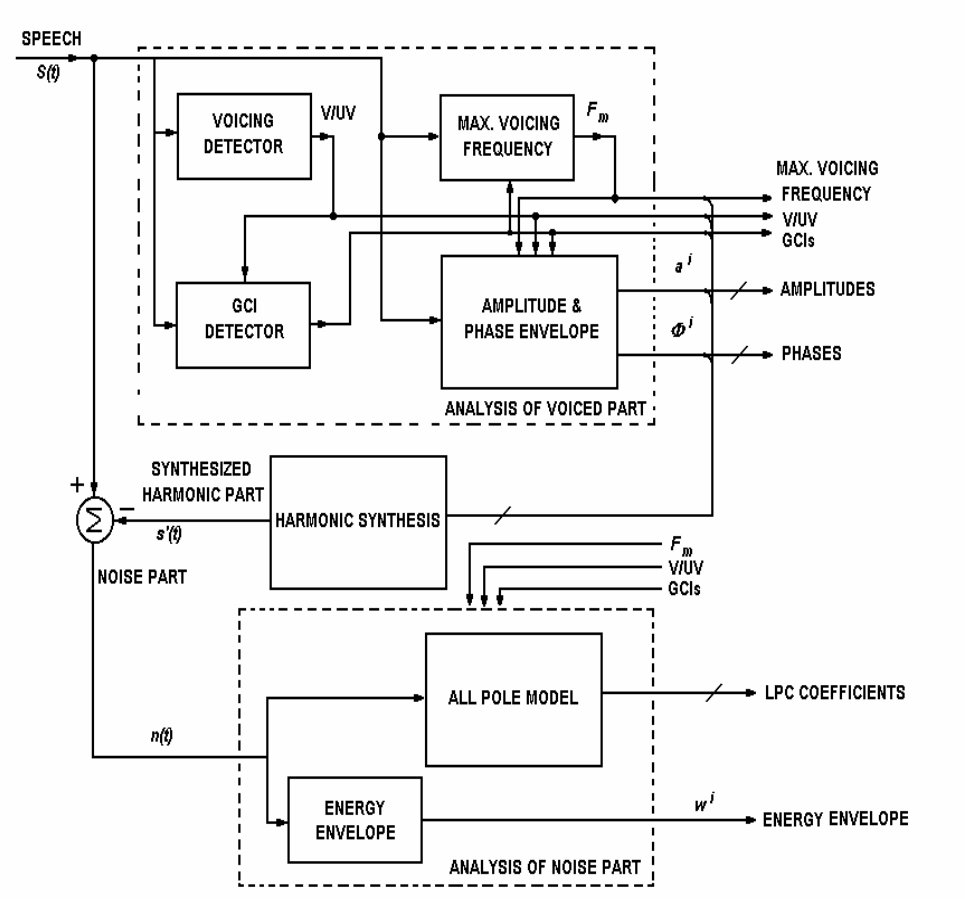

Fig. 3.1 Analysis of speech using HNM, from [22]

For calculating the noise parameters, the synthesized voice part of speech is obtained using Eq.3.2 and noise part is obtained by subtracting this from the original speech. After passing the noise part through a high pass filter having cut off frequency F_m, it is analyzed for obtaining LPC coefficients and energy envelope. The length of the analysis window for noise part is taken as two local pitch periods for both voiced and

unvoiced frames. For voiced frames, the local pitch is the pitch of the frame itself. For unvoiced frames, the pitch of the last voiced frame is taken. The estimation techniques for the parameters related to harmonic part are described in the following subsections. The HNM analysis program used has been earlier developed in our laboratory by Parveen K. Lehana [22].

3.3.1 Voiced/unvoiced decision

The analysis using HNM is carried out on frame-by-frame basis. Hence, the speech signal is first divided into frames of equal length. Then each frame is tested for voiced/unvoiced decision using the method proposed by Childers [25]. In this method, two parameters namely the energy (E) and the first order reflection coefficient (r_1) are used.

We know that the amplitudes of unvoiced speech segments are much lower than the amplitudes of voiced speech segments. Hence, the energy of unvoiced segment is much lower than the energy of voiced segment. Thus, the energy of a speech segment provides a basis for distinguishing voiced speech segments from unvoiced speech segments. For voiced/unvoiced decision, a threshold value (T_e) of 10^7 for input signal range of $\pm 2^{15}$ is used. The first order reflection coefficient (r_1) is defined as the ratio of autocorrelation value of first lag $R_{ss}(1)$ and autocorrelation value of zeroth lag $R_{ss}(0)$. This ratio is significantly high for voiced speech segments than for unvoiced speech segments. Thus, r_1 also gives significant information for distinguishing the voiced and unvoiced frames. The algorithm of this method is described in the following steps.

1. Define 1st order reflection coefficient as:

$$r_1 = \frac{R_{ss}(1)}{R_{ss}(0)} = \frac{\sum_{n=0}^{N-1} s(n)s(n+1)}{\sum_{n=0}^{N-1} s(n)s(n)} \tag{3.4}$$

2. Declare the frame as voiced or unvoiced using the following criterion [4]:

   IF ($r_1 > 0.2$ AND $E > 2T_e$) OR (($r_1 > 0.3$ AND E > T_e) AND Previous frame is voiced)

   {Frame is voiced}

   ELSE

   {Frame is unvoiced}

   END

Finally, the voiced and unvoiced frames are denoted by '1' and '0' respectively. The impractical sequences such as 101 and 010 are replaced by 111 and 000 respectively.

3.3.2 Estimation of glottal closure instants

The analysis and synthesis in HNM are pitch-synchronous, and hence it is necessary to accurately estimate the glottal closure instants (GCIs). Calculation of GCIs involves two steps. In the first step, the pitch periods are estimated. In the second step, these are further refined. Childers and Hu's algorithm is used for these calculations [25]. In this method, the estimation of GCI is based on the concept of LPC residual signal (e_p). GCI is estimated as the location of the first peak in the real cepstrum of the low-pass filtered prediction error (LPC residual). This method of estimation of the pitch is described in the following steps.

1. Low pass e_p to get e_{LP}
2. Calculate $c_e(n)$ = IFFT (|FFT (e_{LP})|), $0 \le n \le N$-1, where N = frame length
3. Find index m for maximum value of $c_e(n)$ in the range $25 \le n \le N$
4. Find index k for maximum value of $c_e(n)$ in the range $25 \le n \le m$-25
5. If $c_e(k) > 0.7c_e(m)$ then k is taken as the pitch period otherwise m is retained

The estimated pitch periods are median filtered to get a smooth pitch period contour. The pitch periods estimated in the first phase are further refined using the peak picking algorithm [25], using following steps.

1. Find most negative peak of e_{LP} in the frame
2. Take 15 samples before and 30 samples after this peak
3. Find the cross correlation of e_{LP} and the segment obtained in step 2
4. Positive peaks of the above sequence are taken as GCIs
5. If the difference between two successive GCIs (obtained in this phase) is less than 25 samples (corresponding to the pitch value of 400 Hz), pitch obtained in the first phase is considered. Otherwise, the GCIs obtained in this phase are considered.

3.3.3 Estimation of maximum voiced frequency

Maximum voiced frequency, F_m, is defined as the boundary between harmonic part and noise part of the speech signal. The harmonic part of the spectrum has a well defined harmonic structure and the part of the spectrum where the harmonic structure is

completely lost is attributed to the noise part. Thus, F_m denotes the frequency of the last harmonic peak (i.e., the last peak declared as voiced), in the spectrum of the speech signal. Identification of harmonic peaks is based on peak magnitude, its position, and area under the log magnitude spectrum segment between the valleys on either side of the peak. The implementation used is essentially the same as used by Stylianou [21], [26]. For identifying the harmonic peaks, we first locate all the peaks. Operations for locating the peaks are carried out on the log magnitude spectrum of each voiced frame. For this purpose, the signal segment in the frame is padded with zero-valued samples to obtain log magnitude spectrum $A(f)$ with a fine sampling of frequency f. The algorithm can be described as the following.

1. Set the starting test voiced frequency $f_c = 0$.
2. Find the highest peak in the frequency range $[f_c+0.5F_0, f_c+1.5F_0]$. The location of this peak is f_c and the magnitude is A_m. Let the peak next to the maximum peak be termed as A_1.
3. Locate all the peaks in the frequency range $[f_c-0.5F_0, f_c+0.5F_0]$. For each of the peak location, calculate the area under the spectral segment between the valleys on either side of the peak. Let the area for the maximum peak be A_c and average of the areas for all the other peaks be $\overline{A_c}$.
4. Declare f_c as voiced if it is within ± 20 % of an integer multiple of F_0 and the harmonic test

$$\left(\frac{A_c}{\overline{A_c}} > 2\right) \text{ or } (A_m - A_1 > 13)$$

   is satisfied.
5. Go to step 2 with the new value of f_c, until the entire band $[0, f_s/2]$ is covered.

The above search gives us a set of peak locations $f_c(m)$ and each is declared as voiced (1) or unvoiced (0). In many cases, the voiced regions of the spectrum are not clearly separated from the unvoiced regions [26]. These two regions are separated by filtering the voiced/unvoiced decision binary vector using a three-point median smoothing filter. The last frequency, in the filtered vector, declared as voiced is labeled as F_m.

3.3.4 Estimation of amplitudes and phases of the harmonics

For analyzing the voiced frames, an analysis window of length twice the local pitch period is centered at the GCIs and amplitudes and phases are calculated by minimizing a weighted time domain least-squares criterion [21] with respect to $a_k(t_i)$. The weighted squared error signal (ε) is given by:

$$\varepsilon = \sum_{t=t_i-T_0}^{t=t_i+T_0} w_{\mathrm{H}}^2(t)[s(t)-s'(t)]^2 \tag{3.5}$$

where T_0=local pitch period, $w_{\mathrm{H}}(t)$ = Hanning window, $s(t)$ = original speech, and $s_{\mathrm{h}}(t)$ = harmonic signal to estimate, and t_i is the centre of the frame. The details of the minimization process are given in [21].

3.4. HNM parameters for speaker recognition

The advantage of the HNM scheme for speaker recognition is the representation of the speech signal as sum of harmonic part and noise part. Since, the effect of contaminations such as background noise on estimation of harmonic magnitudes of the speech signal will be less, the parameters related to harmonic part may be useful for speaker recognition. Among the HNM parameters discussed earlier, maximum voiced frequency and pitch may be useful for speaker recognition.

Maximum voiced frequency (F_m) distinguishes the harmonic part of the speech signal from the noise part in the frequency domain. This means, F_m is the boundary between the harmonic part and the noise part in frequency domain. In general, this boundary will vary from speaker to speaker due to the variations in size and shape of the vocal tract. Thus, maximum voiced frequency is associated with vocal tract and to a certain extent with the excitation and hence contains speaker information that may be useful for speaker recognition. A new parameter, α, defined as the ratio of energy of spectral segment above F_m and total energy, may also give speaker dependent information. Pitch is another parameter useful for speaker recognition. Pitch varies with the size of the vocal chords and hence it also contains speaker information.

The random variations in the parameters F_m and F_0 are speaker dependent and hence may be useful for speaker recognition. These random variations are called as jitter in a parameter. Jitter in a parameter, p, for the n^{th} frame is defined as:

$$J(n) = 100 \frac{|p(n+1) - p(n)|}{p(n)} \tag{3.6}$$

Jitter in pitch (ΔF_0) and jitter in maximum voiced frequency (ΔF_m) can be obtained by using (3.6), where p is replaced by pitch (F_0) and maximum voiced frequency (F_m) respectively. The random variations in sound intensity (A_0) are also speaker dependent. These random variations are called as shimmer in intensity (ΔA_0). Shimmer in intensity for n^{th} frame is defined as:

$$\Delta A_0 = 100 \frac{|A_0(n+1) - A_0(n)|}{A_0(n)} \tag{3.7}$$

The harmonic amplitudes and phases contain basically the same information as provided by LPCC or MFCC, and hence these are not investigated at this stage. The study of variations of the HNM parameters (F_m, F_0, α, ΔF_0, ΔF_m, and ΔA_0) through histogram analysis, moments, and correlation is described in Chapter 4. It is to be noted that all these parameters relate only to the voiced frames. In the present study, we are not using parameters from unvoiced frames.

Chapter 4

STUDY OF PARAMETERS FOR SPEAKER RECOGNITION

For a parameter to be useful for speaker recognition, its inter-speaker variation should be much more than the intra-speaker variation [1], [12]. In order to observe these variations, the study of selected parameters needs to be carried out through statistical methods. Some of the techniques described in this chapter are histogram analysis, statistical moments, and correlation. The results obtained using these techniques are also discussed.

4.1 Histogram analysis

The variation of a parameter can be studied through frequency distribution [27], as a compact representation of the data. In frequency distribution, the frequency of distinct values of the data will be counted. Hence, this technique is suitable for the parameters containing small number of distinct values. But, in our case, each parameter varies over a large continuous range and hence grouped frequency distribution technique [27] can be used. In grouped frequency distribution, the values of a parameter are grouped into classes (also called bins) and the frequency of each class will be counted. The graphical form of this type of distribution is called histogram plot [27], [28], which gives a picture of the general shape of the distribution.

The number of bins will affect the shape of the histogram and should be chosen carefully while obtaining the histograms. The variation in a parameter can be observed clearly with more number of bins. But, this will cause more random variations in the shape of the histogram and hence more deviation from the assumed distributional model.

The effect of variations between the successive samples of a variable can be studied through multi-sample histogram analysis [28]. One-sample histogram analysis is same as the simple histogram analysis. In case of two-sample histogram analysis, the present sample will be accepted into the corresponding bin only if the difference between the present sample and the previous sample is less than half of the bin width. In case of three-sample histogram analysis, the present sample will be accepted into the

corresponding bin only if the difference between the present sample and the previous sample as well as the difference between the present sample and the one before previous sample is less than half of the bin width. The same method can be further extended to four or higher sample histograms.

4.2 Statistical moments

Histogram plot provides a graphical description of variation in a parameter. Statistical moments, namely, mean, RMS deviation about mean, skewness, and kurtosis [27] can be used for quantifying the shape of the histogram. These moments can be obtained directly from the N samples of a variable x as follows:

- Mean value ($\bar{x}$ or M_1):

$$M_1 = \bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{4.1}$$

- RMS deviation about mean (M_2 or σ_x):

$$M_2 = \sigma_x = [\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2]^{1/2} \tag{4.2}$$

- Skewness (M_3) :

$$M_3 = [\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^3] / \sigma_x^{\,3} \tag{4.3}$$

- Kurtosis (M_4):

$$M_4 = [\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^4] / \sigma_x^{\,4} \tag{4.4}$$

Skewness, given by (4.3), is a measure of the lack of symmetry of the data [28]. For normal distribution, skewness is zero and for any symmetric data, skewness should be nearly zero. Negative skewness indicates that the data are skewed left i.e., the left tail in the histogram will be longer than the right tail. Similarly, positive skewness indicates that the data are skewed right i.e., the right tail in the histogram will be longer than the left tail.

Kurtosis, given by (4.4), is a measure of the peakiness of the distribution relative to a normal distribution [28]. For a standard normal distribution, with zero mean and unit

standard deviation, kurtosis is three. The data with high kurtosis (>3) will have a strong peak near the mean, more rapid decay and heavier tails in the histogram.

Significant skewness and kurtosis indicate that the data are not normally distributed. This means, skewness and kurtosis are the measures of the amount of deviation of the data from the normal distribution. The above moments are used to study the pattern of variations in speech parameters across speakers, to investigate their possible use in speaker recognition.

4.3 Correlation coefficient

Another parameter to measure the closeness of a parameter to an assumed distributional model is correlation coefficient (*r*) and is defined as follows [27]:

$$r = \frac{\sum_{i=1}^{K}[(O_i - \overline{O}).(E_i - \overline{E})]}{\sqrt{\sum_{i=1}^{K}(O_i - \overline{O})^2 . \sum_{i=1}^{K}(E_i - \overline{E})^2}} \tag{4.5}$$

where $\overline{O}$ and $\overline{E}$ are the mean values for the observed and estimated distributions respectively. Correlation coefficient (*r*) is a pure number without units or dimensions and it always lies between -1 and +1. Positive values of *r* indicate a tendency of the two variables (*O* and *E*) to increase together and negative values of *r* indicate that large values of one variable are associated with small values of the other variable. The value $r = 0$ indicates that there is no correlation between the two variables. The parameter is said to have an assumed distribution if the value of *r* is close to unity. The distribution of various analysis parameters showed that some of them could be approximated by an exponential distribution while others could be approximated by normal distribution. For exponential distribution, the approximation is given as:

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & , x \geq 0 \\ 0 & , x < 0 \end{cases} \tag{4.6}$$

with $\lambda = 1/\overline{x}$ obtained from the observed values. For normal distribution, the approximation is given as:

$$f(x) = \frac{1}{\sigma_x \sqrt{2\pi}} e^{-(x-\overline{x})^2/(2\sigma_x^2)} \tag{4.7}$$

with $\overline{x}$ and σ_x obtained from the observed values. In both cases, the E_i values were calculated for the corresponding bin by integrating the function over the bin.

4.4 Analysis results

For observing the variations in the histograms and statistical parameters (discussed above) of the speech analysis parameters, namely, peak amplitudes, F_m, F_0, α, ΔF_0, ΔF_m, and ΔA_0, different speech recordings were made at the sampling rate of 10 kSa/s, for 10 male speakers labeled as Speaker-1 to Speaker-10. As the statistical parameters may vary for the same speaker in different utterances of the same phrase (due to various factors such as background noise, variations in handling the microphone and most importantly, the linguistic information in the utterance), the same utterance was recorded five times for each speaker. Then, the histograms and statistical moments were obtained for these utterances. From this, the inter-speaker and intra-speaker variations were calculated. Finally, F-ratio test was carried out on each statistical parameter. The results for the speech analysis parameters (as described in Chapter 3), using histograms and statistical parameters, are presented in the following subsections.

4.4.1 Histogram analysis on peak amplitudes

For analyzing the peak amplitudes using histograms, the speech signal "where were you a year ago" was recorded five times for each of the 10 speakers. The recorded signal was normalized such that the magnitude of the highest peak was scaled to unity. For this normalized signal, peak amplitudes were calculated. Then, multi-sample histograms for the peak amplitudes were obtained.

One-sample histograms of peak amplitudes are shown in Fig 4.1, with one plot for each of the speaker, with the plots of different utterances superimposed together. The *y*-axis represents the frequency in percentage and *x*-axis represents the normalized amplitude, divided into 60 bins (an empirically selected number). The variations within the histogram for different utterances by a single speaker are observed in Fig 4.1. There are clear differences in the plots across speakers. However, differences across the speakers do not appear much larger than the differences across utterances.

The variations in the histogram plots are characterized through statistical parameters. The statistical parameters, namely, moments and correlation coefficients were calculated for peak amplitudes of each normalized signal. Correlation coefficients were calculated by assuming exponential distribution. The mean and standard deviation of each statistical parameter for 10 speakers are given in Table 4.1. It is observed from the table that the inter-speaker variations of each statistical parameter are of the same order as

intra-speaker variations. But, for the application of speaker recognition, the inter-speaker variations should be significantly higher than the intra-speaker variations. F-ratio test shows that the variation across speakers is statistically highly significant ($p < 0.005$). However, F-ratio values in this case are not as large as for HNM parameters for which results are described later. This indicates that although peak amplitude may be indicative of a speaker, their usefulness in speaker recognition is much lower than that of HNM parameters.

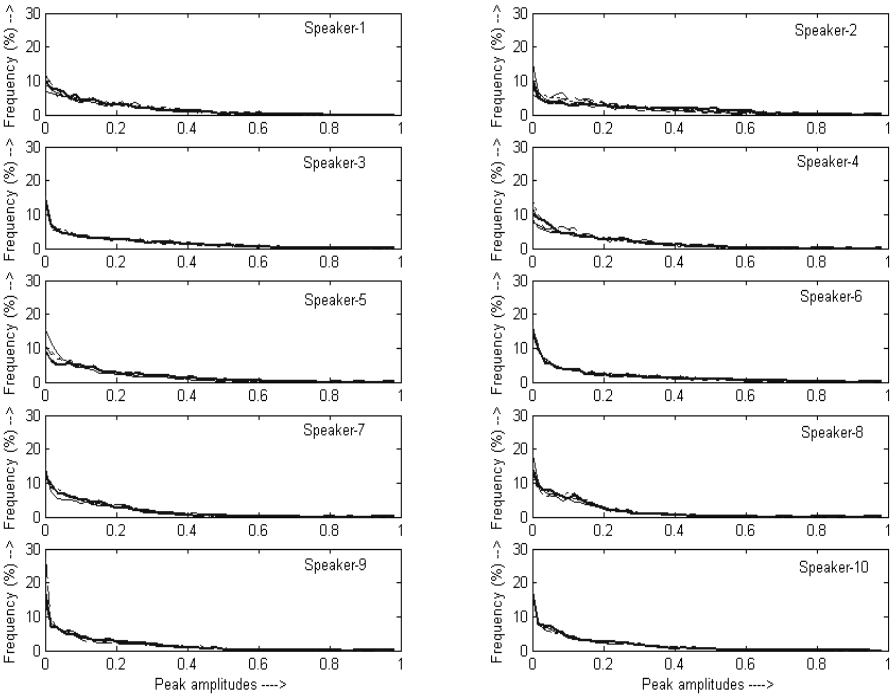

Fig 4.1 One-sample histograms of peak amplitude

The variations in M_1 of peak amplitudes for 10 speakers are shown in Fig 4.2. In this figure, each vertical line represents the variation of M_1 for a particular speaker, center of the line represents mean value of M_1, and length of the line represents standard deviation of M_1 within the speaker for different utterances. It is observed that the speakers are not well separated. The results of two-sample and three-sample histogram analyses are given in Appendix A. They exhibit the same pattern as one-sample histograms.

Table 4.1 Mean and standard deviation of moments for peak amplitude. The correlation coefficient *r* is with respect to exponential distribution. No. of utterances = 5, No. of speakers = 10. Mn = mean, Sd = std. dev., MSA = mean square among speakers, MSW = mean square within speakers. D.F. for F-ratio test: 9, 40. $(F)_{0.005} = 3.22$, $(F)_{0.01} = 2.89$.

Spk. No.	M_1		M_2		M_3		M_4		M_2/M_1		M_4/M_3		*r*	
	Mn	**Sd**	**Mn**	**Sd**	**Mn**	**Sd**	**Mn**	**Sd**	**Mn**	**Sd**	**Mn**	**Sd**	**Mn**	**Sd**
1	0.19	0.02	0.17	0.02	1.29	0.15	4.73	0.71	0.89	0.04	3.66	0.17	0.98	0.01
2	0.23	0.03	0.20	0.02	1.06	0.29	3.75	0.91	0.86	0.05	3.58	0.27	0.93	0.06
3	0.21	0.01	0.20	0.01	1.12	0.16	3.78	0.67	0.94	0.03	3.36	0.13	0.90	0.01
4	0.18	0.02	0.16	0.01	1.35	0.16	5.02	0.62	0.90	0.05	3.72	0.11	0.98	0.01
5	0.18	0.03	0.17	0.01	1.41	0.23	4.94	0.92	0.96	0.09	3.50	0.11	0.98	0.01
6	0.19	0.01	0.19	0.01	1.28	0.13	4.17	0.55	1.02	0.03	3.25	0.11	0.91	0.01
7	0.16	0.01	0.15	0.01	1.69	0.17	6.95	1.11	0.95	0.01	4.09	0.27	0.98	0.02
8	0.14	0.01	0.13	0.01	2.00	0.17	8.86	1.21	0.99	0.05	4.43	0.32	0.97	0.02
9	0.17	0.02	0.17	0.01	1.48	0.13	5.30	0.68	1.03	0.08	3.57	0.18	0.92	0.07
10	0.17	0.01	0.17	0.01	1.47	0.12	5.26	0.62	1.02	0.02	3.57	0.15	0.92	0.01
Mn	**0.18**		**0.17**		**1.41**		**5.27**		**0.95**		**3.67**		**0.95**	
Sd	**0.03**		**0.02**		**0.27**		**1.56**		**0.06**		**0.35**		**0.03**	
MSA	**0.0035**		**0.0021**		**0.3726**		**12.171**		**0.0167**		**0.6038**		**0.0057**	
MSW	**0.0003**		**0.0001**		**0.0313**		**0.6845**		**0.0025**		**0.0392**		**0.0009**	
F	**10.6**		**15.1**		**11.9**		**17.8**		**6.70**		**15.4**		**6.15**	
p	**0.005**		**0.005**		**0.005**		**0.005**		**0.005**		**0.005**		**0.005**	

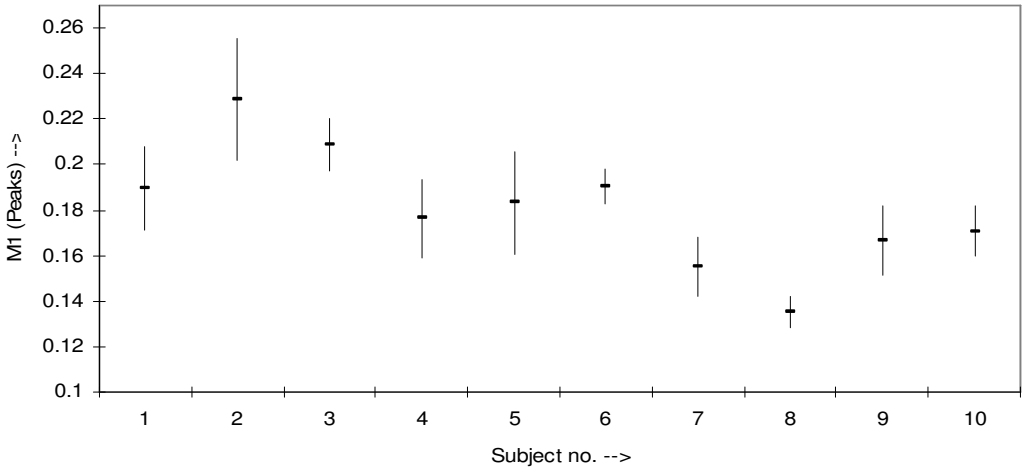

Fig 4.2 Variation of M_1 of peak amplitude for 10 speakers

Hence, peak amplitudes of the speech signal may not be useful for speaker recognition. The sensitiveness of peak amplitudes to various factors such as background noise, variations in handling the microphone, etc., may be causing the difficulty with these parameters for speaker recognition.

4.4.2 Histogram analysis on HNM parameters

For observing the variations in HNM parameters (F_m, F_0, α, ΔF_0, ΔF_m, and ΔA_0) using histograms, the English sentences given in Table 4.2 were recorded five times for each of the 10 speakers. These sentences have been specifically formed to emphasize certain speech features [29].

Histogram analysis was carried out on the concatenation of the 8 sentences given in Table 4.2. The concatenated speech signal was divided into frames of equal length and each frame was analyzed to get the parameters, F_m, F_0, and α. Then, ΔF_0, ΔF_m, and ΔA_0 were calculated. Finally, multi-sample histograms for these parameters were obtained.

Table 4.2 Sentences recorded for the study of HNM parameters

S. No.	Sentence type	Sentence or phrase
1.	Fricatives/Whispers/Affricatives	Bright sunshine shimmers on the ocean.
		His vicious father has seizures.
2.	Unvoiced Consonants	Get a calico cat to keep away.
		Primitive tribes have an upbeat attitude.
3.	Voiced Consonants	Did dad do academic bidding?
		Doctors prescribe drugs too freely.
4.	Predominantly Voiced (Vowels/Diphthongs/Semivowels)	He will allow a rare lie.
		Will Robin wear a yellow lily?

One-sample histograms of F_m, F_0, α, ΔF_0, ΔF_m, and ΔA_0 are shown in Fig 4.3 to 4.8 respectively. In each case, there is one plot for each of the 10 speakers, with the plots corresponding to different recordings superimposed together. The *y*-axis of the plots represents the frequency in percentage; with *x*- axis representing the analysis parameters, with the range divided into 60 bins. For F_m, F_0, and α, the variations within the histogram for different utterances by a single speaker can be observed in Fig 4.3 to 4.5. But, these are small compared to variations across speakers. Each speaker appears to have a different characteristic in the plot. For ΔF_0, ΔF_m, and ΔA_0, the variations across speakers are not as large as for F_m, F_0, and α.

The variations in the histogram plots are characterized through statistical parameters. The statistical parameters, namely, moments and correlation coefficients were calculated for the HNM parameters (F_m, F_0, α, ΔF_0, ΔF_m, and ΔA_0). Correlation

coefficients were calculated by assuming normal distribution for F_m, F_0, and α, while exponential distribution for ΔF_0, ΔF_m, and ΔA_0. The mean and standard deviation of each statistical parameter for 10 speakers are given in Tables 4.3 to 4.8. For F_m, F_0, and α, it can be observed that the inter-speaker variations of each statistical parameter are significantly higher than the intra-speaker variations. F-ratio test shows that the variation across speakers is statistically highly significant ($p < 0.005$). Also, the F-ratio values (particularly for M_1) are very large compared to that of peak amplitude. This indicates that the parameters F_m, F_0, and α are good indicative of a speaker and hence may be useful for speaker recognition. For ΔF_0, ΔF_m, and ΔA_0, the inter-speaker variations are not as large as for F_m, F_0, and α. Also, the values of F-ratio are not as large as for F_m, F_0, and α.

The variations in M_1 of HNM parameters for 10 speakers are shown in Fig 4.9 to 4.14. In these figures, each vertical line represents the variation of M_1 for a particular speaker, center of the line represents mean value of M_1, and length of the line represents standard deviation of M_1 within the speaker for different utterances. It is observed that the characteristics of speakers are well separated. Most importantly, α has shown good indications for speaker recognition when compared to other parameters, through very high inter-speaker variations compared to intra-speaker variations. This is also confirmed by very high F-ratio value and a good separation of speakers shown in the plot in Fig 4.11. The results of two-sample and three-sample histogram analyses for F_m, F_0, and α are given in Appendix A. They show the same pattern as the one-sample histograms.

Hence, HNM parameters (F_m, F_0, α, ΔF_0, ΔF_m, and ΔA_0) may be useful for speaker recognition. The speaker recognition tests with HNM parameters are carried out in Chapter 5 using vector quantization (VQ) technique.

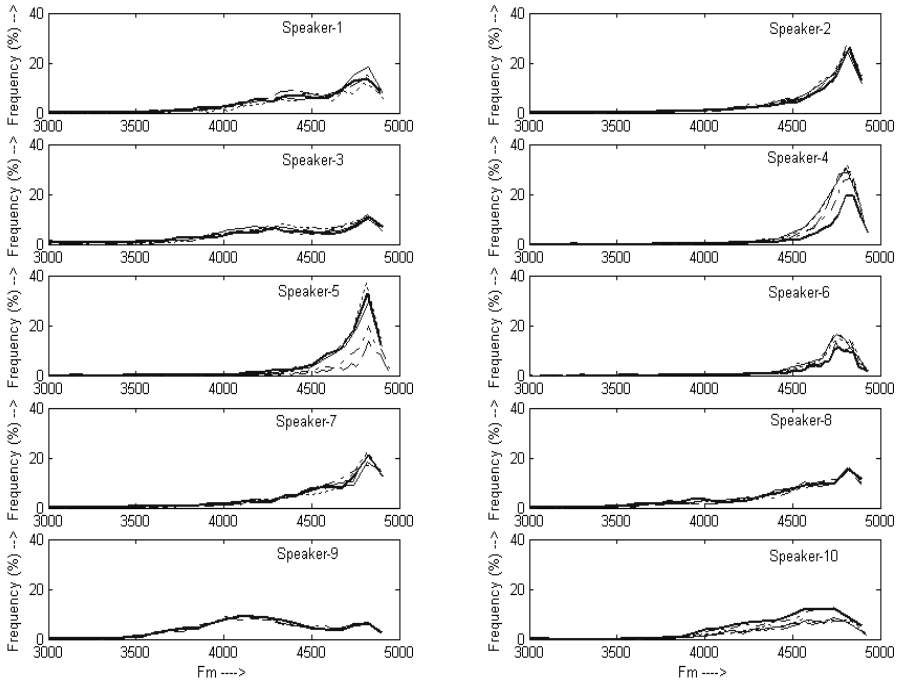

Fig 4.3 One-sample histograms of F_m (Hz)

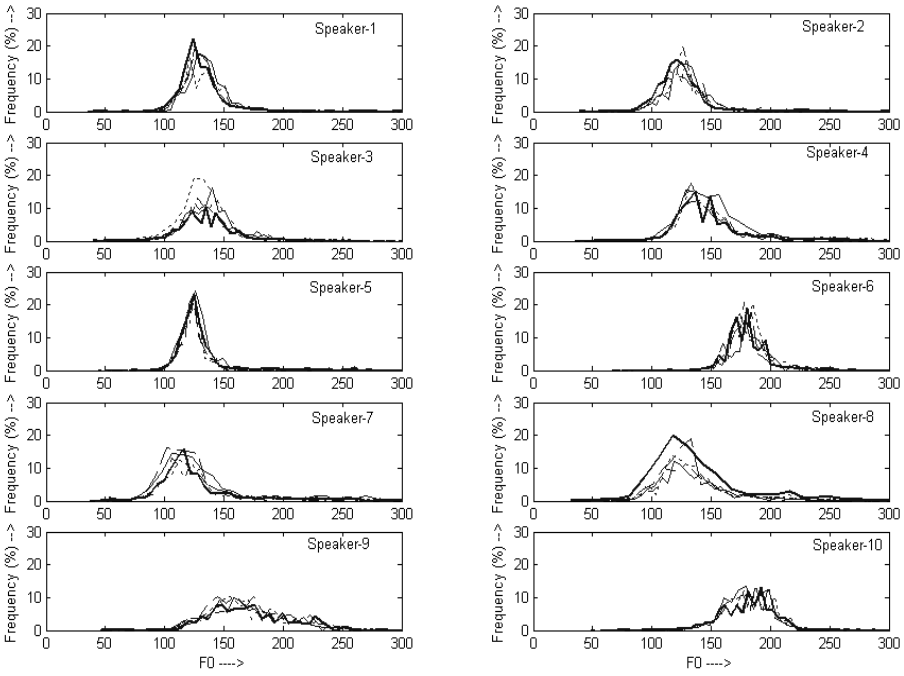

Fig 4.4 One-sample histograms of F_0 (Hz)

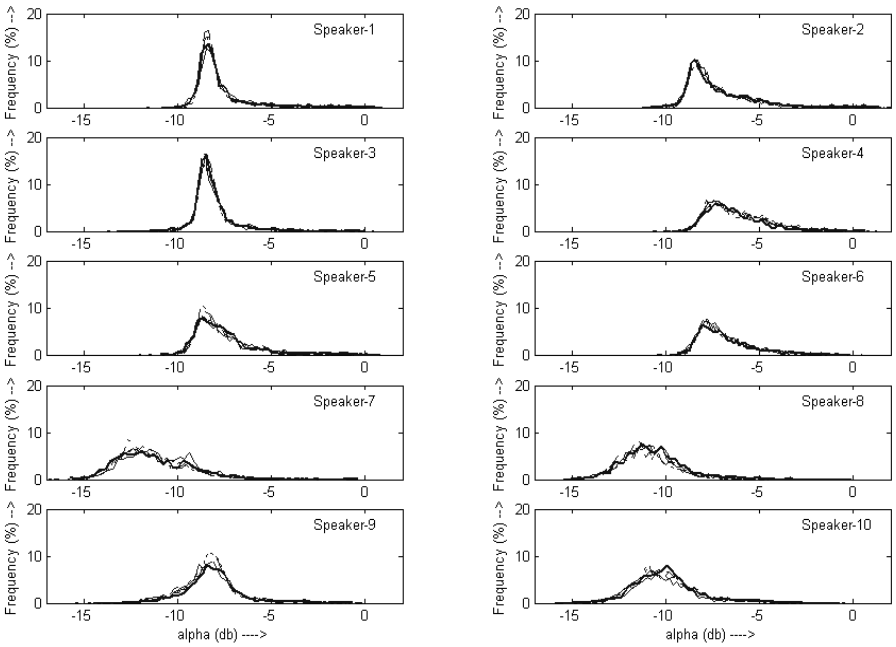

Fig 4.5 One-sample histograms of α (dB)

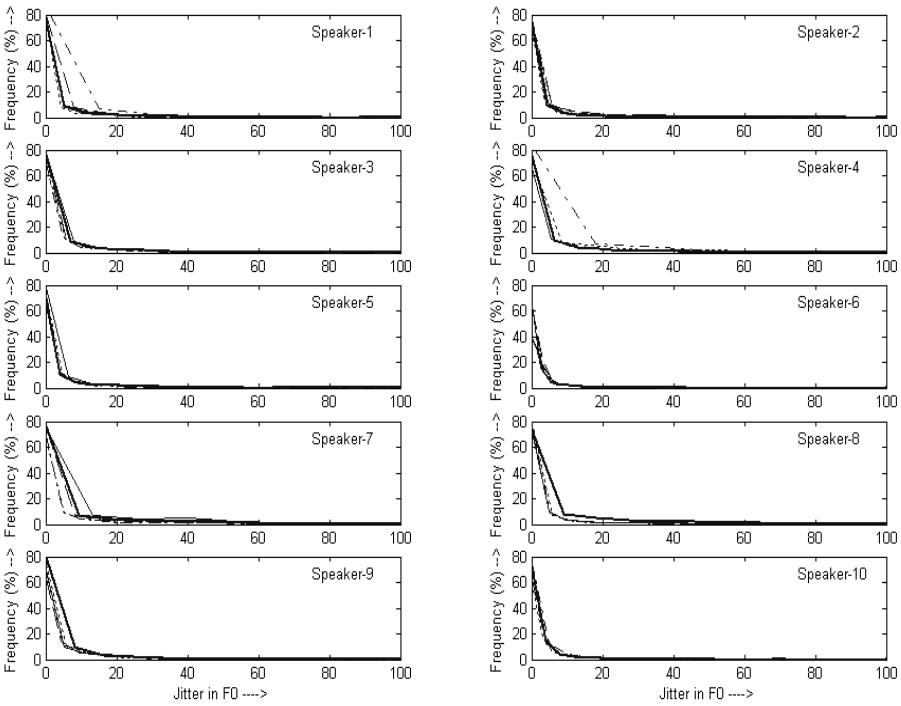

Fig 4.6 One-sample histograms of ΔF_0

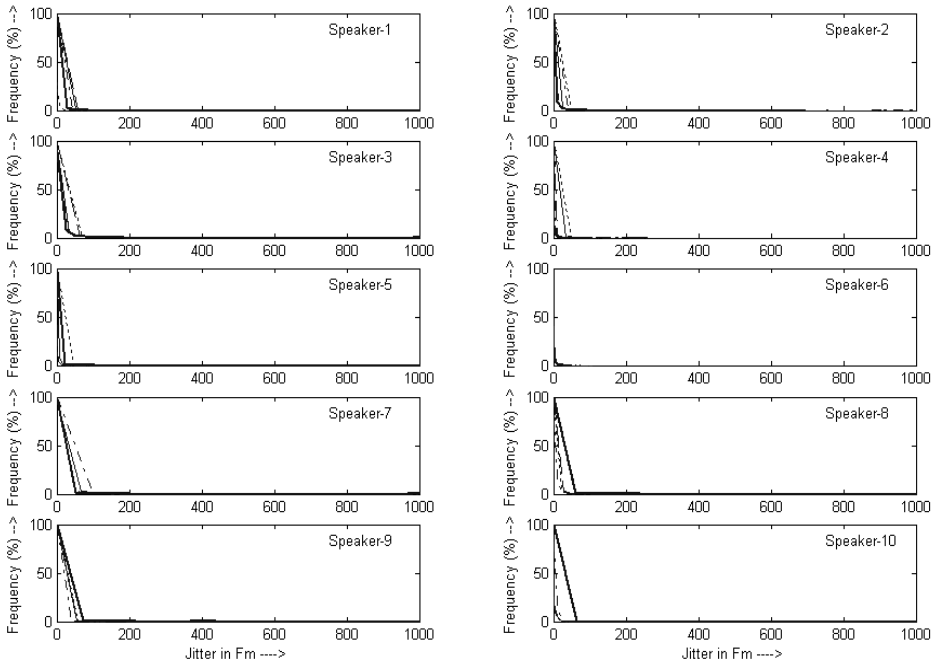

Fig 4.7 One-sample histograms of ΔF_m

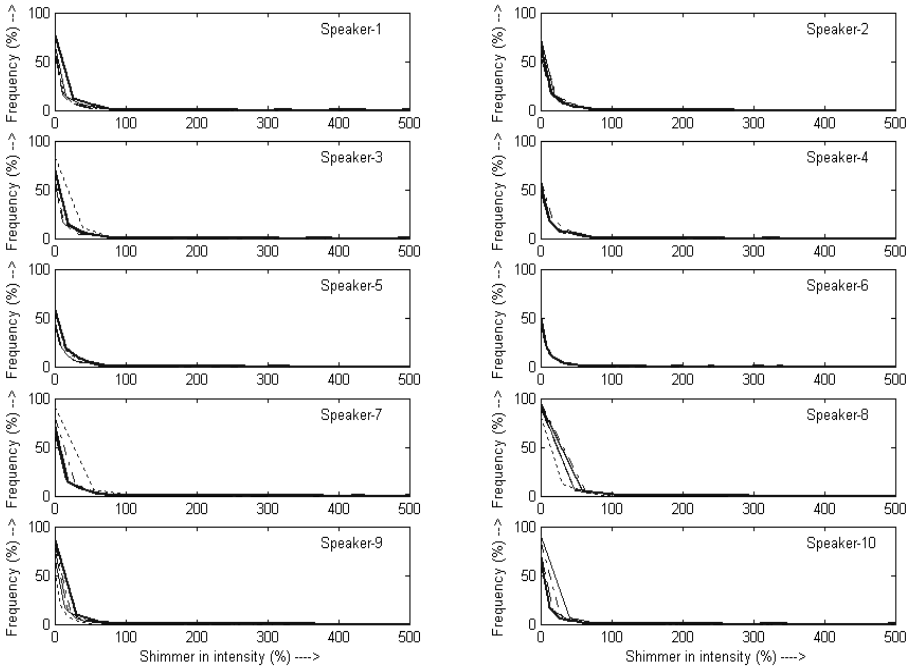

Fig 4.8 One-sample histograms of ΔA_0

Table 4.3 Mean and standard deviation of moments for F_m (Hz). The correlation coefficient r is with respect to normal distribution. No. of utterances = 5, No. of speakers = 10. Mn = mean, Sd = std. dev., MSA = mean square among speakers, MSW = mean square within speakers. D.F. for F-ratio test: 9, 40. $(F)_{0.005} = 3.22$, $(F)_{0.01} = 2.89$.

Spk. No.	M_1		M_2		M_3		M_4		M_2/M_1		M_4/M_3		r	
	Mn	**Sd**	**Mn**	**Sd**	**Mn**	**Sd**	**Mn**	**Sd**	**Mn**	**Sd**	**Mn**	**Sd**	**Mn**	**Sd**
1	4493	33.7	452	48.7	-2.9	0.54	18.4	5.60	0.10	0.01	-6.3	1.00	0.85	0.03
2	4621	18.6	430	36.4	-3.7	0.62	24.9	7.39	0.09	0.01	-6.6	1.00	0.73	0.03
3	4242	53.7	696	43.6	-2.1	0.38	9.47	2.52	0.16	0.01	-4.4	0.41	0.79	0.03
4	4744	13.0	259	33.9	-6.7	1.62	75.9	28.8	0.05	0.01	-11	1.94	0.81	0.01
5	4733	23.5	261	83.0	-5.5	2.75	66.8	62.3	0.06	0.02	-10	5.41	0.78	0.05
6	4694	6.44	223	6.19	-2.9	0.40	16.7	5.27	0.05	0.00	-5.7	1.10	0.81	0.03
7	4494	62.6	705	105	-4.0	0.50	22.1	5.74	0.16	0.03	-5.5	0.79	0.68	0.05
8	4491	39.2	505	104	-3.2	0.69	21.3	4.22	0.11	0.02	-6.6	0.58	0.81	0.03
9	4239	9.51	462	26.6	-2.3	0.47	17.9	3.64	0.11	0.01	-7.9	0.27	0.96	0.00
10	4523	8.05	312	31.1	-2.3	1.46	20.3	19.4	0.07	0.01	-7.4	2.47	0.94	0.01
Mn	**4527**		**430**		**-3.56**		**29.4**		**0.10**		**-7.15**		**0.82**	
Sd	**181**		**173**		**1.51**		**22.6**		**0.04**		**2.02**		**0.08**	
MSA	**16.3E+04**		**14.8E+04**		**11.327**		**2549.3**		**0.0085**		**20.438**		**0.0350**	
MSW	**1073.2**		**3713.3**		**1.4226**		**527.04**		**0.0002**		**4.3517**		**0.0009**	
F	**152**		**40.1**		**7.96**		**4.84**		**41.5**		**4.70**		**40.2**	
p	**0.005**		**0.005**		**0.005**		**0.005**		**0.005**		**0.005**		**0.005**	

Table 4.4 Mean and standard deviation of moments for F_0 (Hz). The correlation coefficient r is with respect to normal distribution. No. of utterances = 5, No. of speakers = 10. Mn = mean, Sd = std. dev., MSA = mean square among speakers, MSW = mean square within speakers. D.F. for F-ratio test: 9, 40. $(F)_{0.005} = 3.22$, $(F)_{0.01} = 2.89$.

Spk. No.	M_1		M_2		M_3		M_4		M_2/M_1		M_4/M_3		r	
	Mn	**Sd**	**Mn**	**Sd**	**Mn**	**Sd**	**Mn**	**Sd**	**Mn**	**Sd**	**Mn**	**Sd**	**Mn**	**Sd**
1	136	1.86	24.7	2.24	3.17	0.53	25.2	4.88	0.18	0.02	7.96	0.78	0.87	0.04
2	131	3.93	26.2	5.56	2.48	0.74	17.6	6.60	0.20	0.04	7.06	2.14	0.87	0.05
3	140	2.72	32.2	8.92	2.04	1.32	14.0	7.87	0.23	0.06	7.18	0.71	0.86	0.07
4	156	3.62	42.0	4.62	2.60	0.21	14.4	3.18	0.27	0.02	5.51	0.77	0.79	0.06
5	131	3.91	23.7	8.31	3.70	0.88	26.1	9.76	0.18	0.06	6.93	1.24	0.77	0.09
6	184	2.43	19.9	1.63	1.14	0.71	12.5	3.79	0.11	0.01	13.9	6.41	0.88	0.03
7	136	5.96	50.1	8.97	2.40	0.41	10.8	3.71	0.37	0.05	4.42	0.71	0.71	0.03
8	144	5.41	46.0	12.2	2.72	0.48	14.8	3.91	0.32	0.07	5.39	0.57	0.81	0.05
9	174	1.45	37.4	1.13	0.97	0.24	6.41	1.63	0.21	0.01	6.59	0.55	0.92	0.01
10	185	2.58	27.6	3.91	0.76	0.88	10.2	1.80	0.15	0.02	-2.8	42.0	0.91	0.03
Mean	**152**		**33.0**		**2.20**		**15.2**		**0.22**		**6.22**		**0.84**	
Sd	**21.6**		**10.4**		**0.97**		**6.29**		**0.08**		**4.08**		**0.07**	
MSA	**2339.1**		**536.28**		**4.7084**		**197.69**		**0.0303**		**83.078**		**0.0230**	
MSW	**13.407**		**45.494**		**0.5123**		**28.410**		**0.0017**		**181.05**		**0.0025**	
F	**174**		**11.8**		**9.19**		**6.96**		**17.8**		**0.46**		**9.13**	
p	**0.005**		**0.005**		**0.005**		**0.005**		**0.005**		**-**		**0.005**	

Table 4.5 Mean and standard deviation of moments for α (dB). The correlation coefficient r is with respect to normal distribution. No. of utterances = 5, No. of speakers = 10. Mn = mean, Sd = std. dev., MSA = mean square among speakers, MSW = mean square within speakers. D.F. for F-ratio test: 9, 40. $(F)_{0.005} = 3.22$, $(F)_{0.01} = 2.89$.

Spk. No.	M_1		M_2		M_3		M_4		M_2/M_1		M_4/M_3		r	
	Mn	**Sd**	**Mn**	**Sd**	**Mn**	**Sd**	**Mn**	**Sd**	**Mn**	**Sd**	**Mn**	**Sd**	**Mn**	**Sd**
1	-7.6	0.08	1.67	0.08	2.45	0.19	9.50	1.08	-0.2	0.01	3.86	0.17	0.66	0.03
2	-7.0	0.07	1.95	0.09	1.85	0.09	7.07	0.61	-0.3	0.02	3.82	0.18	0.74	0.03
3	-7.9	0.07	1.44	0.11	2.49	0.08	11.7	1.01	-0.2	0.02	4.71	0.36	0.74	0.03
4	-6.1	0.09	1.70	0.04	1.07	0.17	4.20	0.59	-0.3	0.01	3.96	0.19	0.87	0.02
5	-7.4	0.09	1.77	0.11	1.79	0.09	6.79	0.61	-0.2	0.02	3.80	0.22	0.80	0.04
6	-6.8	0.05	1.47	0.04	1.32	0.03	4.91	0.19	-0.2	0.01	3.73	0.12	0.84	0.02
7	-11.0	0.17	2.14	0.08	1.23	0.11	5.82	0.34	-0.2	0.01	4.75	0.34	0.92	0.03
8	-11.0	0.18	1.71	0.12	1.11	0.16	6.16	0.96	-0.2	0.01	5.60	0.65	0.95	0.01
9	-8.1	0.08	1.69	0.05	0.71	0.14	5.84	0.32	-0.2	0.01	8.55	1.86	0.91	0.02
10	-9.9	0.08	1.81	0.09	0.95	0.13	5.01	0.45	-0.2	0.01	5.29	0.29	0.95	0.01
Mean	**-8.3**		**1.73**		**1.50**		**6.70**		**-0.2**		**4.81**		**0.84**	
Sd	**1.69**		**0.21**		**0.62**		**2.29**		**0.04**		**1.48**		**0.10**	
MSA	**14.226**		**0.2118**		**1.9273**		**26.212**		**0.0077**		**10.907**		**0.0498**	
MSW	**0.0106**		**0.0075**		**0.0166**		**0.4672**		**0.0001**		**0.438**		**0.0007**	
F	**1340.0**		**28.4**		**116.0**		**56.1**		**51.9**		**24.9**		**76.0**	
p	**0.005**		**0.005**		**0.005**		**0.005**		**0.005**		**0.005**		**0.005**	

Table 4.6 Mean and standard deviation of moments for ΔF_0. The correlation coefficient r is with respect to exponential distribution. No. of utterances = 5, No. of speakers = 10. Mn = mean, Sd = std. dev., MSA = mean square among speakers, MSW = mean square within speakers. D.F. for F-ratio test: 9, 40. $(F)_{0.005} = 3.22$, $(F)_{0.01} = 2.89$.

Spk. No.	M_1		M_2		M_3		M_4		M_2/M_1		M_4/M_3		r	
	Mn	**Sd**	**Mn**	**Sd**	**Mn**	**Sd**	**Mn**	**Sd**	**Mn**	**Sd**	**Mn**	**Sd**	**Mn**	**Sd**
1	8.04	0.60	23.8	4.59	9.77	4.41	162	163	2.95	0.40	14.1	6.86	0.94	0.04
2	7.58	1.32	19.2	3.00	7.06	1.00	76.9	24.2	2.54	0.19	10.7	2.05	0.92	0.03
3	9.58	0.66	24.4	1.35	7.23	1.06	79.9	27.8	2.55	0.13	10.8	2.16	0.92	0.02
4	11.9	1.53	30.2	5.76	7.96	4.02	122	136	2.54	0.34	12.6	6.66	0.92	0.04
5	7.16	1.28	18.2	3.62	7.49	1.58	87.4	36.9	2.53	0.22	11.2	2.67	0.92	0.03
6	4.60	0.26	10.3	1.32	7.09	0.89	75.4	16.8	2.23	0.18	10.5	1.18	0.96	0.01
7	14.3	2.13	35.2	7.30	5.94	1.50	57.6	29.9	2.45	0.21	9.18	2.65	0.88	0.05
8	9.99	2.74	25.0	5.87	5.98	0.41	54.3	8.16	2.52	0.14	9.04	0.79	0.90	0.01
9	8.20	0.78	19.0	1.80	7.71	2.23	104	68.2	2.32	0.16	12.5	4.16	0.94	0.03
10	6.07	0.27	15.3	0.93	7.37	1.17	77.9	24.6	2.51	0.15	10.4	1.63	0.95	0.01
Mean	**8.75**		**22.0**		**7.36**		**89.6**		**2.51**		**11.1**		**0.92**	
Sd	**2.83**		**7.26**		**1.08**		**32.1**		**0.19**		**1.57**		**0.02**	
MSA	**40.170**		**263.70**		**5.7824**		**5166.5**		**0.1759**		**12.335**		**0.0029**	
MSW	**1.9324**		**17.176**		**4.9786**		**5418.0**		**0.0522**		**13.642**		**0.0010**	
F	**20.8**		**15.4**		**1.16**		**0.95**		**3.37**		**0.90**		**2.99**	
p	**0.005**		**0.005**		**-**		**-**		**0.005**		**-**		**0.01**	

Table 4.7 Mean and standard deviation of moments for ΔF_m. The correlation coefficient r is with respect to exponential distribution. No. of utterances = 5, No. of speakers = 10. Mn = mean, Sd = std. dev., MSA = mean square among speakers, MSW = mean square within speakers. D.F. for F-ratio test: 9, 40. $(F)_{0.005} = 3.22$, $(F)_{0.01} = 2.89$.

Spk. No.	M_1		M_2		M_3		M_4		M_2/M_1		M_4/M_3		r	
	Mn	**Sd**	**Mn**	**Sd**	**Mn**	**Sd**	**Mn**	**Sd**	**Mn**	**Sd**	**Mn**	**Sd**	**Mn**	**Sd**
1	9.65	2.22	72.9	42.9	23.1	10.8	670	399	7.00	3.56	25.1	10.4	1.00	0.01
2	7.89	1.48	51.5	31.0	27.5	5.69	875	317	6.28	3.17	31.1	4.99	1.00	0.00
3	17.8	4.15	114	56.9	18.6	3.16	408	162	6.06	1.73	21.3	4.50	1.00	0.00
4	4.96	1.50	35.7	34.4	21.1	12.8	626	563	6.09	4.77	24.5	10.9	1.00	0.01
5	4.91	1.82	30.2	30.4	18.2	13.7	530	606	5.39	5.06	22.0	12.0	0.99	0.01
6	4.28	0.10	5.60	0.28	4.54	1.40	44.5	26.9	1.31	0.08	9.00	3.28	0.99	0.00
7	24.7	8.31	183	37.4	15.1	5.70	295	257	7.70	1.49	17.4	7.31	1.00	0.00
8	12.8	6.04	84.0	65.7	23.4	5.91	656	320	5.74	2.38	26.7	7.44	1.00	0.00
9	13.9	3.97	106	47.2	31.1	10.8	1E3	690	7.36	1.65	33.0	11.7	1.00	0.00
10	7.88	1.93	29.9	44.5	13.7	13.5	407	505	3.06	3.74	19.3	13.4	1.00	0.00
Mean	**10.9**		**71.3**		**19.6**		**564**		**5.60**		**22.9**		**1.00**	
Sd	**6.55**		**52.8**		**7.52**		**304**		**1.98**		**6.91**		**0.00**	
MSA	**214.30**		**13.9E+03**		**282.50**		**46.1E+04**		**19.607**		**239.03**		**2.26E-05**	
MSW	**15.489**		**1807.3**		**87.958**		**18.7E+04**		**9.8742**		**85.213**		**1.49E-05**	
F	**13.8**		**7.70**		**3.21**		**2.46**		**1.99**		**2.81**		**1.51**	
p	**0.005**		**0.005**		**0.01**		**-**		**-**		**-**		**-**	

Table 4.8 Mean and standard deviation of moments for ΔA_0. The correlation coefficient r is with respect to exponential distribution. No. of utterances = 5, No. of speakers = 10. Mn = mean, Sd = std. dev., MSA = mean square among speakers, MSW = mean square within speakers. D.F. for F-ratio test: 9, 40. $(F)_{0.005} = 3.22$, $(F)_{0.01} = 2.89$.

Spk. No.	M_1		M_2		M_3		M_4		M_2/M_1		M_4/M_3		r	
	Mn	**Sd**	**Mn**	**Sd**	**Mn**	**Sd**	**Mn**	**Sd**	**Mn**	**Sd**	**Mn**	**Sd**	**Mn**	**Sd**
1	23.9	1.05	51.4	6.53	7.78	2.85	110	90.7	2.14	0.21	12.7	5.07	0.95	0.02
2	24.9	2.23	53.5	6.62	8.48	1.53	124	42.8	2.16	0.23	14.3	2.62	0.97	0.01
3	27.5	1.11	68.1	9.88	8.83	4.18	138	151	2.47	0.25	13.0	6.89	0.96	0.02
4	27.7	1.32	49.6	3.08	5.93	1.37	61.5	30.2	1.79	0.12	9.93	2.39	0.96	0.01
5	26.6	1.13	53.5	4.36	6.34	0.76	61.2	18.6	2.01	0.11	9.49	1.85	0.96	0.01
6	19.0	0.78	37.1	2.18	5.77	0.34	46.2	4.89	1.96	0.06	8.00	0.39	0.97	0.01
7	24.3	0.67	65.5	16.1	13.4	7.41	310	327	2.69	0.60	19.3	8.79	0.98	0.01
8	23.6	1.88	88.6	14.9	24.9	4.65	781	257	3.75	0.51	30.6	4.90	1.00	0.00
9	19.8	1.15	49.3	10.5	10.7	3.93	194	134	2.48	0.43	16.4	5.73	0.98	0.01
10	20.1	0.46	55.3	10.5	12.3	5.90	257	243	2.74	0.50	17.7	8.46	0.98	0.02
Mean	**23.7**		**57.2**		**10.4**		**208**		**2.42**		**15.2**		**0.97**	
Sd	**3.19**		**14.0**		**5.72**		**219**		**0.57**		**6.56**		**0.01**	
MSA	**50.870**		**979.56**		**163.65**		**24.0E+04**		**1.6019**		**214.98**		**0.0009**	
MSW	**1.6401**		**91.911**		**15.732**		**28.4E+03**		**0.1243**		**29.497**		**0.0002**	
F	**31.0**		**10.7**		**10.4**		**8.46**		**12.9**		**7.29**		**4.90**	
p	**0.005**		**0.005**		**0.005**		**0.005**		**0.005**		**0.005**		**0.005**	

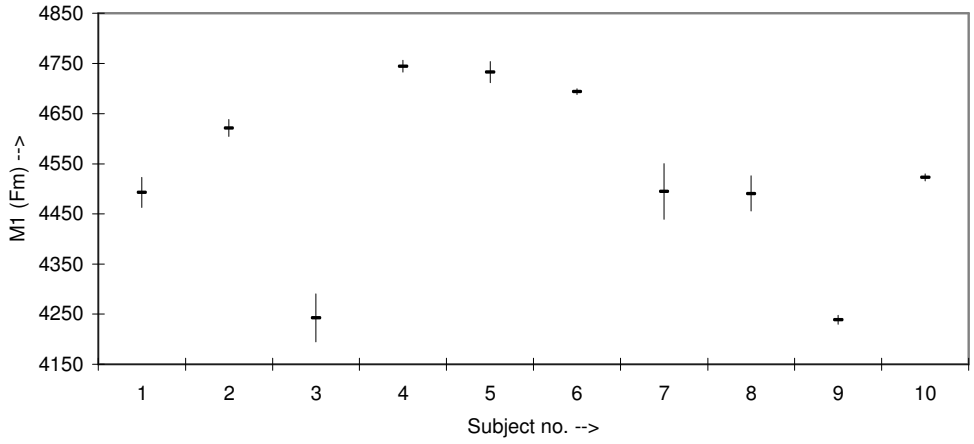

Fig 4.9 Variation of M_1 of F_m (Hz) for 10 speakers

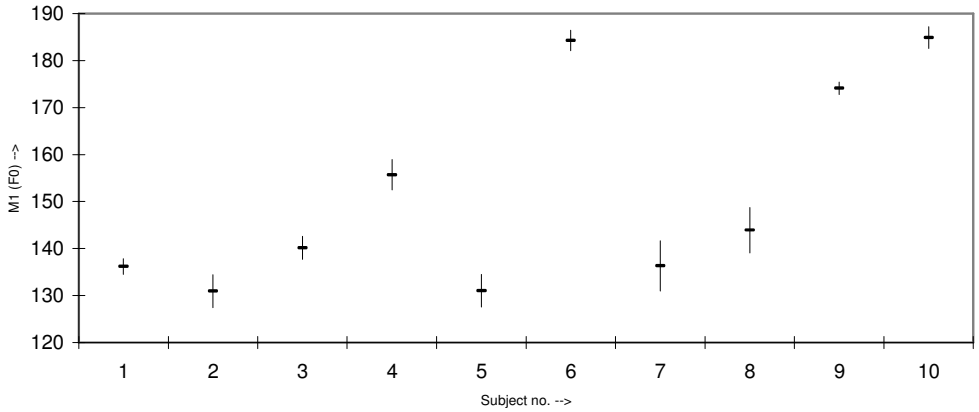

Fig 4.10 Variation of M_1 of F_0 (Hz) for 10 speakers

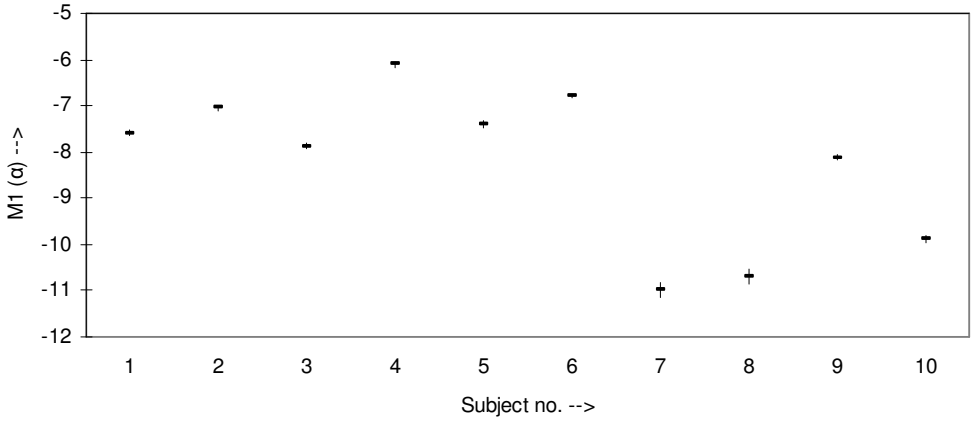

Fig 4.11 Variation of M_1 of α (dB) for 10 speakers

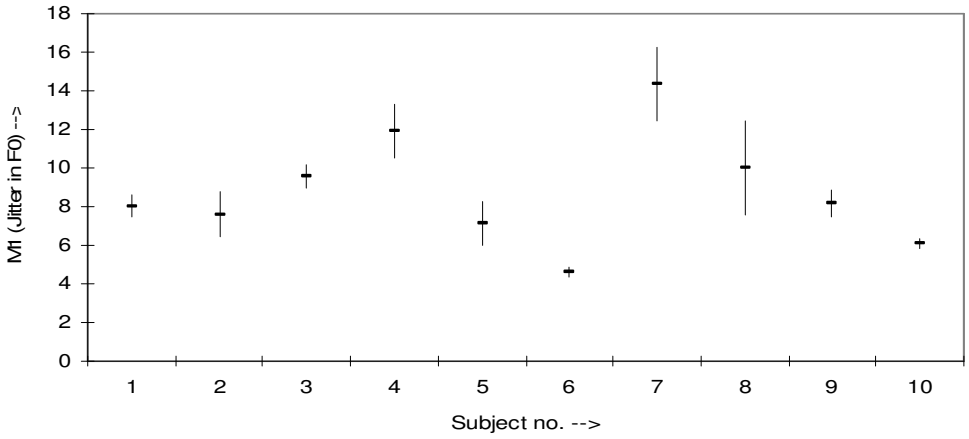

Fig 4.12 Variation of M_1 of ΔF_0 for 10 speakers

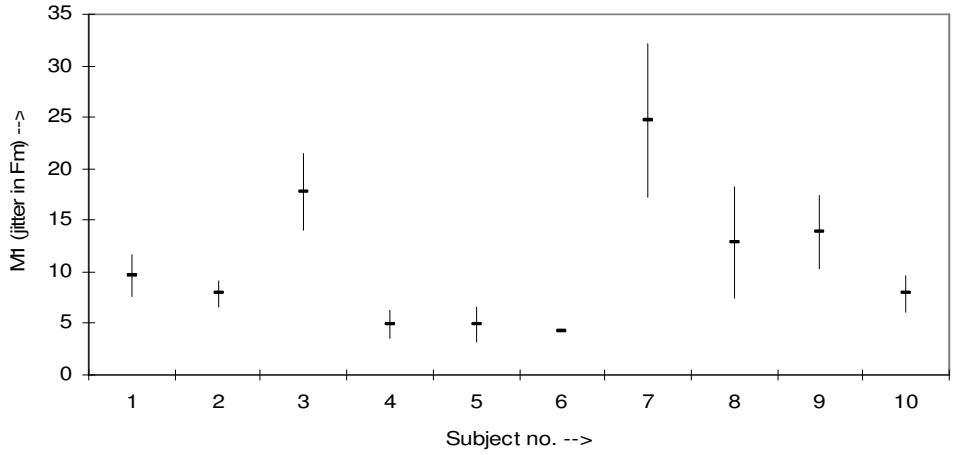

Fig 4.13 Variation of M_1 of ΔF_m for 10 speakers

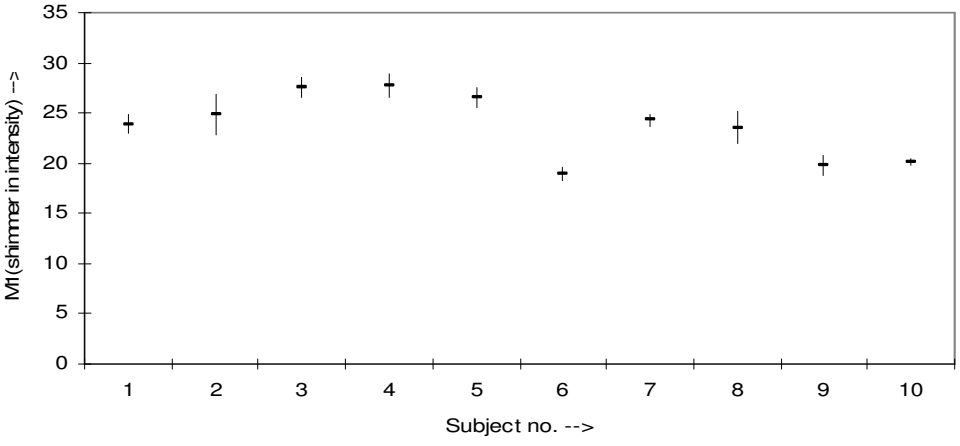

Fig 4.14 Variation of M_1 of ΔA_0 for 10 speakers

Chapter 5

SPEAKER MODELING AND RECOGNITION

The techniques for extracting the speaker dependent features, namely, peak amplitudes and HNM parameters (F_m, F_0, and α) were discussed in Chapter 3. Also, the use of these parameters for speaker recognition was studied through histogram analysis and statistical parameters in Chapter 4. The next step in training stage of the speaker recognition process, after extracting the speaker dependent features, is to create a model of the speaker from the extracted speaker dependent features. Some of the commonly used techniques for speaker modeling are minimum distance classifier, vector quantization (VQ), HMM and GMM [1], [7]. Among these, the first two are template models in which the pattern matching is deterministic and the latter two are stochastic models in which the pattern matching is probabilistic leading to probabilistic measure.

A problem with the minimum distance classifier is that it does not distinguish between acoustic classes, i.e., it uses an average of feature vectors per speaker computed over all sound classes [1], [20]. Hence, individual speech events are blurred. This results in low performance of the system using this technique. This problem can be solved using VQ technique in which the average of feature vectors is taken over distinct sound classes, but without identification or labeling. Hence, VQ performs better than minimum distance classifier [11]. In case of HMM and GMM, large amount of data are required for training. In case of GMM, if amount of data are small, the system can get numerically unstable. Also, GMM requires a lot of time for training. With limited amount of data available, VQ method is more robust and faster than HMM and GMM [8], [7]. Hence, VQ method is considered for speaker modeling in this project. In this chapter, training and testing algorithms using VQ technique are described. The recognition experiments carried out on HNM parameters using VQ method are also discussed and the results are compared with well established MFCC features. For this work, the HNM analysis software developed by Parveen K. Lehana [22] and the software of VQ and MFCC algorithms developed by P. Pradeep Kumar [30] have been used.

5.1 Vector quantization (VQ)

Vector quantization is the most commonly used technique for speaker modeling [8], [10], [11], [13]. In this method, the feature vectors extracted from the training utterance of a particular speaker are divided into clusters using K-means clustering algorithm. Then, mean vectors, also called centroids, are calculated for each cluster. Each centroid in the clustering represents an acoustic class, but without identification or labeling. These centroids are also called as code-vectors and the collection of such code-vectors is known as codebook. The number of code vectors represents the size of the codebook. The codebook represents the model for a particular speaker. The K-means clustering algorithm for creating the codebook of a speaker is described in the following steps.

1. The size of the codebook is assumed as K, i.e., K number of clusters will be formed. Also, K feature vectors are randomly chosen from the training feature vectors as initial centroids to represent K number of clusters.
2. The nearest neighbors of each centroid are determined by using Mahalanobis (or Euclidean) distance measure. A particular feature vector is determined as the neighbor of i^{th} centroid, if the distance between the feature vector and the i^{th} centroid is small compared to the distances with other centroids.
3. The centroids of the new clusters formed in the above step are again determined.
4. The above two steps (step 2 and step 3) are repeated until the centroids of previous and present iterations are same, i.e., until the process converges.

Thus, a group of centroids called codebook is formed that represents the model for a particular speaker. The codebooks are determined for all speakers and are stored in the database.

The training stage of the speaker recognition process is completed with the creation of the speaker models from the extracted speaker dependent features. The next stage of the speaker recognition process is testing stage. During the testing stage, the features extracted from the test utterance are to be compared with the existing models of speakers in the database and a decision is to be taken based on the comparison. The algorithm for the testing stage of the recognition process using VQ technique [6], [19] is described in the following steps.

1. The speaker dependent feature vectors are extracted from the test utterance.

2. An acoustic class (represented by centroid) is chosen for each test feature vector from the codebook of a particular speaker by finding the minimum distance with respect to various centroids.
3. The average of the minimum distances over all test feature vectors is computed.
4. The above three steps are repeated for all known speakers.
5. Finally, the speaker with smallest average minimum distance is identified as the claimed speaker.

The speaker recognition experiments were conducted on various speaker dependent features, namely, HNM parameters (F_m, F_0, and α) and MFCC using VQ method and the results are described in the following section.

5.2 Results and discussion

As discussed in Chapter 4, HNM parameters have shown good indications for the application of speaker recognition. Hence, speaker recognition experiments (by assuming closed set identification) are carried out on these features using VQ algorithm. For this purpose, the sentences shown in Table 4.2 were recorded for 10 male speakers at the sampling rate of 10 kSa/s. These sentences are repeated, for convenience, in Table 5.1. Each speaker has uttered these sentences five times. The procedure of the experiments carried out on these 10 speakers during training and testing stages are described in the following subsections.

Table 5.1 Sentences recorded for speaker recognition experiments (same as Table 4.1)

S. No.	Sentence type	Sentence or phrase
1.	Fricatives/Whispers/Affricatives	Bright sunshine shimmers on the ocean.
		His vicious father has seizures.
2.	Unvoiced Consonants	Get a calico cat to keep away.
		Primitive tribes have an upbeat attitude.
3.	Voiced Consonants	Did dad do academic bidding?
		Doctors prescribe drugs too freely.
4.	Predominantly Voiced (Vowels/Diphthongs/Semivowels)	He will allow a rare lie.
		Will Robin wear a yellow lily?

5.2.1 Training

During the training stage, the HNM parameters, namely, F_m, F_0, and α were extracted from the training utterance using the methods described in Chapter 3. These features, for each frame of the training utterance, were converted into a vector denoted by $[F_m \quad F_0 \quad \alpha]$. These vectors are called as feature vectors. The feature vectors were then applied to the VQ training algorithm (also called K-means clustering algorithm) by considering the codebook size as 18. This codebook was represented as a model for a particular speaker. The codebooks were created for 10 speakers and were stored in database.

5.2.2 Testing

During testing, the HNM parameters namely F_m, F_0, and α were extracted from the test utterance. The extracted features, for each frame of the test utterance, were then converted into a feature vector denoted by $[F_m \quad F_0 \quad \alpha]$. The feature vectors were then applied to VQ testing algorithm in which the average of minimum distances for test feature vectors with each speaker model in the database were calculated. Finally, the one with smallest average minimum distance was identified as the target speaker. The results of different recognition experiments are discussed below.

5.2.3 Speaker recognition results with HNM parameters

Initially, the recognition experiments were conducted on the training sentence that consists of concatenation of the 8 sentences given in Table 5.1. Each speaker is tested on 4 different utterances of the same concatenated sentence. In order to observe which one of the three HNM parameters (F_m, F_0, and α) contains more speaker information, the experiments were conducted on different feature combinations such as $[F_m \quad F_0]$, $[F_m \quad \alpha]$, and $[F_0 \quad \alpha]$. The experiments were carried with both Euclidean and Mahalanobis distance measures that are used in VQ clustering algorithm. The performance of recognition with Mahalanobis distance measure was better than Euclidean distance measure. This is due to the fact that the HNM feature vector contains the parameters, F_m, F_0, and α with different variances. Hence, results from Mahalanobis distance only will be discussed.

The experiments conducted using the feature vectors $[F_m \quad F_0]$, $[F_m \quad \alpha]$, $[F_0 \quad \alpha]$, and $[F_m \quad F_0 \quad \alpha]$ have produced 80 %, 95 %, 85 %, and 95 % recognition respectively. From this result, the following observations can be made.

- The above experiment has shown that HNM parameters contain some useful information for speaker recognition.
- The feature combinations that contain α have produced good recognition compared to other feature combinations. Hence, the parameter α may contain more speaker information compared to other HNM parameters.

In order to know how far the HNM parameters are useful for speaker recognition, the performance of speaker recognition using HNM parameters is compared with MFCC features. For this purpose, the experiments are carried out on the HNM feature vector $[F_m \quad F_0 \quad \alpha]$ using Mahalanobis distance measure. The comparison tests are described below.

5.2.4 Comparison of HNM parameters with MFCC features

For comparing the performance of HNM parameters in speaker recognition with well established parameters, the recognition experiments were also carried out on MFCC features. For this purpose, MFCC features were extracted using the method described in Chapter 2. The length of the MFCC feature vector was considered as 13, i.e., 13 mel cepstral coefficients were calculated for each frame of the sentence. The specifications of the mel filter bank used in MFCC calculation are given below [30].

Lowest frequency: 133 Hz

Number of linear filters: 13

Linear frequency spacing: 66.7 Hz

Number of log filters: 27

Log frequency spacing: 1.06

Initially, the experiments were conducted on 10 speakers with training on concatenated sentence. Testing is carried out with different features such as HNM parameters, MFCC features, and the combination of HNM parameters and MFCC features, on the following four different sentences.

Sentence-1: Bright sunshine shimmers on the ocean.

Sentence-2: His vicious father has seizures.

Sentence-3: He will allow a rare lie.

Sentence-4: Will Robin wear a yellow lily?

As stated earlier in Section 3.4, all the HNM parameters investigated for their usefulness for speaker recognition correspond to voiced frames only. HNM parameters correspond to unvoiced frames have not been used. When we combine HNM parameters with MFCC parameters, we leave the unvoiced frames. The results of this test are shown in Table 5.2. From this test, the following can be observed.

- MFCC features have performed better than HNM parameters.
- Combination of HNM parameters and MFCC features has performed better than HNM parameters alone. But, the combinational features have not shown much improvement in the performance compared to MFCC features alone. Even in some cases, the performance with the combinational features is degraded compared to MFCC features alone. This may be due to the fact that the MFCC features were ignored for the unvoiced frames while using along with the HNM parameters.

Table 5.2 Results of speaker recognition on different sentences, with training on concatenated sentence

Test sentence	Performance in percentage		
	HNM parameters (F_m, F_0, α)	MFCC features	HNM (F_m, F_0, α) and MFCC parameters
Sentence-1	75	95	95
Sentence-2	77.5	100	90
Sentence-3	75	100	92.5
Sentence-4	80	100	100

From the above observations, it is clear that MFCC features perform very well in the cases where the duration of the training utterance is around 20 s. In these cases, the HNM parameters or the combination of HNM parameters with MFCC are not of much use for speaker recognition. Hence, the recognition experiments with small duration training sentences were carried out. For this purpose, the experiments were conducted in such a way that training is carried out on one of the four sentences mentioned above as Sentence-1 to Sentence-4 and testing is carried out on all the 4-sentences. The results of these tests are shown in Table 5.3. From these results, the following observations can be made.

- MFCC features have performed better than HNM parameters only in text-dependent cases, i.e. when the training and test sentences are the same. But, MFCC features have not performed better than HNM parameters in text-independent experiments, when the training and test sentences are different.
- The combination of HNM parameters and MFCC features has shown good improvement in the performance in text-independent recognition.

Table 5.3 Results of speaker recognition with different combinations of training and test sentences. Row with the same training and test sentences is indicated in bold.

Training sentence	Test sentence	Performance in percentage		
		HNM parameters (F_m, F_0, α)	MFCC features	HNM (F_m, F_0, α) and MFCC parameters
Sentence-1	Sentence-1	**80**	**100**	**100**
	Sentence-2	74	74	82
	Sentence-3	74	58	84
	Sentence-4	64	62	76
Sentence-2	Sentence-1	78	54	84
	Sentence-2	**95**	**97.5**	**100**
	Sentence-3	72	60	68
	Sentence-4	76	72	80
Sentence-3	Sentence-1	64	54	74
	Sentence-2	58	42	84
	Sentence-3	**72.5**	**100**	**100**
	Sentence-4	66	80	88
Sentence-4	Sentence-1	56	46	66
	Sentence-2	64	66	76
	Sentence-3	66	78	78
	Sentence-4	**90**	**97.5**	**97.5**

The above experiments were carried out only on the three HNM parameters, F_m, F_0, and α. The jitter in F_m and F_0, and shimmer in intensity are also speaker dependent features and hence recognition experiments were carried out on ΔF_0, ΔF_m, and ΔA_0. The results of speaker recognition experiments carried out with different combinations of ΔF_0, ΔF_m, and ΔA_0 along with F_m, F_0, α, and MFCC are presented in Table 5.4. In this table, S-1 represents Sentence-1; S-2 represents Sentence-2, and so on. The results lead to the following observations.

- The performance of speaker recognition is improved with the use of parameters, ΔF_0, ΔF_m, and ΔA_0 along with F_m, F_0, α, and MFCC.
- The performance of recognition is in particular good with two combinations, namely, F_m, F_0, α, ΔF_0, ΔF_m, and MFCC, and F_m, F_0, α, ΔF_0, ΔF_m, ΔA_0, and MFCC.

Table 5.4 Results of speaker recognition using ΔF_0, ΔF_m, and ΔA_0 along with F_m, F_0, α, and MFCC, with different combinations of training and test sentences. Row with the same training and test sentences is indicated in bold.

Tr. Sen.	**Test sen.**	**Performance in percentage**						
		MFCC	**F_m, F_0, α, MFCC**	**F_m, F_0, α, ΔF_0, MFCC**	**F_m, F_0, α, ΔF_0, ΔF_m, MFCC**	**F_m, F_0, α, ΔF_0, ΔA_0, MFCC**	**F_m, F_0, α, ΔF_m, ΔA_0, MFCC**	**F_m, F_0, α, ΔF_0, ΔF_m, ΔA_0 MFCC**
S-1	S-1	**100**	**100**	**100**	**100**	**100**	**100**	**100**
	S-2	74	82	86	92	88	70	94
	S-3	58	84	82	88	80	82	82
	S-4	62	76	76	84	88	70	86
S-2	S-1	54	84	80	88	78	82	88
	S-2	**97.5**	**100**	**100**	**100**	**100**	**100**	**100**
	S-3	60	68	70	76	70	74	80
	S-4	72	80	84	86	76	80	90
S-3	S-1	54	74	66	70	68	68	72
	S-2	42	84	90	86	88	86	82
	S-3	**100**	**100**	**97.5**	**100**	**97.5**	**97.5**	**100**
	S-4	80	88	92	92	94	90	96
S-4	S-1	46	66	58	64	62	64	64
	S-2	66	76	80	80	76	76	74
	S-3	78	78	78	82	76	80	76
	S-4	**97.5**	**97.5**	**97.5**	**97.5**	**97.5**	**97.5**	**97.5**

Generally, MFCC fails in text independent speaker recognition systems that use small duration utterances for training. For better performance with MFCC, the duration of training utterance should be at least 24 s and the test utterance should be 5 s [6], [13]. From our results, it is clear that the performance of the system can be improved by using the combination of HNM and MFCC parameters, particularly in the cases where MFCC fails.

Chapter 6

SUMMARY AND CONCLUSION

6.1 Summary

Objective of this project was to study speech parameters for improving the performance of speaker recognition system. For this purpose, peak amplitudes of the speech signal and three parameters of HNM analysis (F_m, F_0, and α) were selected. In order to study intra-speaker and inter-speaker variability of these parameters, distribution of these parameters were obtained as histograms, and these were characterized through statistical moments and correlation coefficients with an exponential or normal distributions, on speech recordings from 10 male speakers.

For speaker modeling, VQ method was used. The speaker recognition experiments were carried out on HNM parameters using VQ method for 10 speakers. In order to establish relative usefulness of the parameters for speaker recognition, the performance was compared with that of the MFCC features, and a combination of HNM parameters and MFCC features.

6.2 Conclusion

The study of the investigated parameters, peak amplitudes of the speech signal and three HNM parameters (F_m, F_0, and α), using several statistical measures on the distribution has shown that peak amplitudes of the speech signal may not be useful for speaker recognition. This may be due to the fact that the peak amplitudes of the speech signal are highly sensitive to various factors such as background noise, transmission medium characteristics, etc. But, the HNM parameters have shown significant variations across the speakers as compared to intra-speaker variations and hence may be useful for speaker recognition.

The speaker recognition experiments conducted on HNM parameters using VQ method have shown that HNM parameters are useful for speaker recognition. Also, the parameter α contains more speaker information compared to the other two investigated

HNM parameters, F_m and F_0. The performance of speaker recognition based on HNM parameters is comparable to that of MFCC features. However, the speaker dependent feature vector with HNM parameters contains only 3 elements while the MFCC feature vector contains more number of elements (13 in our case). Hence, the recognition with MFCC involves more number of computations (particularly in testing stage) than with HNM parameters. Also, MFCC features do not perform well when limited amount of data are available for training. It has been observed that in addition to F_m, F_0, and α; we can also use jitter in F_0 and F_m as additional parameters. The performance of the speaker recognition system can be improved by using the HNM parameters along with MFCC features, particularly in the cases where MFCC fails.

6.3 Suggestions for future work

The speaker recognition experiments, with HNM parameters, are to be carried out on large databases. Further, the performance of the speaker recognition systems based on HNM parameters may be improved by using GMM for speaker modeling, provided large amount of data are available for training. The amplitudes and phases of the harmonics obtained in the harmonic part of the HNM may also be used for speaker recognition by representing them with GMM. The performance of recognition based on HNM parameters has to be compared with that of MFCC, in the cases where speech contains a moderate level of background noise.

Appendix A

RESULTS OF MULTI-SAMPLE HISTOGRAM ANALYSIS

The experimental results of two-sample and three-sample histogram analyses carried out on peak amplitudes and HNM parameters (F_m, F_0, and α) for 10 male speakers are presented here. Also, the statistical parameters, namely, moments and correlation coefficients, calculated for characterizing the two-sample and three-sample histograms are presented. The variation of each statistical parameter across 10 speakers was observed through different plots. The plots showing the variation of M_1 only are presented here. Similar results were observed from the plots showing the variation of other statistical parameters.

A.1 Two-sample and three-sample histogram plots

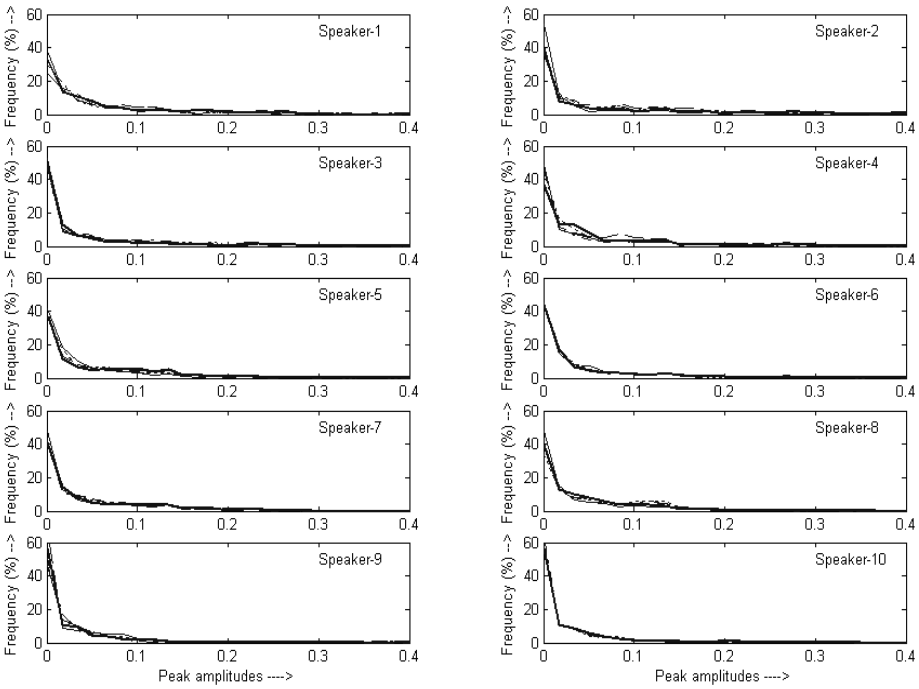

Fig. A.1 Two-sample histograms of peak amplitude

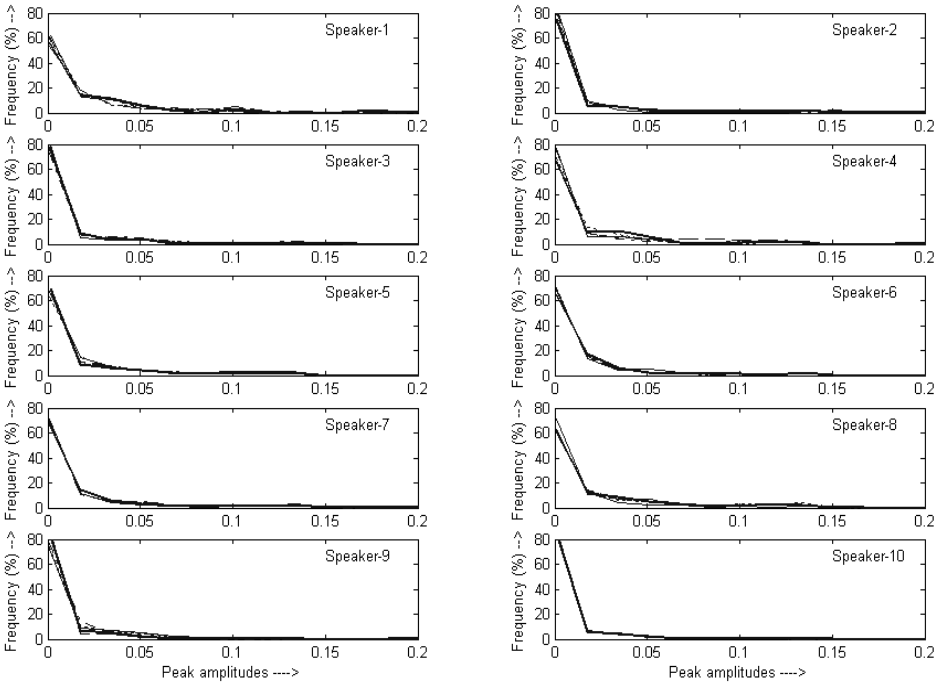

Fig. A.2 Three-sample histograms of peak amplitude

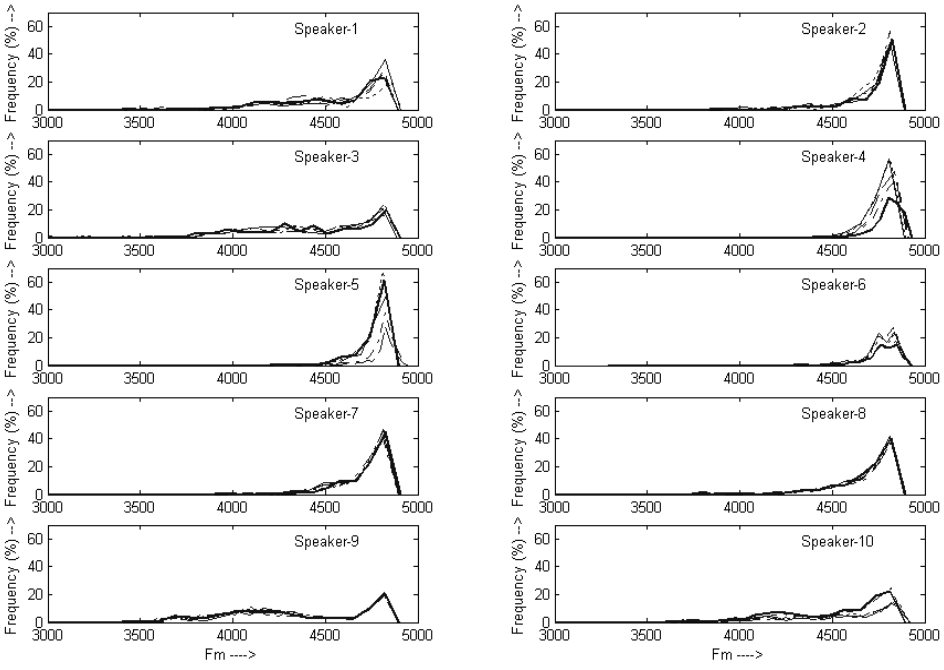

Fig. A.3 Two-sample histograms of F_m (Hz)

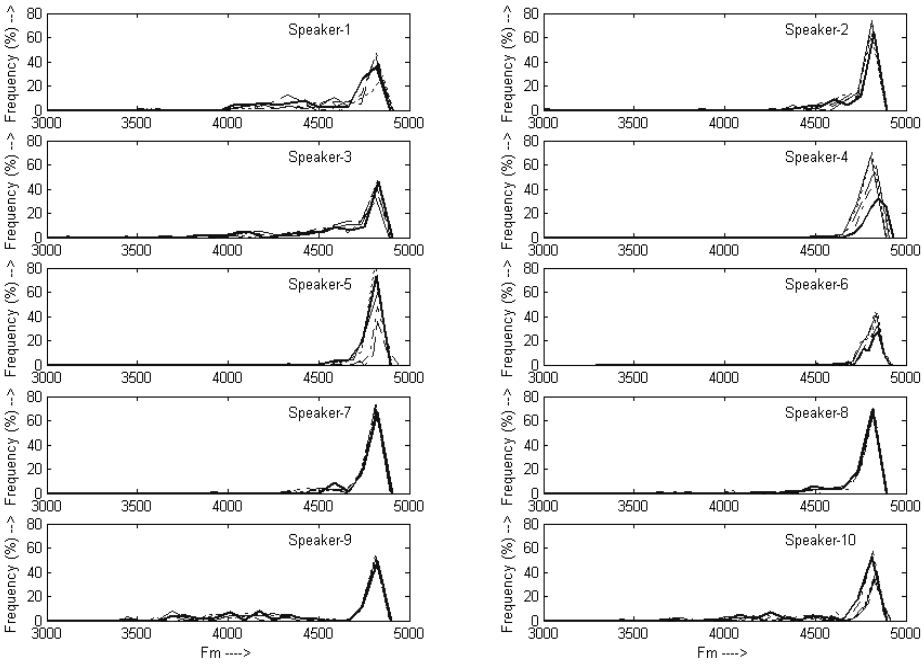

Fig. A.4 Three-sample histograms of F_m (Hz)

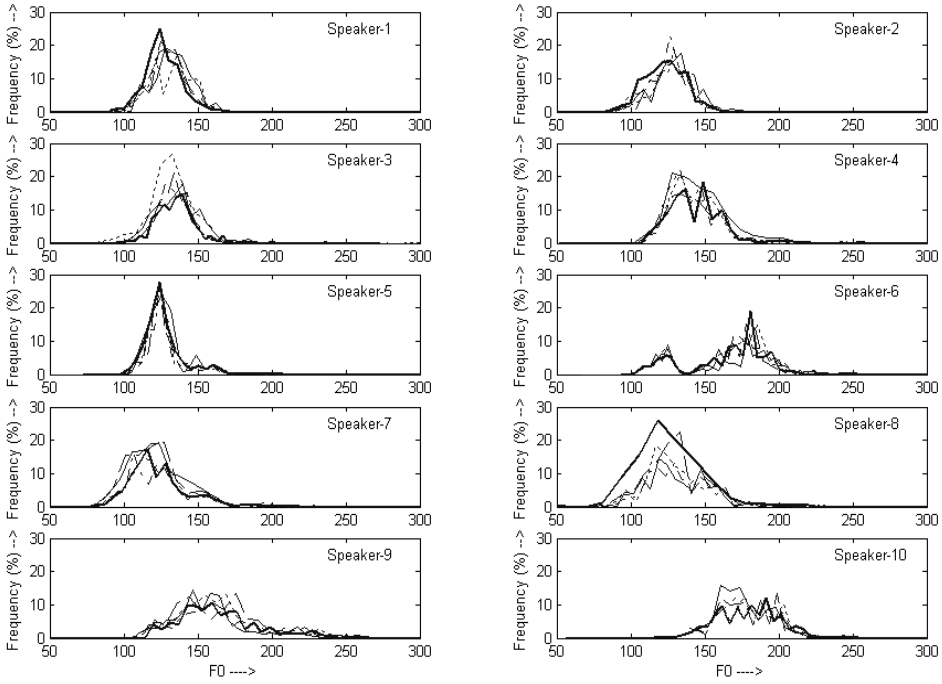

Fig. A.5 Two-sample histograms of F_0 (Hz)

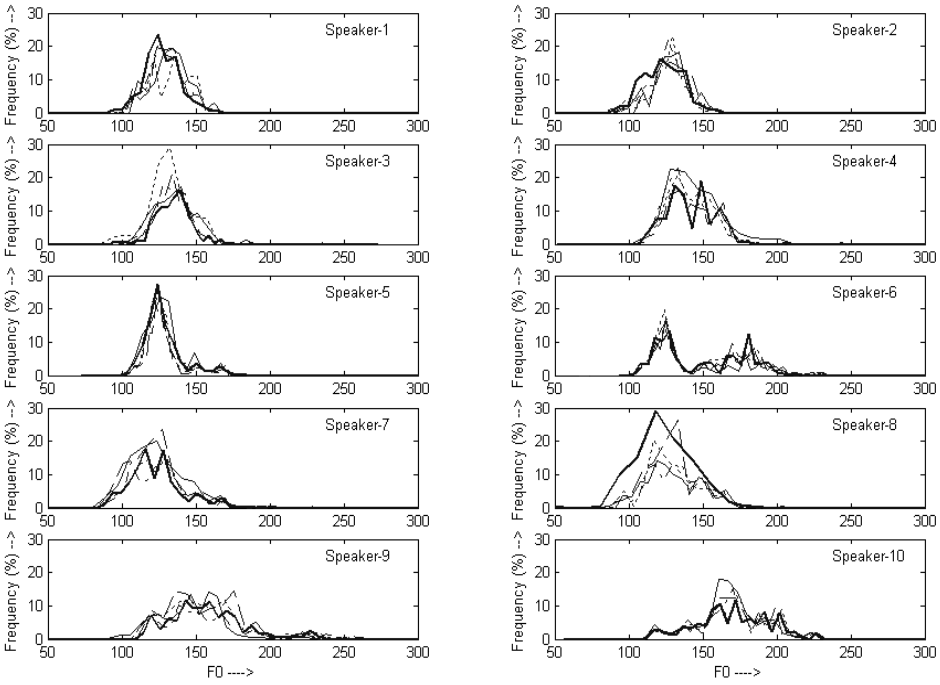

Fig. A.6 Three-sample histograms of F_0 (Hz)

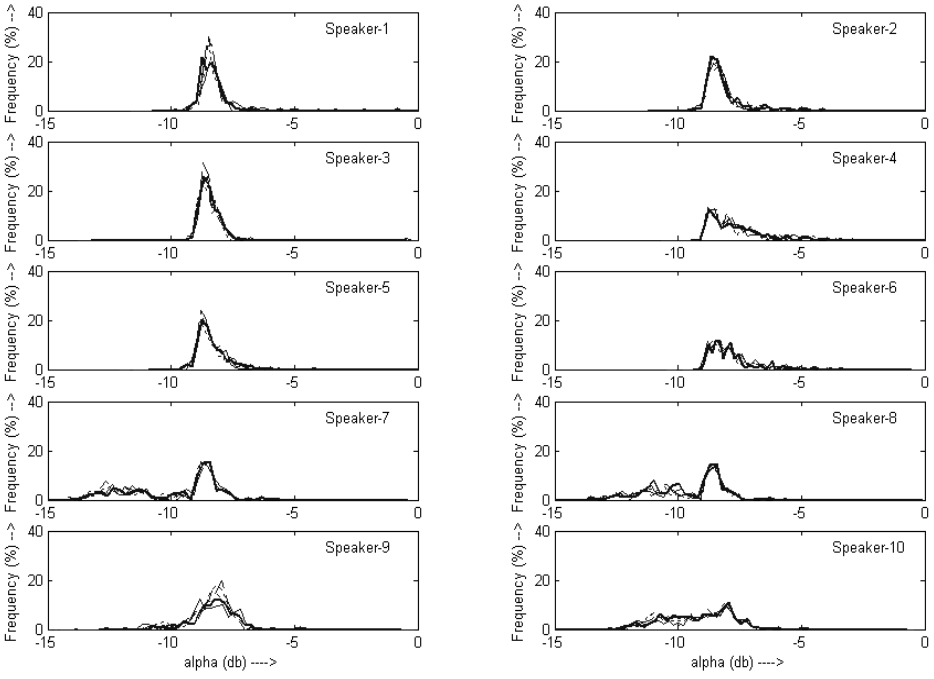

Fig. A.7 Two-sample histograms of α (dB)

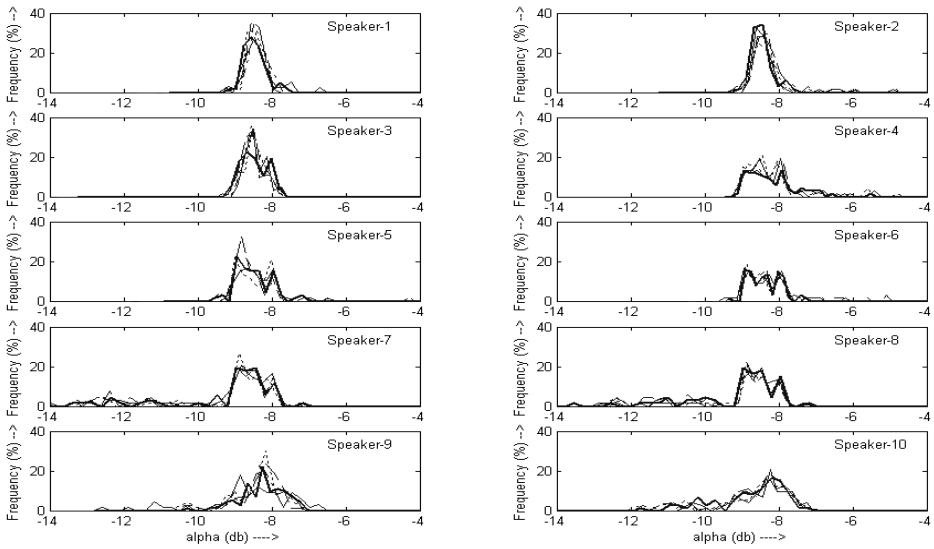

Fig. A.8 Three-sample histograms of α (dB)

A.2 Variation of first moment (M_1) across speakers

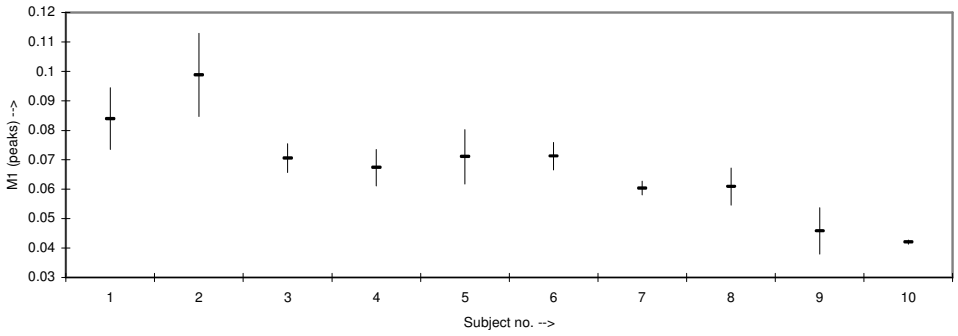

Fig. A.9 Variation of M_1 of peak amplitude for two-sample histograms

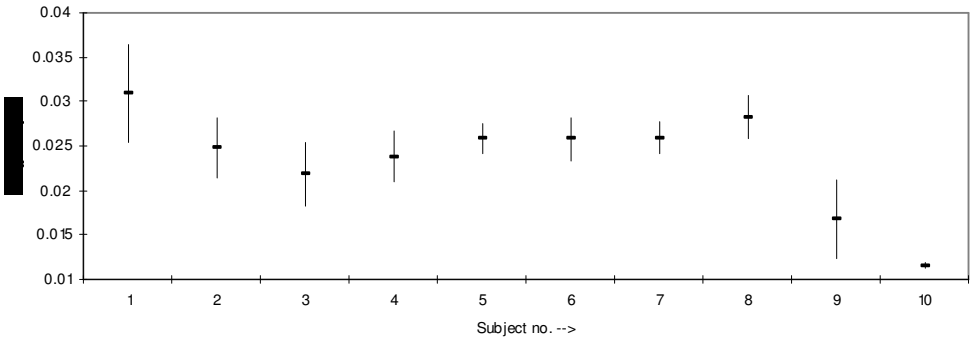

Fig. A.10 Variation of M_1 of peak amplitude for three-sample histograms

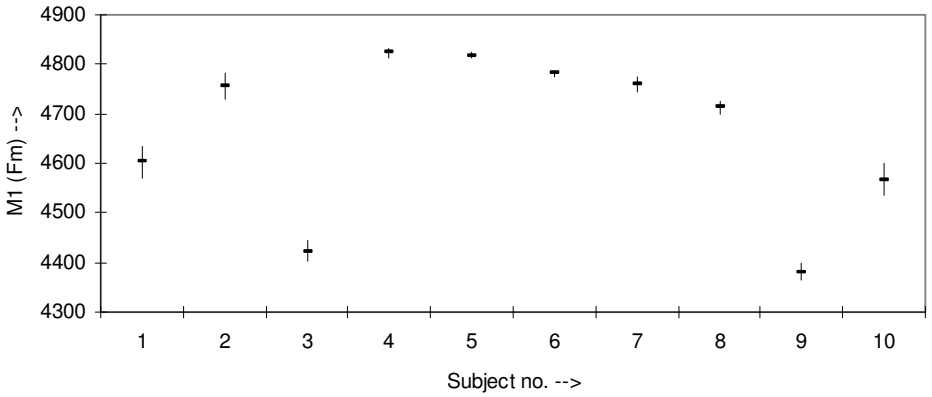

Fig. A.11 Variation of M_1 of F_m (Hz) for two-sample histograms

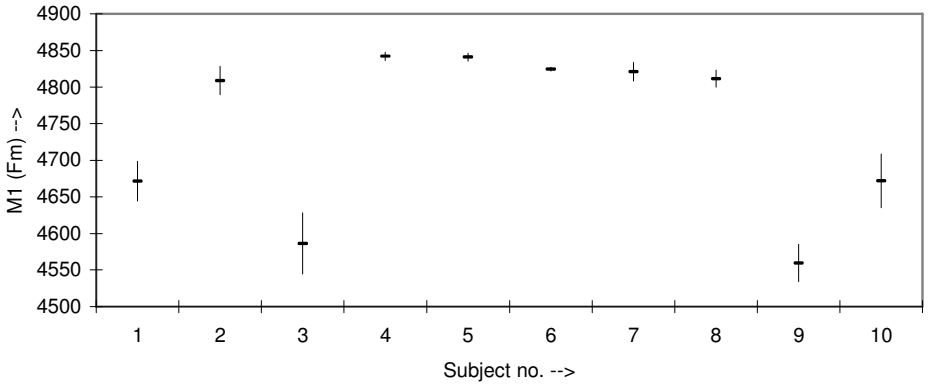

Fig. A.12 Variation of M_1 of F_m (Hz) for three-sample histograms

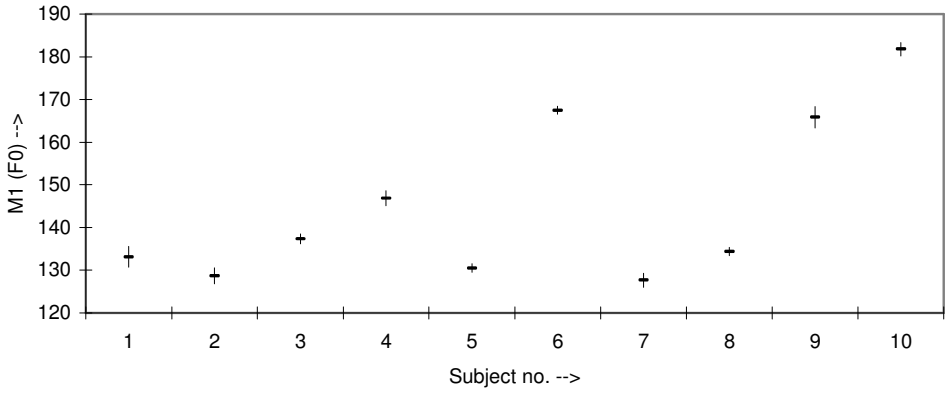

Fig. A.13 Variation of M_1 of F_0 (Hz) for two-sample histograms

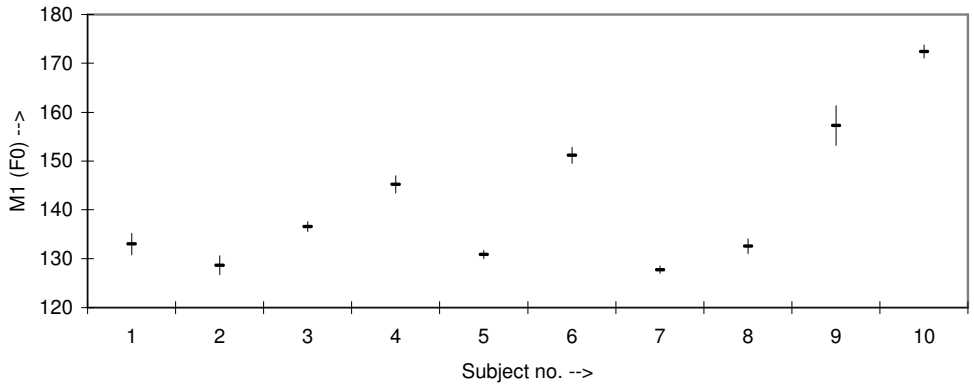

Fig. A.14 Variation of M_1 of F_0 (Hz) for three-sample histograms

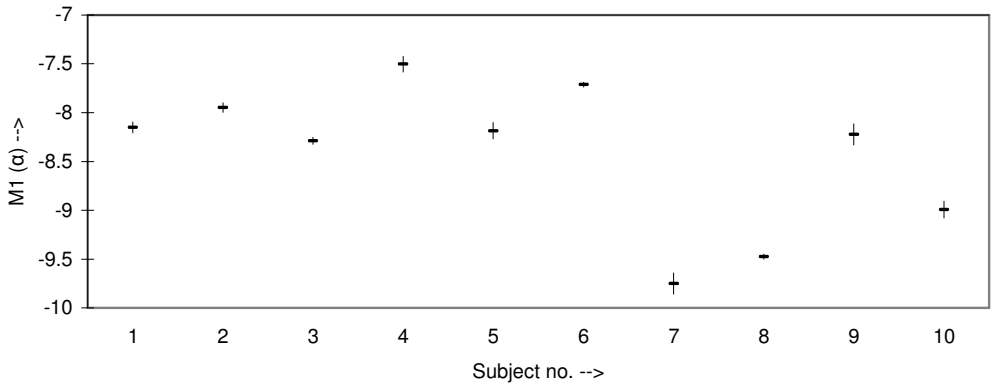

Fig. A.15 Variation of M_1 of α (dB) for two-sample histograms

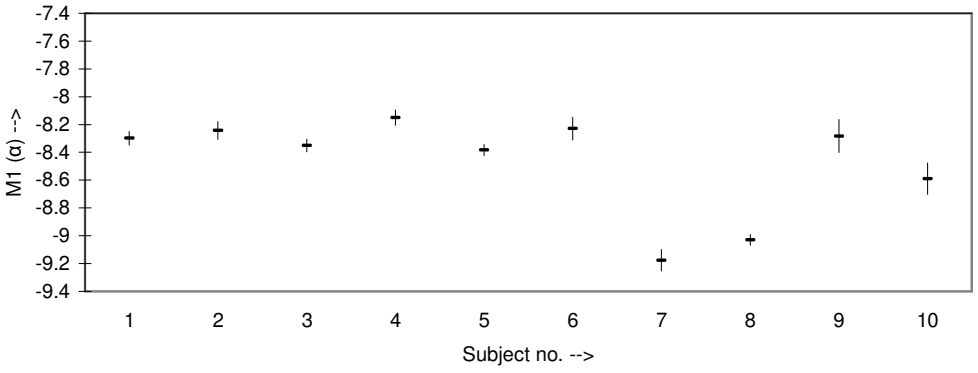

Fig. A.16 Variation of M_1 of α (dB) for three-sample histograms

A.3 Statistical parameters for histograms

Table A.1 Mean and standard deviation of moments for two-sample histograms of peak amplitude. The correlation coefficient *r* is with respect to exponential distribution. No. of utterances = 5, No. of speakers = 10. Mn = mean, Sd = std. dev., MSA = mean square among speakers, MSW = mean square within speakers. D.F. for F-ratio test: 9, 40. $(F)_{0.005} = 3.22$, $(F)_{0.01} = 2.89$.

Spk. No.	M_1		M_2		M_3		M_4		M_2/M_1		M_4/M_3		*r*	
	Mn	**Sd**	**Mn**	**Sd**	**Mn**	**Sd**	**Mn**	**Sd**	**Mn**	**Sd**	**Mn**	**Sd**	**Mn**	**Sd**
1	0.08	0.01	0.11	0.01	2.27	0.15	9.15	1.35	1.30	0.09	4.03	0.46	0.91	0.01
2	0.10	0.02	0.14	0.02	2.08	0.29	7.74	1.28	1.39	0.21	3.72	0.28	0.78	0.05
3	0.07	0.01	0.11	0.01	2.61	0.21	10.9	1.07	1.57	0.09	4.16	0.13	0.80	0.02
4	0.07	0.01	0.09	0.01	2.34	0.33	10.3	3.75	1.33	0.08	4.31	0.97	0.87	0.06
5	0.07	0.01	0.10	0.01	2.92	0.55	14.7	5.99	1.39	0.09	4.89	0.93	0.90	0.03
6	0.07	0.01	0.11	0.01	2.65	0.31	10.6	2.44	1.57	0.06	3.94	0.42	0.85	0.02
7	0.06	0.00	0.08	0.01	2.09	0.22	8.38	1.65	1.26	0.09	3.99	0.50	0.89	0.02
8	0.06	0.01	0.07	0.01	2.53	0.59	13.7	6.97	1.23	0.11	5.17	1.36	0.91	0.02
9	0.05	0.01	0.08	0.01	3.18	0.40	14.8	3.30	1.67	0.17	4.62	0.42	0.90	0.01
10	0.04	0.00	0.07	0.00	3.46	0.14	18.8	2.66	1.78	0.07	5.43	0.59	0.88	0.01
Mean	**0.07**		**0.10**		**2.61**		**11.9**		**1.45**		**4.43**		**0.87**	
Sd	**0.02**		**0.02**		**0.46**		**3.49**		**0.19**		**0.57**		**0.05**	
MSA	**0.0014**		**0.0023**		**1.0502**		**60.831**		**0.1742**		**1.6524**		**0.0105**	
MSW	**7.3E-05**		**0.0001**		**0.1238**		**12.973**		**0.0131**		**0.4923**		**0.0009**	
F	**19.0**		**20.8**		**8.48**		**4.69**		**13.3**		**3.36**		**12.1**	
p	**0.005**		**0.005**		**0.005**		**0.005**		**0.005**		**0.005**		**0.005**	

Table A.2 Mean and standard deviation of moments for three-sample histograms of peak amplitude. The correlation coefficient *r* is with respect to exponential distribution. No. of utterances = 5, No. of speakers = 10. Mn = mean, Sd = std. dev., MSA = mean square among speakers, MSW = mean square within speakers. D.F. for F-ratio test: 9, 40. $(F)_{0.005} = 3.22$, $(F)_{0.01} = 2.89$.

Spk. No.	M_1		M_2		M_3		M_4		M_2/M_1		M_4/M_3		*r*	
	Mn	**Sd**	**Mn**	**Sd**	**Mn**	**Sd**	**Mn**	**Sd**	**Mn**	**Sd**	**Mn**	**Sd**	**Mn**	**Sd**
1	0.03	0.01	0.05	0.02	3.65	1.09	21.9	11.7	1.48	0.25	5.68	1.41	0.95	0.02
2	0.02	0.00	0.06	0.01	4.67	1.23	28.5	12.9	2.28	0.54	5.88	1.16	0.91	0.03
3	0.02	0.00	0.04	0.02	3.98	1.19	23.0	12.0	1.91	0.42	5.50	1.31	0.93	0.02
4	0.02	0.00	0.04	0.00	2.88	0.85	13.4	6.67	1.50	0.16	4.39	1.13	0.94	0.02
5	0.03	0.00	0.04	0.01	3.98	1.75	30.3	26.0	1.57	0.18	6.60	2.93	0.94	0.02
6	0.03	0.00	0.05	0.01	5.61	1.42	48.7	23.8	1.84	0.20	8.29	2.18	0.95	0.01
7	0.03	0.00	0.04	0.01	2.99	0.65	13.6	6.27	1.51	0.18	4.38	1.05	0.94	0.02
8	0.03	0.00	0.04	0.00	2.85	1.60	16.6	20.5	1.38	0.23	4.65	2.61	0.94	0.01
9	0.02	0.00	0.03	0.01	8.09	1.23	105	44.8	1.90	0.14	12.6	3.66	0.97	0.02
10	0.01	0.00	0.02	0.00	7.15	2.28	87.7	45.1	1.74	0.23	11.5	3.41	0.99	0.00
Mean	**0.02**		**0.04**		**4.59**		**38.9**		**1.71**		**6.94**		**0.94**	
Sd	**0.01**		**0.01**		**1.83**		**32.2**		**0.28**		**2.93**		**0.02**	
MSA	**0.0002**		**0.0005**		**16.686**		**5173.3**		**0.3815**		**42.902**		**0.0022**	
MSW	**1.3E-05**		**0.0001**		**1.9594**		**622.86**		**0.0789**		**5.2568**		**0.0003**	
F	**12.7**		**4.70**		**8.52**		**8.31**		**4.83**		**8.16**		**7.15**	
p	**0.005**		**0.005**		**0.005**		**0.005**		**0.005**		**0.005**		**0.005**	

Table A.3 Mean and standard deviation of moments for two-sample histograms of F_m(Hz). The correlation coefficient r is with respect to normal distribution. No. of utterances = 5, No. of speakers = 10. Mn = mean, Sd = std. dev., MSA = mean square among speakers, MSW = mean square within speakers. D.F. for F-ratio test: 9, 40. $(F)_{0.005} = 3.22$, $(F)_{0.01} = 2.89$.

Spk.	M_1		M_2		M_3		M_4		M_2/M_1		M_4/M_3		r	
No.	Mn	Sd	Mn	Sd	Mn	Sd	Mn	Sd	Mn	Sd	Mn	Sd	Mn	Sd
1	4602	36.5	319	43.2	-1.5	0.57	6.05	3.03	0.07	0.01	-3.9	0.52	0.72	0.04
2	4755	29.8	252	72.3	-3.7	0.70	21.1	7.01	0.05	0.02	-5.5	0.91	0.66	0.07
3	4422	23.4	536	44.9	-1.8	0.21	6.79	1.34	0.12	0.01	-3.7	0.32	0.70	0.05
4	4822	8.99	80.7	9.96	-2.1	0.66	12.5	5.88	0.02	0.00	-5.7	0.87	0.92	0.03
5	4818	5.54	90.7	10.0	-2.3	0.37	10.2	2.96	0.02	0.00	-4.4	0.59	0.83	0.02
6	4781	5.35	111	9.77	-2.2	0.50	11.1	3.57	0.02	0.00	-4.9	0.55	0.89	0.01
7	4758	16.9	232	64.7	-5.8	3.74	67.5	65.9	0.05	0.01	-9.2	4.64	0.75	0.05
8	4712	15.7	308	42.1	-5.1	2.21	47.0	35.1	0.07	0.01	-8.1	3.00	0.70	0.02
9	4382	18.9	438	45.8	-1.1	0.87	7.55	5.76	0.10	0.01	-6.9	0.98	0.70	0.01
10	4566	36.0	313	25.4	-0.9	0.15	2.94	0.44	0.07	0.01	-3.2	0.07	0.72	0.03
Mean	**4662**		**268**		**-2.7**		**19.3**		**0.06**		**-5.6**		**0.76**	
Sd	**161**		**149**		**1.68**		**21.2**		**0.03**		**1.99**		**0.09**	
MSA	**12.9E+04**		**11.1E+04**		**14.060**		**2239.8**		**0.0059**		**19.792**		**0.0398**	
MSW	**509.62**		**1810.3**		**2.1310**		**571.85**		**8.86E-05**		**3.4109**		**0.0015**	
F	**255**		**61.1**		**6.60**		**3.92**		**66.4**		**5.80**		**26.3**	
p	**0.005**		**0.005**		**0.005**		**0.005**		**0.005**		**0.005**		**0.005**	

Table A.4 Mean and standard deviation of moments for three-sample histograms of F_m(Hz). The correlation coefficient r is with respect to normal distribution. No. of utterances = 5, No. of speakers = 10. Mn = mean, Sd = std. dev., MSA = mean square among speakers, MSW = mean square within speakers. D.F. for F-ratio test: 9, 40. $(F)_{0.005} = 3.22$, $(F)_{0.01} = 2.89$.

Spk.	M_1		M_2		M_3		M_4		M_2/M_1		M_4/M_3		r	
No.	Mn	Sd	Mn	Sd	Mn	Sd	Mn	Sd	Mn	Sd	Mn	Sd	Mn	Sd
1	4671	30.4	261	25.9	-1.3	0.28	4.17	1.25	0.06	0.01	-3.1	0.46	0.61	0.05
2	4809	21.5	157	91.7	-3.1	1.56	15.5	11.6	0.03	0.02	-4.5	1.27	0.69	0.10
3	4586	46.7	511	43.6	-2.6	0.26	9.86	1.50	0.11	0.01	-3.7	0.21	0.53	0.06
4	4842	6.19	51.8	3.92	-1.0	0.33	4.76	1.86	0.01	0.00	-4.5	0.49	0.96	0.01
5	4841	6.01	57.0	5.62	-2.4	0.77	12.3	5.82	0.01	0.00	-4.9	0.74	0.89	0.05
6	4824	3.00	57.3	4.61	-2.1	0.13	10.8	1.24	0.01	0.00	-5.1	0.58	0.92	0.02
7	4821	14.1	110	43.1	-2.7	0.82	11.5	5.87	0.02	0.01	-4.0	0.79	0.75	0.08
8	4811	13.2	134	39.6	-3.1	0.68	14.6	5.32	0.03	0.01	-4.6	0.67	0.71	0.07
9	4560	28.8	432	57.1	-1.5	0.76	6.03	6.10	0.09	0.01	-3.5	1.42	0.39	0.01
10	4672	41.2	296	47.3	-1.6	0.26	4.83	1.17	0.06	0.01	-2.9	0.24	0.47	0.05
Mean	**4744**		**207**		**-2.20**		**9.42**		**0.04**		**-4.10**		**0.69**	
Sd	**110**		**163**		**0.75**		**4.21**		**0.04**		**0.75**		**0.19**	
MSA	**60.8E+03**		**13.3E+04**		**2.7875**		**88.555**		**0.0065**		**2.8087**		**0.1889**	
MSW	**654.49**		**1996.1**		**0.5073**		**27.867**		**9.60E-05**		**0.6132**		**0.0032**	
F	**92.9**		**66.8**		**5.49**		**3.18**		**67.6**		**4.58**		**58.2**	
p	**0.005**		**0.005**		**0.005**		**0.01**		**0.005**		**0.005**		**0.005**	

Table A.5 Mean and standard deviation of moments for two-sample histograms of F_0(Hz). The correlation coefficient r is with respect to normal distribution. No. of utterances = 5, No. of speakers = 10. Mn = mean, Sd = std. dev., MSA = mean square among speakers, MSW = mean square within speakers. D.F. for F-ratio test: 9, 40. $(F)_{0.005} = 3.22$, $(F)_{0.01} = 2.89$.

Spk. No.	M_1		M_2		M_3		M_4		M_2/M_1		M_4/M_3		r	
	Mn	**Sd**	**Mn**	**Sd**	**Mn**	**Sd**	**Mn**	**Sd**	**Mn**	**Sd**	**Mn**	**Sd**	**Mn**	**Sd**
1	133	2.72	12.6	0.82	0.59	1.11	9.11	12.8	0.09	0.01	11.7	17.9	0.97	0.03
2	129	2.13	14.4	1.20	0.41	0.64	6.27	4.27	0.11	0.01	-12	44.3	0.97	0.01
3	137	1.24	16.9	3.21	2.98	2.87	32.3	44.6	0.12	0.02	8.55	3.05	0.95	0.03
4	147	1.99	20.6	1.21	3.11	0.75	29.1	11.7	0.14	0.01	9.07	1.80	0.93	0.03
5	131	1.10	15.3	0.87	2.35	0.41	12.9	6.65	0.12	0.01	5.25	1.63	0.82	0.04
6	167	1.06	29.0	2.18	-0.3	0.19	2.76	0.78	0.17	0.01	-11	9.87	0.72	0.07
7	128	1.84	24.9	6.20	2.34	1.71	15.8	17.4	0.19	0.05	5.77	1.72	0.91	0.03
8	134	1.08	24.7	5.21	2.33	2.66	22.3	35.2	0.18	0.04	7.07	2.85	0.92	0.01
9	166	2.78	29.5	1.78	1.03	0.13	4.79	1.05	0.18	0.01	4.63	0.45	0.91	0.03
10	182	1.75	23.9	2.19	1.15	0.48	9.66	1.94	0.13	0.01	8.87	1.47	0.91	0.02
Mean	**145**		**21.2**		**1.60**		**14.5**		**0.14**		**3.79**		**0.90**	
Sd	**19.4**		**6.13**		**1.18**		**10.3**		**0.03**		**8.33**		**0.08**	
MSA	**1880.3**		**187.88**		**6.9928**		**525.87**		**0.0061**		**347.16**		**0.0292**	
MSW	**3.5077**		**9.3038**		**2.0890**		**389.87**		**0.0005**		**240.39**		**0.0011**	
F	**536**		**20.2**		**3.35**		**1.35**		**12.9**		**1.44**		**26.5**	
p	**0.005**		**0.005**		**0.005**		**-**		**0.005**		**-**		**0.005**	

Table A.6 Mean and standard deviation of moments for three-sample histograms of F_0(Hz). The correlation coefficient r is with respect to normal distribution. No. of utterances = 5, No. of speakers = 10. Mn = mean, Sd = std. dev., MSA = mean square among speakers, MSW = mean square within speakers. D.F. for F-ratio test: 9, 40. $(F)_{0.005} = 3.22$, $(F)_{0.01} = 2.89$.

Spk. No.	M_1		M_2		M_3		M_4		M_2/M_1		M_4/M_3		r	
	Mn	**Sd**	**Mn**	**Sd**	**Mn**	**Sd**	**Mn**	**Sd**	**Mn**	**Sd**	**Mn**	**Sd**	**Mn**	**Sd**
1	133	2.48	12.1	0.87	0.02	0.16	2.83	0.38	0.09	0.01	26.8	21.8	0.96	0.04
2	129	2.19	12.8	1.07	-0.0	0.32	4.01	2.07	0.10	0.01	13.1	49.0	0.96	0.02
3	137	1.15	13.6	1.34	0.55	0.64	7.98	4.82	0.10	0.01	-17	44.9	0.97	0.01
4	145	1.96	16.4	0.95	1.03	0.62	6.62	4.37	0.11	0.01	6.6	1.45	0.94	0.03
5	131	0.98	13.9	0.24	1.68	0.24	6.61	0.78	0.11	0.00	3.96	0.17	0.84	0.05
6	151	1.85	29.7	1.03	0.44	0.17	1.94	0.20	0.20	0.01	4.83	1.13	0.61	0.05
7	128	0.86	22.6	3.06	4.47	2.99	60.5	49.8	0.18	0.02	10.8	5.10	0.93	0.03
8	133	1.66	19.1	0.76	0.67	0.32	4.69	2.05	0.14	0.01	7.17	0.90	0.95	0.03
9	157	4.55	26.7	2.70	1.07	0.24	5.59	1.75	0.17	0.01	5.25	1.23	0.92	0.03
10	172	1.51	25.3	1.71	0.65	0.35	7.43	1.73	0.15	0.01	12.7	2.94	0.88	0.01
Mean	**142**		**19.2**		**1.05**		**10.8**		**0.13**		**7.42**		**0.90**	
Sd	**14.8**		**6.46**		**1.30**		**17.6**		**0.04**		**10.9**		**0.11**	
MSA	**1088.0**		**208.83**		**8.4810**		**1544.9**		**0.0071**		**592.97**		**0.0581**	
MSW	**4.6959**		**2.5840**		**1.0221**		**253.62**		**0.0001**		**493.08**		**0.0009**	
F	**232**		**80.8**		**8.30**		**6.09**		**63.8**		**1.20**		**62.6**	
p	**0.005**		**0.005**		**0.005**		**0.005**		**0.005**		**-**		**0.005**	

Table A.7 Mean and standard deviation of moments for two-sample histograms of α(dB). The correlation coefficient r is with respect to normal distribution. No. of utterances = 5, No. of speakers = 10. Mn = mean, Sd = std. dev., MSA = mean square among speakers, MSW = mean square within speakers. D.F. for F-ratio test: 9, 40. $(F)_{0.005} = 3.22$, $(F)_{0.01} = 2.89$.

Spk. No.	M_1		M_2		M_3		M_4		M_2/M_1		M_4/M_3		r	
	Mn	**Sd**	**Mn**	**Sd**	**Mn**	**Sd**	**Mn**	**Sd**	**Mn**	**Sd**	**Mn**	**Sd**	**Mn**	**Sd**
1	-8.2	0.06	0.87	0.13	5.02	1.07	36.8	11.0	-0.1	0.02	7.19	0.86	0.77	0.06
2	-7.9	0.06	1.18	0.09	3.53	0.56	19.9	5.17	-0.1	0.01	5.54	0.69	0.69	0.03
3	-8.3	0.04	0.76	0.20	5.87	1.62	50.9	22.3	-0.1	0.02	8.36	1.86	0.82	0.09
4	-7.5	0.09	1.37	0.12	1.74	0.27	6.44	1.59	-0.2	0.02	3.67	0.34	0.73	0.05
5	-8.2	0.09	0.86	0.21	2.66	0.67	14.0	5.00	-0.1	0.03	5.18	0.68	0.82	0.06
6	-7.7	0.03	1.07	0.07	1.90	0.26	7.45	1.57	-0.1	0.01	3.88	0.33	0.78	0.03
7	-9.8	0.12	1.98	0.07	-0.3	0.11	2.12	0.23	-0.2	0.01	-7.3	2.15	0.56	0.06
8	-9.5	0.03	1.63	0.05	-0.2	0.06	2.66	0.16	-0.2	0.01	-11	2.38	0.67	0.04
9	-8.2	0.12	1.04	0.15	0.24	0.75	8.18	2.97	-0.1	0.02	-31	66.9	0.90	0.03
10	-9.0	0.10	1.54	0.05	0.20	0.14	4.03	0.19	-0.2	0.00	63.3	102	0.88	0.03
Mean	**-8.4**		**1.23**		**2.06**		**15.2**		**-0.1**		**4.80**		**0.76**	
Sd	**0.74**		**0.39**		**2.20**		**16.3**		**0.04**		**23.8**		**0.10**	
MSA	**2.7569**		**0.7692**		**24.263**		**1334.9**		**0.0068**		**2835.6**		**0.0534**	
MSW	**0.0065**		**0.0163**		**0.5257**		**68.253**		**0.0003**		**1484.5**		**0.0026**	
F	**421**		**47.2**		**46.2**		**19.6**		**27.4**		**1.91**		**20.7**	
p	**0.005**		**0.005**		**0.005**		**0.005**		**0.005**		**-**		**0.005**	

Table A.8 Mean and standard deviation of moments for three-sample histograms of α(dB). The correlation coefficient r is with respect to normal distribution. No. of utterances = 5, No. of speakers = 10. Mn = mean, Sd = std. dev., MSA = mean square among speakers, MSW = mean square within speakers. D.F. for F-ratio test: 9, 40. $(F)_{0.005} = 3.22$, $(F)_{0.01} = 2.89$.

Spk. No.	M_1		M_2		M_3		M_4		M_2/M_1		M_4/M_3		r	
	Mn	**Sd**	**Mn**	**Sd**	**Mn**	**Sd**	**Mn**	**Sd**	**Mn**	**Sd**	**Mn**	**Sd**	**Mn**	**Sd**
1	-8.3	0.06	0.42	0.34	1.57	3.03	13.4	21.4	-0.1	0.04	7.56	11.9	0.89	0.14
2	-8.2	0.07	0.48	0.20	1.93	1.06	8.81	4.37	-0.1	0.02	4.72	0.68	0.86	0.10
3	-8.4	0.05	0.44	0.32	1.56	2.90	12.3	21.7	-0.1	0.04	1.59	11.8	0.89	0.14
4	-8.2	0.06	0.67	0.11	1.71	0.65	7.09	3.11	-0.1	0.01	4.08	0.45	0.85	0.05
5	-8.4	0.05	0.51	0.12	1.37	0.97	7.07	6.22	-0.1	0.01	4.93	0.96	0.87	0.05
6	-8.2	0.09	0.66	0.22	1.83	1.14	8.86	5.08	-0.1	0.03	5.23	1.04	0.83	0.08
7	-9.2	0.09	1.55	0.10	-1.4	0.17	3.65	0.63	-0.2	0.01	-2.6	0.15	0.55	0.05
8	-9.0	0.04	1.27	0.06	-1.4	0.14	4.03	0.79	-0.1	0.01	-3.0	0.28	0.62	0.05
9	-8.3	0.13	0.75	0.24	-1.1	0.21	4.53	0.47	-0.1	0.03	-4.2	0.73	0.87	0.05
10	-8.6	0.13	1.01	0.10	-1.1	0.15	3.70	0.83	-0.1	0.01	-3.3	0.40	0.82	0.01
Mean	**-8.5**		**0.78**		**0.50**		**7.34**		**-0.1**		**1.50**		**0.80**	
Sd	**0.35**		**0.38**		**1.51**		**3.52**		**0.04**		**4.37**		**0.12**	
MSA	**0.6249**		**0.7314**		**11.367**		**61.846**		**0.0080**		**95.441**		**0.0713**	
MSW	**0.0068**		**0.0415**		**2.1473**		**102.61**		**0.0006**		**28.388**		**0.0068**	
F	**92.0**		**17.6**		**5.29**		**0.60**		**13.4**		**3.36**		**10.5**	
p	**0.005**		**0.005**		**0.005**		**-**		**0.005**		**0.005**		**0.005**	

REFERENCES

[1] J. P. Campbell, "Speaker recognition: A tutorial," *Proc. IEEE*, vol. 85, pp. 1437-1462, Sept. 1997.

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

[3] D. A. Reynolds, "An overview of automatic speaker recognition technology," in *Proc. IEEE ICASSP*'02, vol. 4, pp. 4072-4075, May 2002.

[4] M. R. Sambur, "Selection of acoustic features for speaker identification," *IEEE Trans. Acoustics*, *Speech, and Signal Processing*, vol. 23, pp. 176-182, Apr. 1975.

[5] J. L. Flanagan, *Speech Analysis Synthesis and Perception*, New York: Springer-Verlag, 1972.

[6] D. A. Reynolds, "Experimental evaluation of features for robust speaker identification," *IEEE Trans. Speech and Audio Processing*, vol. 2, pp. 639-643, Oct. 1994.

[7] K. Yu, J. Mason, and J. Oglesby, "Speaker recognition using hidden Markov models, dynamic time warping and vector quantization," *Proc. IEE on Vision, Image and Signal Processing*, vol. 142, pp. 313-318, Oct. 1995.

[8] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture models," *IEEE Trans. Speech and Audio Processing*, vol. 3, pp. 72-83, Jan. 1995.

[9] B. Gold and N. Morgan, *Speech and Audio Signal Processing*, New York: John Wiley, 2000.

[10] T. Matsui and S. Furui, "Comparison of text-independent speaker recognition methods using VQ-distortion and discrete/continuous HMM's," *IEEE Trans. Speech and Audio Processing*, vol. 2, pp. 456-459, July 1994.

[11] F. K. Soong, A. E. Rosenberg, L. R. Rabiner, and B. H. Juang, "A vector quantization approach to speaker recognition," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 10, pp. 387-390, Apr. 1985.

[12] D. O'Shaughnessy, *Speech Communication – Human and Machine*, New York: Addison-Wesley, 1987.

[13] L. Liu, J. He, and G. Palm, "Signal modeling for speaker identification," in *Proc. IEEE ICASSP*'96, vol. 2, pp. 665-668, May 1996.

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

[15] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*, Englewood Cliffs, New Jersey: Prentice-Hall, 1978.

[16] J. G. Proakis and D. G. Manolakiis, *Digital Signal Processing – Principles, Algorithms, and Applications*, New Delhi: Prentice-Hall, 1996.

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

[18] L. R. Rabiner and B. H. Juang, "An introduction to hidden Markov models," *IEEE ASSP Mag.*, vol. 3, pp. 4-16, Jan. 1986.

[19] R. D. Peacocke and D. H. Graf, "An introduction to speech and speaker recognition," *IEEE Computer Mag.*, vol. 23, pp 26-33, Aug. 1990.

[20] T. F. Quatieri, *Discrete-Time Speech Signal Processing – Principles and Practice*, Singapore: Pearson Education, 2004.

[21] Y. Stylianou, "On the implementation of the harmonic plus noise model for concatenative speech synthesis," *IEEE Trans. Acoustics*, *Speech*, *and Signal Processing*, vol. 2, pp. 11957-11960, June 2000.

[22] P. K. Lehana and P.C. Pandey, "Harmonic plus noise model based speech synthesis in Hindi and pitch modification," in *Proc. 18th International Congress on Acoustics, ICA* 2004, Kyoto, Japan, pp. 3333-3336, Apr. 2004.

[23] Y. Stylianou, "Applying the harmonic plus noise model in concatenative speech synthesis," *IEEE Trans. Speech and Audio Processing*, vol. 9, pp. 21-29, Jan. 2001.

[24] Y. Stylianou, "HNS: Speech modification based on a harmonic + noise model," *IEEE Trans. Acoustics*, *Speech*, *and Signal Processing*, vol. 2, pp. 550-553, Apr. 1993.

[25] D. G. Childers, *Speech Processing and Synthesis Toolboxes*, New York: John Wiley, 2000.

[26] Y. Stylianou, "A pitch and maximum voiced frequency estimation technique adapted to harmonic models of speech," in *Proc. IEEE Nordic Signal Processing Symposium*, Helsinki, Finland, Sept. 1996.

[27] G. W. Snedecor and W. G. Cochran, *Statistical Methods*, Ames, Iowa: Iowa State Univ. Press, 1980.

[28] J. J. Filliben and A. Heckert, "Exploratory data analysis," in *NIST/SEMATECH e-Handbook of Statistical Methods*, Information Technology Laboratory, National Institute of Standards and Technology, U. S. Commerce Dept., Washington, D.C. Available at website http://www.itl.nist.gov/div898/handbook/eda/section3/eda35f.htm, accessed in Oct. 2005.

[29] P. P. Patwardhan / P. Rao (Supervisor), "Low bit rate speech coding," *4th annual PhD progress seminar report* (unpublished), Dept. of Electrical Engineering, IIT Bombay, Aug. 2004.

[30] P. P. Kumar / P. Rao (Supervisor), "Speaker recognition systems," *1st annual PhD progress seminar report* (unpublished), Dept. of Electrical Engineering, IIT Bombay, Aug. 2003.