M.Tech. (Computer Science) Dissertation Series

Speaker Recognition

a dissertation submitted in partial fulfilment of the
requirements for the M.Tech. (Computer Science)
degree of the Indian Statistical Instutute

By

Sulabh Nangalia

under the supervision of

Dr. Mandar Mitra
CVPRU

INDIAN STATISTICAL INSTITUTE
203, Barrackpore Trunk Road
Calcutta - 700 108

Certificate of Approval

This is to certify that this dissertation thesis titled ***"Speaker Recognition"*** submitted by **Mr. Sulabh Nangalia**, in partial fulfilment of the requirements for the M.Tech. (Computer Science) degree of the Indian Statistical Institute, Kolkata, embodies the work done under my supervision.

Mandar Mitra

Dr. Mandar Mitra
CVPRU, ISI,
Kolkata.

Acknowledgements

A dissertation of this type requires a lot of input from people, and I have been fortunate enough to have all the support I needed.

First and foremost, I must acknowledge the real debt of gratitude that I owe to my guide Dr. Mandar Mitra (Assistant Professor, CVPRU) for his incessant support and guidance.

It was a nice experience to collaborate and learn with Dr. Amita Pal (Associate Professor, ASU), Dr. Smarajit Bose (Associate Professor, Stat-Math Unit) and Prof. Debapriya Sengupta (Stat-Math Unit).

I also thank Dr. Anil K. Ghosh (Visiting Scientist, Stat-Math Unit) and Mrs. Madhuri Panda (Project Personnel, ASU) for their valuable advice and help.

Last, but not the least, I should not forget to mention about my friends who have always been with me and made this journey, though a tough one, very pleasent and exiciting.

Sulabh Nangalia

Abstract

We have concentrated on the *speaker identification* part of the speaker recognition problem. Here, we have made a study which involves the classification and identification of the speakers using the Gaussian mixture models (GMM) and the mel frequency cepstral coefficients (MFCC). Due to its reported superior performance, especially under adverse conditions, MFCC is becoming an increasingly popular choice as feature extraction front end to spoken language systems. The individual Gaussian components of a GMM are shown to represent some general speaker-dependent spectral shapes that are effective for modelling speaker identity. A complete experimental evaluation is conducted on two sets of data of 7 speakers and 21 speakers. The GMM attains 100% accuracy on the 7 speaker data and 97.3% on the 21 speaker data using clean speech utterances.

Contents

List of Figures

List of Tables

Chapter 1

An Introduction to Speaker Recognition

The recent development of technology has raised the interest in science fiction inspired biometric recognition i.e., recognition based on an individual's biological features. Numerous measurements and signals have been proposed and investigated for use in biometric recognition systems [JAI99] such as fingerprints, retinal scan, face, written signature, DNA-analysis, smell and voice. Among the most popular measurements are fingerprint, face, and voice. Perhaps the greatest advantage of biometric recognition is that you can forget a PIN-code, but you will never forget your body. Moreover, if the biometric properties are unique then recognition could be rather safe provided the technology can measure these properties accurately.

1.1 Voice as a Biometric

The speech signal conveys several levels of information. Primarily, the speech signal conveys the words or messages being spoken, but on a secondary level, the signal also conveys information about the identity of the speaker. How?

Speech is a complicated signal produced as a result of several transformations occurring at several different levels: semantic, linguistic, articulatory, and acoustic. Differences in these transformations appear as differences in the acoustic properties of the speech signal. Speaker-related differences are a result of a combination of anatomical differences inherent in the vocal tract and the learned speaking habits of different individuals. In speaker recognition, all these differences can be used to discriminate between speakers.

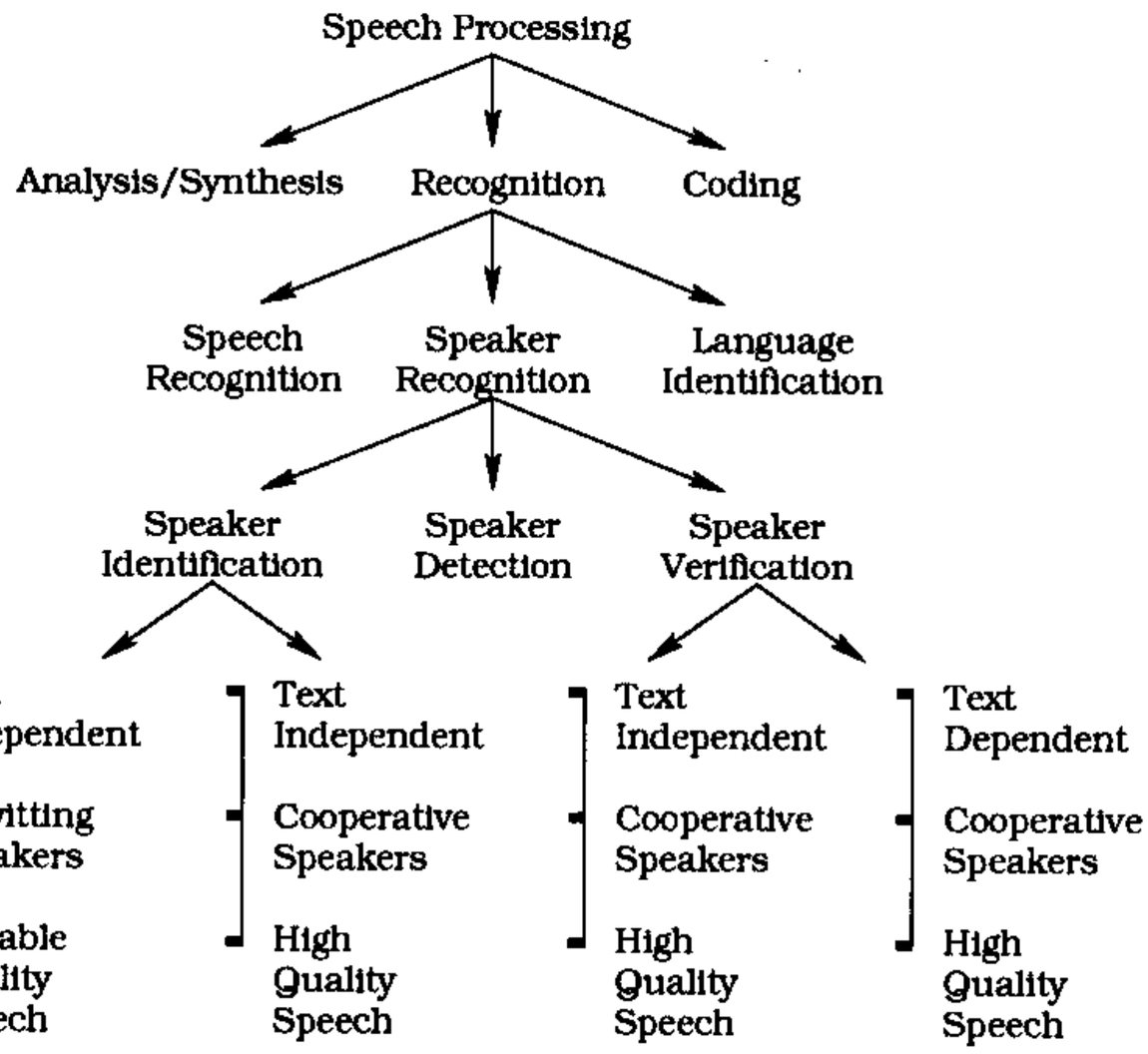

Figure 1.1: Speech Processing

1.2 Automatic Speaker Recognition

Speech processing is a diverse field with many applications. Figure 1.1 shows a few of these areas and how speaker recognition relates to the rest of the field.

Automatic speaker recognition is the use of a machine to recognize a person from a spoken phrase. These systems can operate in two modes: *speaker identification* or *speaker verification.* Speaker recognition methods can also be divided into *text-dependent*, *text-independent* and *text-prompted* methods. Text-prompted method is a nothing but a special case of text-dependent method.

A speaker known to the speaker recognition system who is correctly claiming his/her identity is labeled as *claimant* and a speaker unknown to the system who is posing as a known speaker is labeled as *imposter.* A known speaker is also reffered to as a *target speaker*, while an imposter is alternately called a *background speaker.*

There are two types of errors in speaker recognition systems: false acceptances, where an imposter is accepted as a claimant, and false rejections, where claimants are rejected as imposters. There are several factors that can contribute to these errors. Table 1.1 lists some of the human and environmental factors that contribute to these errors.

- Mis-spoken or misread prompted phrases
- Extreme emotional states (e.g., stress or duress)
- Time varying microphone placement
- Poor or inconsistent room acoustics (e.g., multipath and noise)
- Channel mismatch (e.g., using different microphones for enrollment and recognition)
- Sickness (e.g., head cold can alter the vocal tract)
- Aging (the vocal tract can drift away from models with age)

Table 1.1: Sources of error in speaker recognition

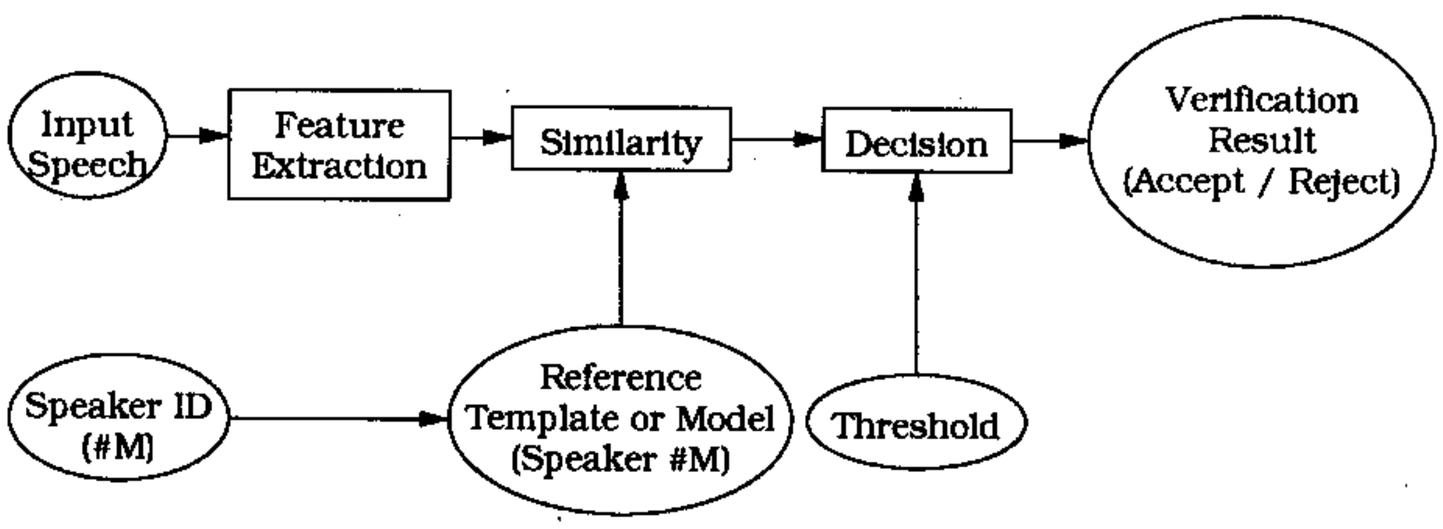

Figure 1.2: Speaker Verification

1.2.1 Speaker Verification

Speaker verification is defined as deciding if a speaker is whom he/she claims to be. The literature abounds with different terms for speaker verification, including voice verification, speaker authentication, voice authentication, talker authentication, and talker verification. Figure 1.2 shows the basic structure of a speaker verification system.

1.2.2 Speaker Identification

In speaker identification, there is no *a priori* identity claim, and the system decides who the person is, what group the person is a member of, or that the person is unknown in case of an open set. In open set identification, the

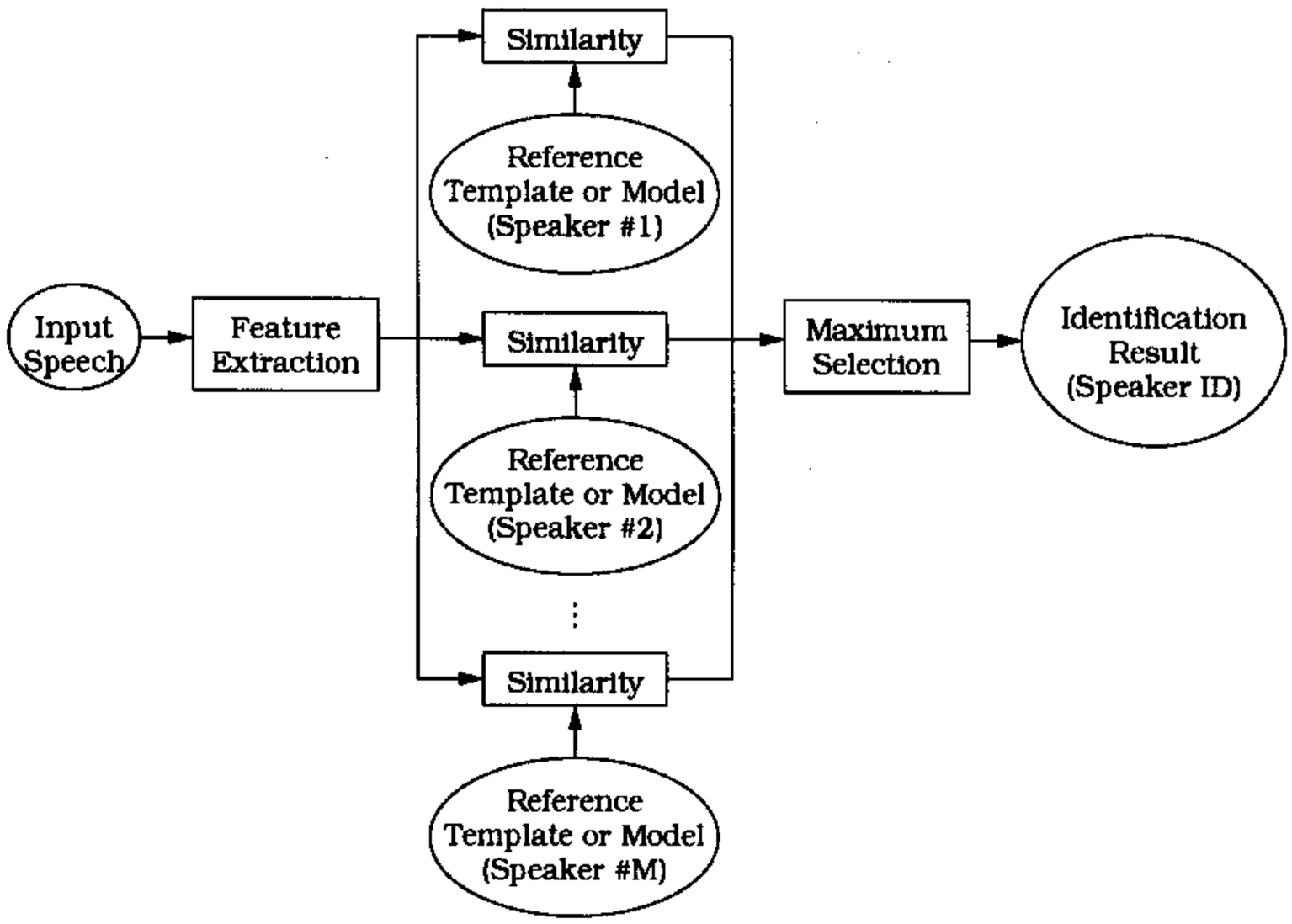

Figure 1.3: Speaker Identification

reference model for an unknown speaker may not exist. This is usually the case in forensic applications. In this situation, an additional decision alternative, *the unknown does not match any of the models*, is required. In both verification and identification processes, an additional threshold test can be used to determine if the match is close enough to accept the decision or if more speech data are needed. Figure 1.3 shows the basic structure of a speaker identification system.

1.2.3 Text Dependent System

In text dependent systems, the users are required to pronounce the test sentence that contains the same text or vocabulary as the training sentences. It is done to reduce the intraspeaker variability. Here, the knowledge of knowing words or word sequence can be exploited to improve the performance.

1.2.4 Text Independent System

In text independent systems, there is no such restriction on the user to pronounce the same text as that pronounced during the enrollment.

1.2.5 Text Prompted System

In the text-prompted speaker recognition method, the recognition system prompts each user with a new key sentence every time the system is used and accepts the input utterance only when it decides that it was the registered speaker who repeated the prompted sentence. The sentence can be displayed as characters or spoken by a synthesized voice. Because the vocabulary is unlimited, prospective impostors cannot know in advance what sentence will be requested. Not only can this method accurately recognize speakers, but it can also reject utterances whose text differs from the prompted text, even if it is spoken by the registered speaker. A recorded voice can thus be correctly rejected.

1.3 Why Speaker Recognition?

As speech interaction with computers becomes more pervasive in activities, the utility of automatically recognizing a speaker based solely on vocal characretistics increases.

While each biometric has pros and cons relative to accuracy and deployment, there are two main factors that have made voice a compelling biometric.

First, speech is a natural signal to produce that is not considered threatening by users to provide. In many applications, speech may be the main (or only, e.g., telephone transactions) modality, so users do not consider providing a speech sample for authentication as a separate or intrusive step.

Second, the telephone system provides a ubiquitous, familiar network of sensors for obtaining and delivering the speech signal. For telephone based applications, there is no need for special signal transducers or networks to be installed at application access points since a cell phone gives one access almost anywhere. Even for non-telephone applications, sound cards and microphones are low-cost and readily available.

Additionally, speaker recognition has found its way into a large number of applications such as voice dialing, phone banking, telephone shopping, database access services, voice mail, security control of confidential information, remote access to computers, information services and forensics.

1.4 Outline of Report

This dissertation work concentrates on the text-independent speaker identification field of the speaker recognition problem. Chapter 2 deals with the

production and classification of the speech signals. The speech analysis for extracting the feature representation used in this work is presented in Chapter 3. Next, the reference model of a speaker used for speaker identification is described in Chapter 4. The experimental evaluation and the analysis of the results obtained is done in Chapter 5 and finally we conclude with Chapter 6.

Chapter 2

The Speech Signal

2.1 Speech Production

Speech signals are composed of a sequence of sounds. These sounds and the transitions between them serve as a symbolic representation of information. The arrangement of these sounds (symbols) is governed by the rules of language. The study of these rules and their implications in human communication is the domain of *linguistics*, and the study and classification of the sounds of speech is called *phonetics*.

2.1.1 The Mechanism of Speech Production

A schematic diagram of the human vocal mechanism is shown in Figure 2.1 [FLA72]. Air enters the lungs via the normal breathig mechanism. As air is expelled from the lungs through the *trachea* (or windpipe), the tensed vocal cords within the *larynx* are caused to vibrate (in the mode of a relaxation oscillator) by the air flow. The air flow is chopped into quasi-periodic pulses which are then modulated in frequency in passing through the *pharynx* (the throat cavity), the mouth cavity, and possibly the nasal cavity. depending on the positions of the various articulators (i.e., jaw, tongue, velum, lip, mouth), different sounds are produced.

A simplified representation of the complete physiological mechanism for creating speech is shown in Figure 2.2 [RAB03]. The human vocal mechanism is driven by an excitation source, which also contains speaker-dependent information. The lungs and the associated muscles act as the source of air for exciting the vocal mechanism. The muscle force pushes air out of the lungs (shown schematically as a piston pushing up within a cylinder) and through the bronchi and trachea. The excitation can be characterized as phonation,

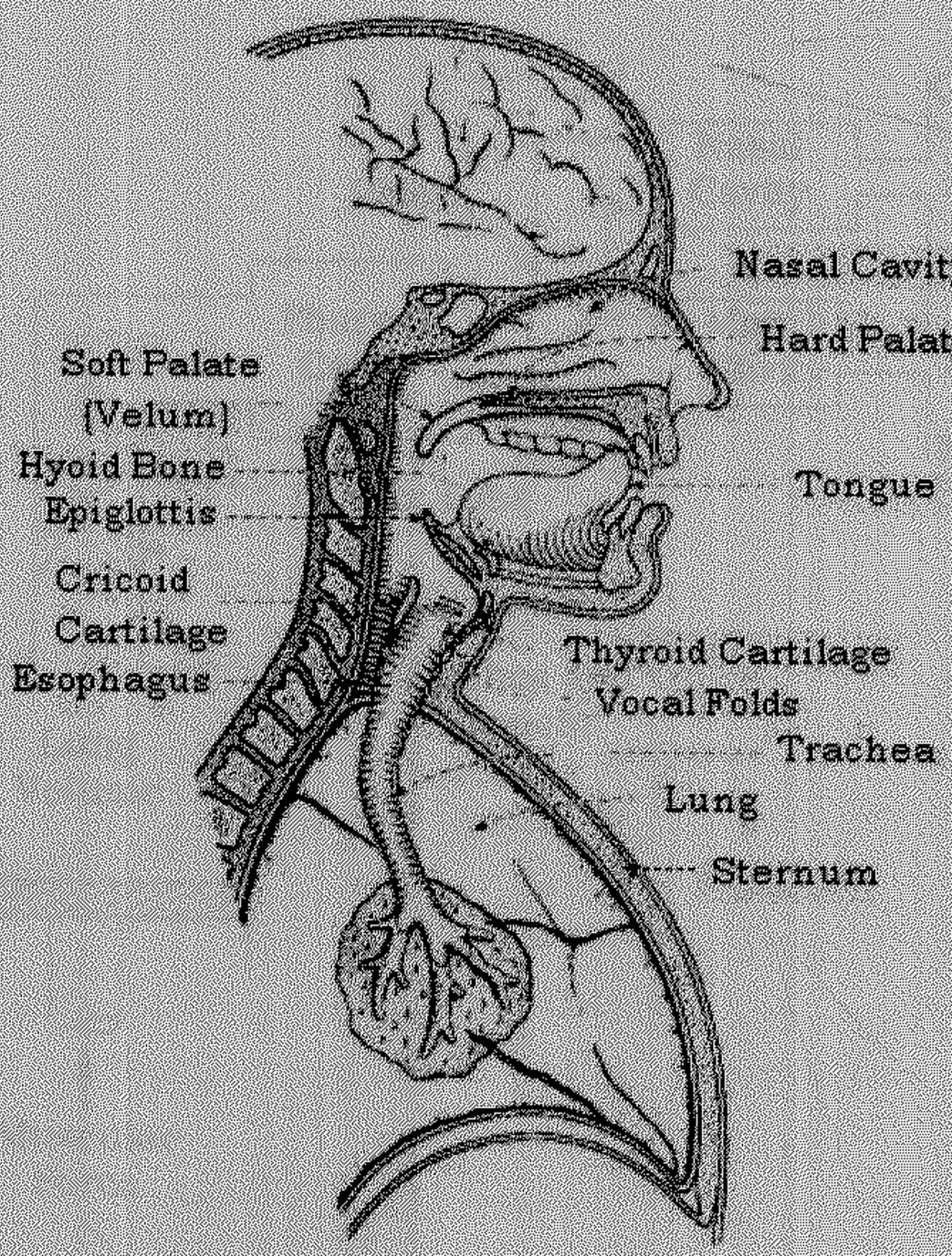

Figure 2.1: Human vocal system

whispering, frication, compression, vibration, or a combination of these. Details of these can be found in [QUA04, RAB03, CAM74].

There are two main sources of speaker-specific characteristics of speech: physical and learned.

The vocal tract is generally considered as the speech production organs above the vocal folds. It comprises of the oral cavity from the larynx to the lips and the nasal passage that is coupled to the oral tract by by way of the velum. Vocal tract shape is an important physical distinguishing factor of speech. As the acoustic wave passes through the vocal tract, its frequency content (spectrum) is altered by the resonances of the vocal tract. Vocal tract resonances are called *formants*. Thus, the vocal tract shape can be estimated from the spectral shape (e.g., formant location and spectral tilt) of the voice signal. Voice verification systems typically use features derived only from the vocal tract.

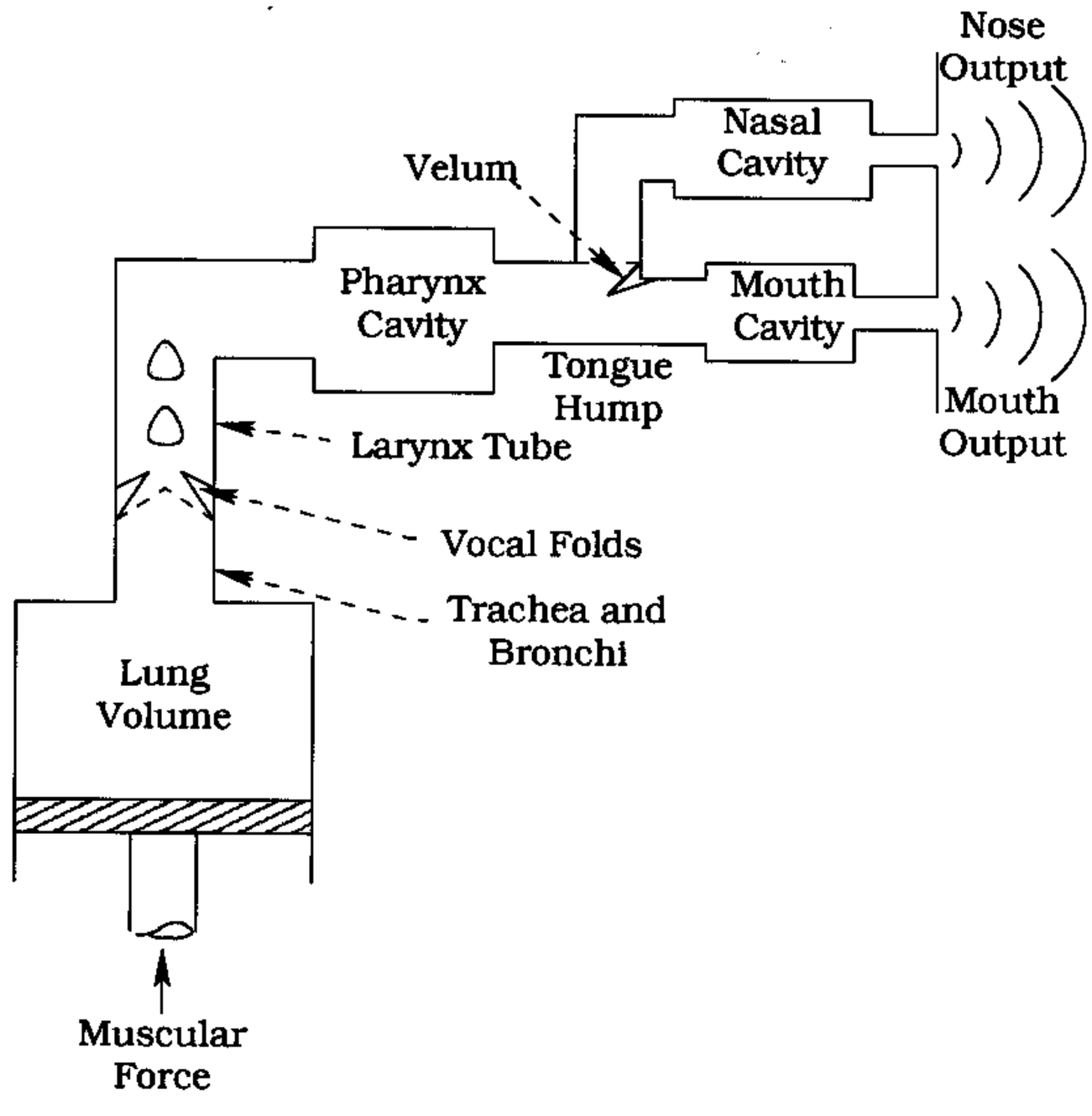

Figure 2.2: Schematic representation of the complete physiological mechanism of speech production.

Other aspects of speech production are learned characteristics, including speaking rate, prosodic effects[1] , and dialect (which might be captured spectrally as a systematic shift in formant frequencies).

2.2 Classification of Speech Sounds

Most languages can be described in terms of a set of distinctive sounds, or *phonemes.* There are a variety of ways of studying phonetics; e.g., linguists study the distinctive features or charecteristics of the phonemes.

The four broad classes of sounds are *vowels*, *dipthongs*, *semivowels* and *consonants* [RAB04]. Each of these classes may be further broken down into

[1]Long-time variations, i.e., changes extending over more than one phoneme, in pitch (intonation), amplitude (loudness) and timing (articulation rate or rhythm). See [QUA04].

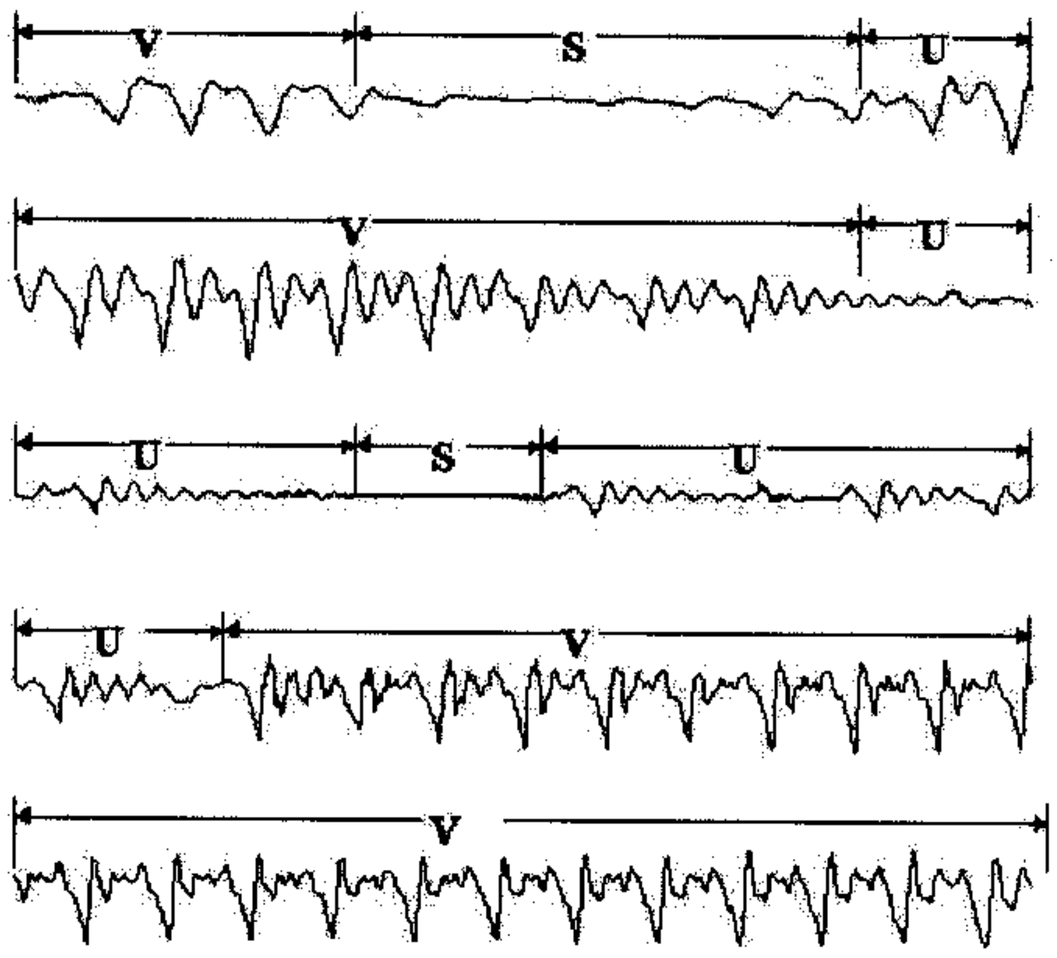

Figure 2.3: Waveform of an utterance. Each line corresponds to 100 msec. S, U and V stand for silence, unvoiced and voiced respectively.

sub-classes which are related to the manner, and place of articulation of the sound within the vocal tract. This classification of sound is generally used in the fields of speech recognition, language identification and speech synthesis.

Speech sounds can also be classified into 3 distinct classes according to their mode of excitation [RAB04] (shown in Figure 2.3). These mechanism of excitation are:

1. Air flow from the lungs is modulated by the vocal cord vibration, resulting in a quasi-periodic pulse-like exciattion. These are called *voiced sounds.*

2. Air flow from the lungs become turbulent as the air passes through a constriction in the vocal tract, resulting in a noise-like excitation. These are called *fricatives* or *unvoiced sounds.*

3. Air flow builds up pressure behind a point of total closure in the vocal tract. The rapid release of this pressure, by removing the constriction, causes a transient excitation. These are called *plosive sounds.*

In Chapter 3, we will see how we can use these attributes for the extraction of features relevant for speaker identification.

Chapter 3

Feature Extraction

Chapter 2 shows a variety of voice attributes that characterize a speaker. In viewing these attributes, both from the perspective of the human and the machine for recognition, speaker-dependent voice characteristics can be categorized as "high-level" and "low-level." High-level voice attributes include "clarity," "roughness," "magnitude" and "animation" [REY92, VOI64]. Other high-level attributes are prosody and dialect. These attributes can be difficult to extract by machine for automatic speaker recognition. In contrast, low-level attributes, being of an accoustic nature, are more measurable. These attributes include primarily the vocal tract spectrum and, to a lesser extent, instantaneous pitch and glottal flow excitation, as well as temporal properties such as source event onset times and modulations in formant trajectories.

We focus on features (derived from spectral measurements) identifying the formants in the speech, which represent the changes in the vocal tract of a speaker, and *mel-frequency cepstrum coefficients* (MFCC) is one such example we have used for speaker identification.

3.1 Mel Scale and MFCC

A pure tone is uniquely defined by its intensity and frequency. The perceptual counterparts of these quantities are termed loudness and pitch respectively. Pitch is difficult to define. Mostly we agree that pure tones can be ordered in such a way that one tone is higher or lower than another. Pitch is the criterion that we use to make such decisions. Like loudness, it is a complex, non-linear function of both frequency and intensity. Stevens, Volkmann and Newman defined the Mel (*melody*) scale, which relates pitch to frequency as depicted in Fig. 3.1. It was later refined by Stevens and Volkmann in their

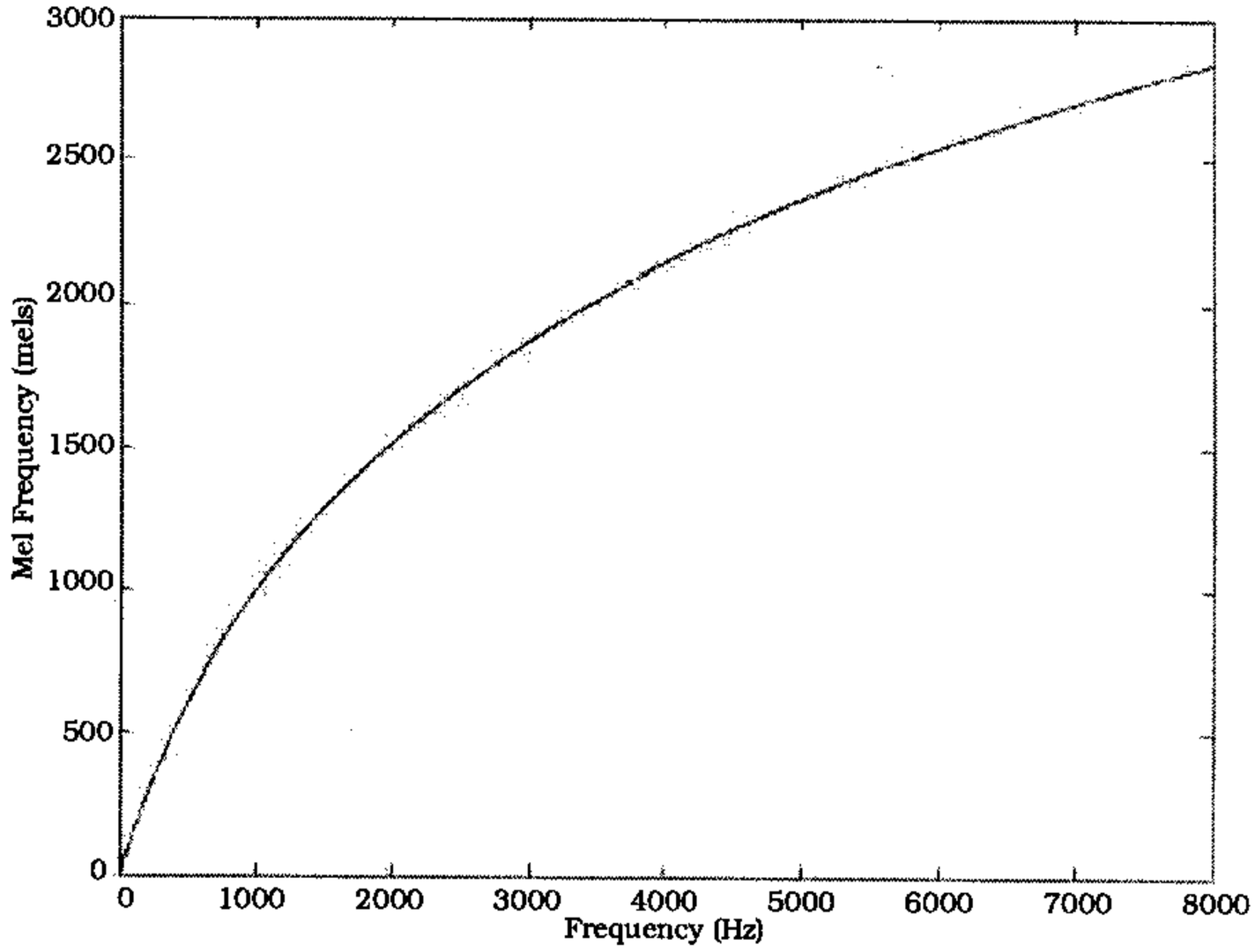

Figure 3.1: Linear frequency vs. Mel frequency

classical paper [STE40]. The form of the curve was determined by perceptual experiments designed to find a linear relation among perceived pitches. A pitch of 2000 mels is therefore subjectively twice as high as a pitch of 1000 mels. The numeric range of the mel scale and its relation to sound intensity was fixed by defining a 40dB tone with a frequency of 1000 Hz as having a pitch of 1000 mels. 1 mel represents one-thousandth of the pitch of 1 kHz.

MFCC features are derived from the FFT [OPP98] magnitude spectrum by applying a filter bank which has filters evenly spaced on a warped frequency scale. The logarithm of the energy in each filter is calculated and accumulated before a Discrete Cosine Transform (DCT) [OPP98] is applied to produce the MFCC feature vector. The frequency warping scale used for filter spacing in MFCC is the mel scale.

3.2 Pre Processing of Speech Signal

Firstly, the amplitude of the input signal is normalized to remove the effect of varying intensity. The speech signal as shown in Figure 2.3 consists of silence, voiced and unvoiced part. The silence part does not contain any relevant infor-

mation about the speaker. Thus we need to remove the silence part before extracting any features from the speech. We used the short-term energy [RAB04] to remove the silence part which is described as follows.

Short-term energy allows us to calculate the amount of energy in a sound at a specific instance in time, and is defined as

$$E[n] = \sum_{m=n-N+1}^{n} (x[m]\, w[n-m])^2 \tag{3.1}$$

where x is the speech signal, w is the window function, n is the sample that the window is centered on, and N is the window size.

We used Hamming window [OPP98] which is given as

$$w[n] = \begin{cases} \alpha - (1-\alpha)\cos\frac{2\pi n}{N-1} & 0 \le n \le N-1,\ \alpha = 0.54 \\ 0 & \text{otherwise} \end{cases} \tag{3.2}$$

where N is the window size.

We segmented the speech signal into several non overlappinp parts of fixed size and calculated the short-term energy of every segment. The segment having energy less than a fixed threshold value was discarded for further consideration. This removes the silence, weak fricatives and weak plosives from the speech whose energies are approximately same as that of the silence part.

3.3 Computing MFCC

The block diagram for computing MFCC is shown in Figure 3.2. Let the N-sample speech signal be x[n], n=0,1,...,N-1. The step-by-step procedure for computing MFCC is explained below [PHA, COM].

3.3.1 Pre Emphasis

The spectral characteristics of speech at higher frequencies are subdued in relation to the lower frequencies. In order to enhance their weight in the extracted parameters, a pre-emphasis filter is applied which is a first order high-pass filter described by:

$$y[n] = x[n] - ax[n-1], \quad 0.9 \le a \le 1.0 \tag{3.3}$$

where the parameter a is not critical. We have taken $a = 0.95$.

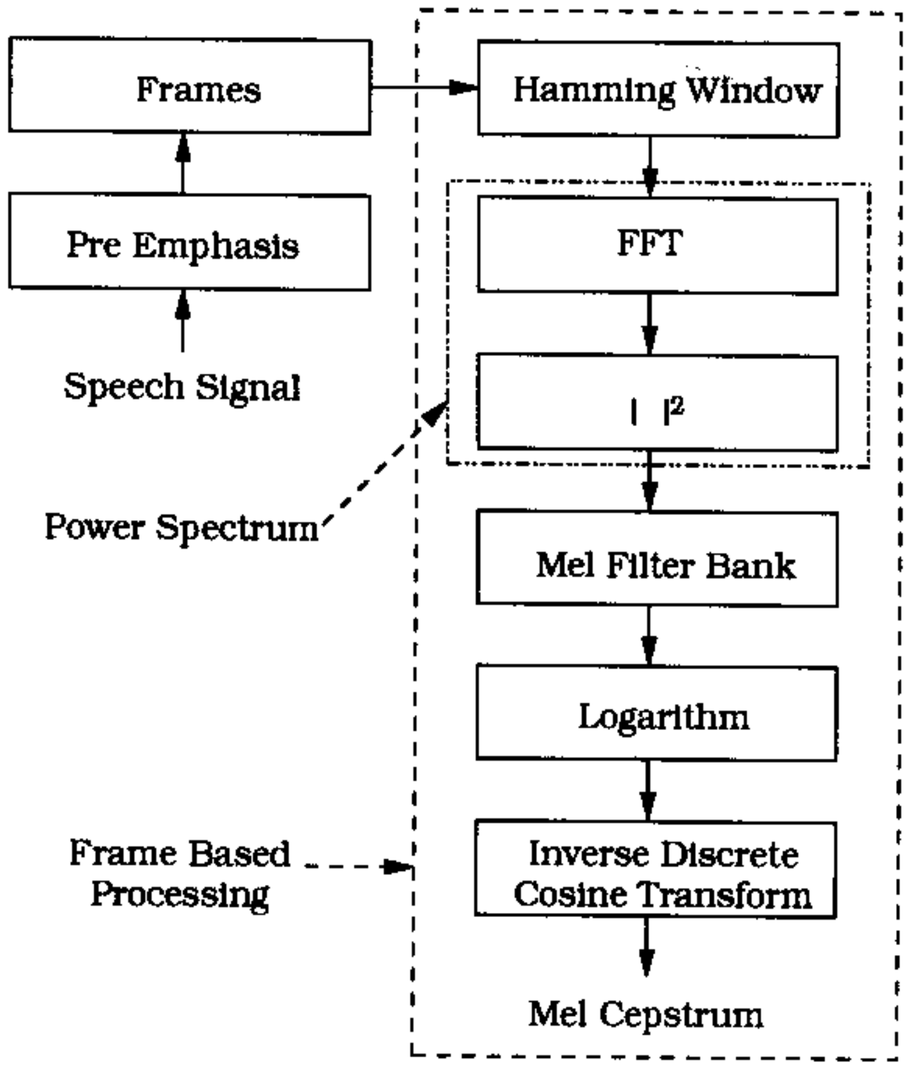

Figure 3.2: Block diagram for computing MFCC

3.3.2 Framing

The speech signal is a slowly time varying signal in the sense that, when examined over a sufficiently short period of time (between 5 and 100 msec), its characteristics are fairly stationary. An illustration of this effect can be seen in Figure 2.3. Thus, we can isolate and process short segments of the speech signal as if they were short segments from a sustained sound with fixed properties.

So, the normalized and filtered signal is broken into M overlapping frames with step size V and frame size W. Frame size usually range from 10 msec to 20 msec. and step size between 20 and 50 percent of the frame size.

3.3.3 Windowing

Each frame is multiplied with a window function w to minimize signal discontunities in the time domain and the resulting spectral artifacts due to *picket-fences* effect.

$$y[n] = y[n]\, w[n], \quad n = 0, 1, \ldots, W-1 \tag{3.4}$$

We have used the Hamming window given by Eq. 3.2.

3.3.4 Power Spectrum

The power spectrum of each frame is calculated by calculating a discrete fourier transform (DFT) [OPP98] of specified size U and then computing its magnitude squared. The frame size W may not be the same as U, and to overcome this problem the frame could be zero padded or readjusted according to U. The DFT is given by

$$Y[k] = \sum_{n=0}^{U-1} y[n]\, e^{-j(2\pi/U)}, \quad k = 0, 1, \ldots, U-1 \tag{3.5}$$

where $j = \sqrt{-1}$ and the magnitude square is given by

$$S[k] = (\text{real}(Y[k]))^2 + (\text{imag}(Y[k]))^2, \quad k = 0, 1, \ldots, U-1 \tag{3.6}$$

The DFT can be computed efficiently using the FFT [OPP98] algorithm.

3.3.5 Mel Spectrum

The mel spectrum of the power spectrum is computed by multiplying the power spectrum by each of the of the triangular mel weighting filters (described later in this section) and integrating the result.

$$\tilde{S}[l] = \sum_{k=0}^{U-1} S[k] m_l[k], \quad l = 0, 1, \ldots, L-1 \tag{3.7}$$

where L is the total number of triangular mel weighting filters and m_l is the l^{th} filter.

Mel Filter Bank

The mel scale filterbank is a series of L triangular bandpass filters (Figure 3.3) that have been designed to simulate the bandpass filtering believed to occur in the auditory system. This corresponds to series of bandpass filters with constant bandwidth and spacing on a mel frequency scale. On a linear frequency scale, this filter spacing is appoximately linear up to 1 kHz and logarithmic at higher frequencies. The following warping function tranforms linear frequencies to mel frequencies:

$$\text{mel}(f) = 2595 \log_{10}\left(1 + \frac{f}{700}\right) \tag{3.8}$$

If f_{min} and f_{max} are the minimum and maximum frequencies in Hz, respectively, f_c the centre frequency of a filter, and the low and high cutoff frequencies, f_l

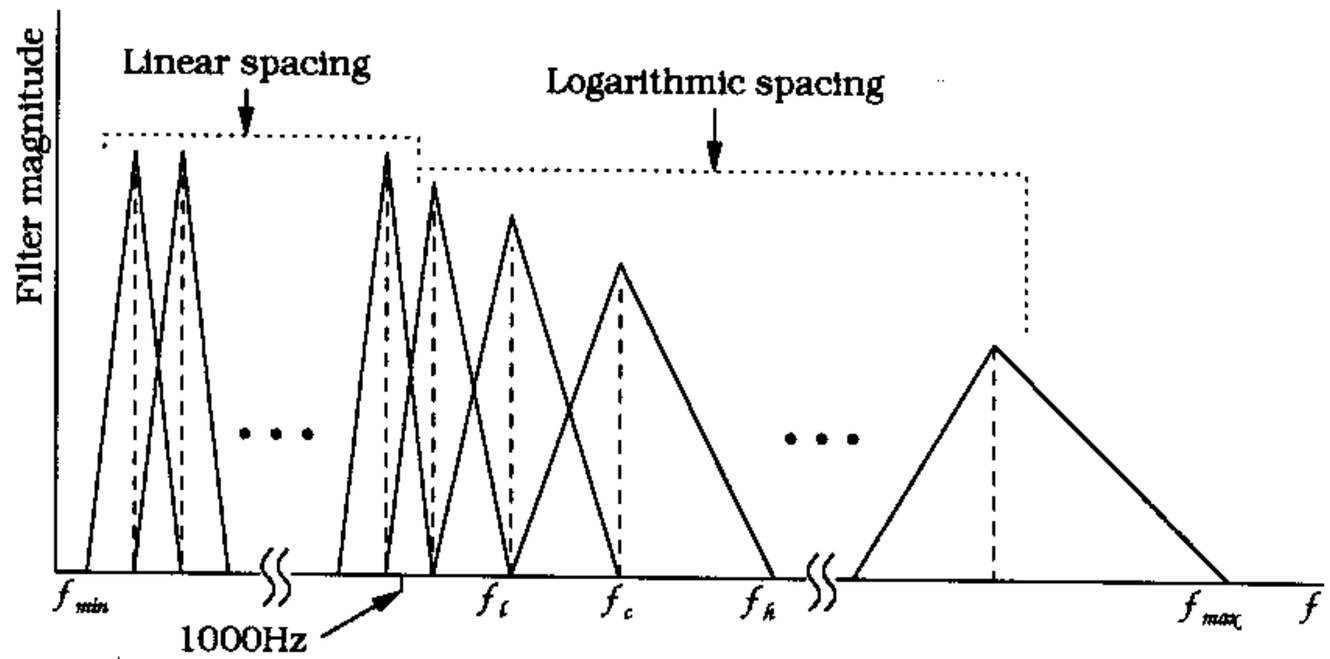

Figure 3.3: The mel scale filter bank

and f_h, as the centre frequencies of the two adjacent filters (Figure 3.3), then the mel filter bank m is defined by the following equations.

$$\gamma = \frac{W-1}{f_{\max} - f_{\min}} \tag{3.9}$$

$$\begin{aligned} I_l &= \gamma(f_l - f_{\min}) \\ I_c &= \gamma(f_c - f_{\min}) \\ I_r &= \gamma(f_h - f_{\min}) \end{aligned} \tag{3.10}$$

$$m_i[k] = \begin{cases} \frac{1}{f_c - f_l}\left(\frac{k}{\gamma} + f_{\min} - f_l\right) & \text{if } \lceil I_l \rceil \le k \le \lfloor I_c \rfloor \\ 1 + \frac{1}{f_c - f_h}\left(\frac{k}{\gamma} + f_{\min} - f_l\right) & \text{if } \lceil I_c \rceil \le k \le \lfloor I_r \rfloor \\ 0 & \text{otherwise} \end{cases} \tag{3.11}$$

for $i = 0, 1, \ldots, L-1$.

$f_{\min}$ is generally taken as 20 Hz (0 mels) and $f_{\max}$ as half the sampling frequency. The centre frequencies f_c can be determined by dividing the mel frequency scale, between the minimum mel frequency and the maximum mel frequency ($[f_{\min}, f_{\max}]$), into L equal parts and then applying Eq. 3.8 to get f as f_c. The number of filters is usually between 13 and 24. We have used 20 filters.

3.3.6 Log of Filter Coefficients

This part of the process is completed by taking the logarithm of each filter coefficient to model the non-linear intensity-loudness relationship which is log-

arithmic in nature.

$$\tilde{S}[l] = \log_{10} \tilde{S}[l], \quad l = 0, 1, \ldots, L-1 \tag{3.12}$$

3.3.7 Inverse Discrete Cosine Transform

The inverse discrete cosine transform (IDCT) [OPP98] is used to orthogonalize the filter vectors. Because of this orthogonalization step, the information of the filter vector is compacted into the first number of components and we can shorten the vector to C components. The IDCT of the filter vector gives the mel cepstrum c as

$$c[n] = \sqrt{\frac{2}{L}} \sum_{l=0}^{L-1} \tilde{S}[l] \cos\left((l - 0.5)\frac{\pi n}{L}\right), \quad n = 0, 1, \ldots, C-1 \tag{3.13}$$

where C is the desired number of cepstral coefficients. C is chosen to be less than L, usually somewhere betwee 9 and 15. We have used 13 mel cepstral coefficients. The first mel-scaled cepstral coefficient is dropped as it represents the mean energy in each frame.

3.4 Why MFCC?

The mel warping transforms the frequency scale to place less emphasis on high frequencies. The primary reason for effectiveness of MFCC is that, it models the non-linear auditory response of the human ear which resolves frequencies on a log scale. The mel cepstrum can be considered as the spectrum of the log spectrum. Removing its mean reduces the effects of linear time-invariant filtering (e.g., channel distortion). The cepstrum's density has the benefit of being modeled well by a linear combination of Gaussian densities as used in the Gaussian mixture model.

Chapter 4

Pattern Matching and Classification

The pattern-matching task involves computing a match score, which is a measure of the similarity, of the input feature vectors to some model. Speaker models are constructed from the features extracted from the speech signal. To enroll users into the system, a model of the voice, based on the extracted features, is generated and stored. Then, the matching algorithm compares/scores the incoming speech signal with the model of the claimed user. There are two types of models: *stochastic models* and *template models.*

In stochastic models, the pattern matching is probabilistic and results in a measure of the likelihood, or conditional probability, of the observation given the model. For template models, the pattern matching is deterministic. The observation is assumed to be an imperfect replica of the template, and the alignment of observed frames to template frames is selected to minimize a distance measure d.

We have used Gaussian mixture model (GMM) (Section 4.1) which is an example of the stochastic model in this dissertation work to construct the speaker models for the speaker identification problem.

4.1 Gaussian Mixture Model

The Gaussian probability density function in one dimension is a bell shaped curve defined by two parameters, mean μ and variance σ^2. In the D-dimensional space it is defined in a matrix form as

$$\mathcal{N}(\vec{x}; \vec{\mu}, \Sigma) = \frac{1}{(2\pi)^{D/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(\vec{x} - \vec{\mu})'\Sigma^{-1}(\vec{x} - \vec{\mu})\right] \qquad (4.1)$$

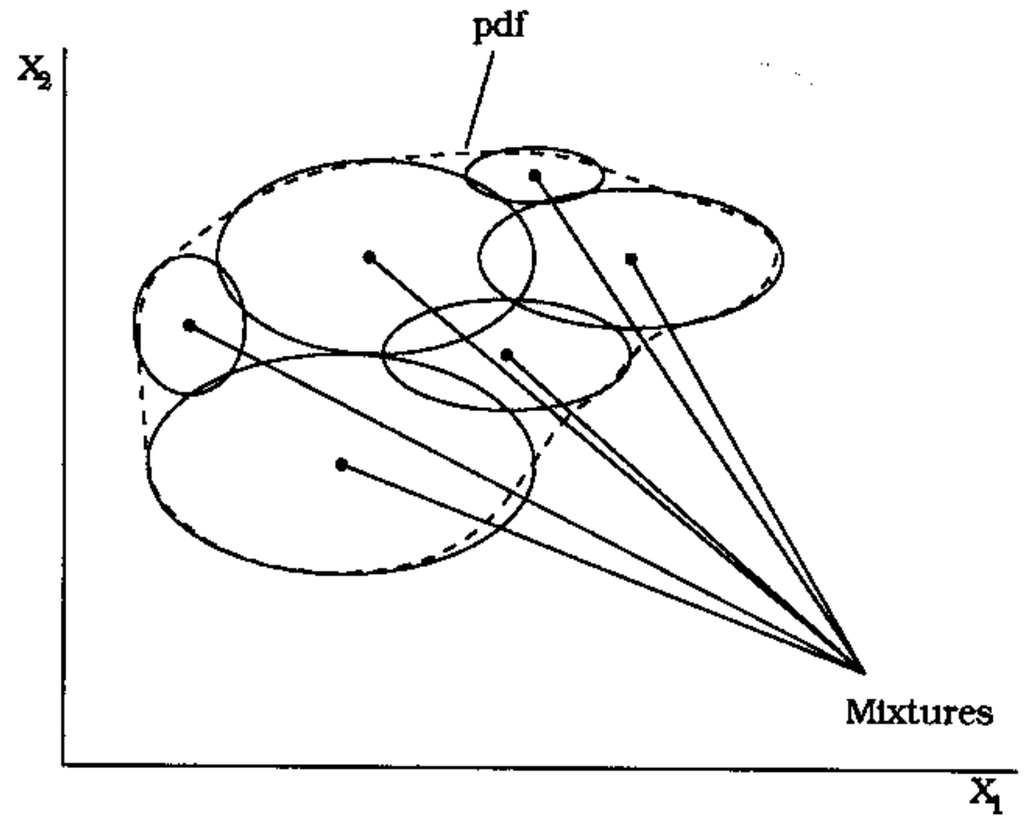

Figure 4.1: The Gaussian mixture model is a union of Gaussian pdfs.

where $\vec{x}$ is a D-dimensional random vector, $\vec{\mu}$ is the mean vector and Σ the covariance matrix. The Gaussian distribution is usually quite good approximation for a class model shape in a suitably selected feature space. It is a mathematically sound function and extends easily to multiple dimensions. In the Gaussian distribution lies an assumption that the class model is truly a model of one basic class. If the actual model, the actual probability density function, is multimodal, it fails.

Gaussian mixture model (GMM) shown in Figure 4.1 is a mixture of several Gaussian distributions and can therefore represent different subclasses inside one class. The probability density function is defined as a weighted sum of Gaussians

$$p(\vec{x}|\lambda) = \sum_{c=1}^{M} p_c \mathcal{N}(\vec{x}; \vec{\mu}_c, \Sigma_c) \tag{4.2}$$

where p_c is the weight of the component c, $0 < p_c < 1$ for all components, and $\sum_{c=1}^{M} p_c = 1$. Thus the complete Gaussian mixture density is parameterized by the mean vectors, covariance matrices and mixture weights from all component densities. These parameters are collectively represented by the notation

$$\lambda = \{p_c, \vec{\mu}_c, \Sigma_c\} \quad c = 1, \ldots, M \tag{4.3}$$

Given data set $\mathcal{X} = \{\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_T\}$ of size T, the aim is to estimate the parameters of the GMM, λ, which in some sense best matches the distribution of the training feature vectors. There are several techniques, available in the

literature, for estimating the parameters of a GMM [DUD01, LAC88]. By far the most popular and well-established method is the *maximum likelihood (ML) estimation*. This method is described in section 4.1.1

4.1.1 Maximum Likelihood Parameter Estimation

Given a density function $p(\vec{x}|\lambda)$ that is governed by the set of parameters λ, and a data set $\mathcal{X} = \{\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_T\}$ of T independent samples drawn from this distribution. That is, these data vectors are independent and identically distributed (i.i.d.) with distribution p. Therefore, the resulting density for the samples is

$$p(\mathcal{X}|\lambda) = \prod_{t=1}^{T} p(\vec{x}_t|\lambda) = \mathcal{L}(\mathcal{X};\lambda) \tag{4.4}$$

The function, $\mathcal{L}(\mathcal{X};\lambda)$, is called the *likelihood* of the parameters given the data, or just the likelihood function. The likelihood is thought of as a function of the parameters λ where the data $\mathcal{X}$ is fixed. In the maximum likelihood problem, the goal is to find the λ that maximizes $\mathcal{L}$, i.e., to find λ^* where

$$\lambda^* = \arg\max_{\lambda} \mathcal{L}(\mathcal{X};\lambda) \tag{4.5}$$

Unfortunately, this expression is nonlinear function of the parameters λ and direct maximization is not possible. Usually this function is not maximized directly but the logarithm

$$L(\mathcal{X};\lambda) = \log \mathcal{L}(\mathcal{X};\lambda) = \sum_{t=1}^{T} \log p(\vec{x}_t|\lambda) \tag{4.6}$$

called the *log-likelihood* function is taken which is analytically easier to handle. Because of the monotonicity of the logarithm function the solution to Eq. 4.5 is the same using $\mathcal{L}(\mathcal{X};\lambda)$ or $L(\mathcal{X};\lambda)$.

Depending on $p(\vec{x}|\lambda)$ it might be possible to find the maximum analytically by setting the derivatives of the log-likelihood function to zero and solving λ. But usually the analytical approach is intractable. In practice an iterative method such as the *expectation maximization* (EM) algorithm [DEM77, DUD01, BIL97] is used. This algorithm is described in section 4.1.2 and has been used for estimating the parameters for the speaker recognition problem.

4.1.2 The EM Algorithm for GMM

The expectation maximization (EM) algorithm is an iterative method for calculating maximum likelihood distribution parameter estimates from incomplete data (some elements missing in some feature vectors). It can also be

used to handle cases where an analytical approach for maximum likelihood estimation is infeasible, such as Gaussian mixtures with unknown and unrestricted covariance matrices and means.

The EM algorithm starts from an initial guess λ^0 for the distribution parameters and the log-likelihood is guaranteed to increase on each iteration until it converges. The convergence leads to a local or global maximum.

On each EM iteration, the following reestimation formulas are used which guarantee a monotonic increase in the model's likelihood value.

$$p_c^{i+1} = \frac{1}{T}\sum_{t=1}^{T} p(c|\vec{x}_t, \lambda^i) \tag{4.7}$$

$$\vec{\mu}_c^{i+1} = \frac{\sum_{t=1}^{T} p(c|\vec{x}_t, \lambda^i)\,\vec{x}_t}{\sum_{t=1}^{T} p(c|\vec{x}_t, \lambda^i)} \tag{4.8}$$

$$\Sigma_c^{i+1} = \frac{\sum_{t=1}^{T} p(c|\vec{x}_t, \lambda^i)\,\vec{x}_t\vec{x}_t'}{\sum_{t=1}^{T} p(c|\vec{x}_t, \lambda^i)} - \mu_c^{i+1}(\mu_c^{i+1})' \tag{4.9}$$

where the *a posteriori* probability for class c is given by

$$p(c|\vec{x}_t, \lambda^i) = \frac{p_c^i \mathcal{N}(\vec{x}_t; \vec{\mu}_c^i, \Sigma_c^i)}{\sum_{k=1}^{M} p_k^i \mathcal{N}(\vec{x}_t; \vec{\mu}_k^i, \Sigma_k^i)} \tag{4.10}$$

for $c = 1, 2, \ldots, M$. The superscripts i and $i+1$ denote the i^{th} and $(i+1)^{th}$ iterations respectively.

The interpretation of the Eqs. 4.7-4.9 is actually quite intuitive. The weight p_c of a component is the portion of samples belonging to that component. It is computed by approximating the component-conditional pdf with the previous parameter estimates and taking the posterior probability of each sample point belonging to the component c (Eq. 4.10). The component mean $\vec{\mu}_c$ and covariance matrix Σ_c are estimated in the same way. The samples are weighted with their probabilities of belonging to the component, and then the sample mean and sample covariance matrix are computed.

The initialization is one of the problems of the EM algorithm. The selection of λ^0 (partly) determines where the algorithm converges or hits the boundary of the parameter space producing singular, meaningless results. Some solutions use multiple random starts or a clustering algorithm for initialization [FIG02]. We have used random initialization as the starting value of λ^0. The other critical factor is the selection of the order M of the mixture. There are no good theoretical means to determine this order, so it is best experimentally determined for a given task.

4.2 GMM for Speaker Identification

Let us consider S speakers. The features for each speaker are extracted as described in Chapter 3. These featurs are used to construct a GMM for each speaker and the speaker is referred by his/her model λ. This is called the *training* phase of a pattern recognition system. These models $(\lambda_1, \lambda_2, \ldots, \lambda_S)$ are stored and used for classifying an unknown speech sample as described in Section 4.4.

4.3 Why GMM?

There are two principle motivations for using Gaussian mixture densities as a representative of speaker identity.

Speech production is not "deterministic" in that a particular sound (e.g., a phone) is never produced by a speaker with exactly the same vocal tract shape and glottal flow, due to context, coarticulation, and anatomical and fluid dynamical variations. One way to represent this variability is probabilistically through a multi-dimensional Gaussian pdf. The Gaussian pdf is *state-dependent* in that there is assigned a different Gaussian pdf for each acoustic sound class. We can think of these states at a very broad level such as quasi-periodic, noise-like and impulse-like sounds or on a very fine level such as individual phonemes. The spectral shape of the c^{th} acoustic class can in turn be represented by the mean $\vec{\mu}_c$ of the c^{th} component density, and variations of the average spectral shape can be represented by the covariance matrix Σ_c.

The second motivation for using Gaussian mixture densities for speaker identification is that a linear combination of Gaussian basis functions is capable of representing a large class of sample distributions. One of the powerful attributes of the GMM is its ability to form smooth approximates to arbitrary-shaped densities.

4.4 Classification and Speaker Identification

Suppose, for a group of S speakers $\mathcal{S} = \{1, 2, \ldots, S\}$, we have estimated their Gaussian models models $\lambda_1, \lambda_2, \ldots, \lambda_S$. Then for each test utterance, features are extracted as described in Chapter 3.

We have used *maximum a posterior probability classification*, where we compute the probability of each speaker model given the features

$$\mathcal{X} = \{\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_T\},$$

i.e., $\Pr(\lambda_k|\mathcal{X})$ for $k = 1, 2, \ldots, S$. The speaker with the highest probability is chosen. Formally,

$$\hat{S} = \arg \max_{1 \leq k \leq S} \Pr(\lambda_k|\mathcal{X}) \tag{4.11}$$

Using Bayes' formula, we can write

$$\Pr(\lambda_k|\mathcal{X}) = \frac{p(\mathcal{X}|\lambda_k) \Pr(\lambda_k)}{p(\mathcal{X})} \tag{4.12}$$

Assuming equally likely speakers (i.e., $\Pr(\lambda_k) = 1/S$ for $k = 1, 2, \ldots S$) and noting that $p(\mathcal{X})$ is same for all the speaker models, Eq. (4.11) can be simplified to

$$\hat{S} = \arg \max_{1 \leq k \leq S} p(\mathcal{X}|\lambda_k) \tag{4.13}$$

$$= \arg \max_{1 \leq k \leq S} p(\{\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_T\}|\lambda_k) \tag{4.14}$$

which is nothing but maximizing the likelihood as in Eq. 4.5.

If we assume that the feature vectors are independent, then the likelihood for an utterance is simply the product of likelihoods for each feature vector.

$$p(\{\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_T\}|\lambda_k) = \prod_{t=1}^{T} p(\vec{x}_t|\lambda_k) \quad k = 1, 2, \ldots, S \tag{4.15}$$

The assumption of feature vector independence is a very strong one; an implication is that both the speaker models λ_k and the likelihoods calculated above do not depend on the *order* of the feature vectors. Dynamics of feature vectors over time are thus not considered here. By applying the logarithm, we can write the speaker identification solution as

$$\hat{S} = \arg \max_{1 \leq k \leq S} \sum_{t=1}^{T} \log p(\vec{x}_t|\lambda_k) \tag{4.16}$$

Chapter 5

Experimental Evaluation

5.1 Performance Evaluation

The evaluation of the speaker identification experiment was conducted in the following manner. The test speech was first processed using the front-end analysis to produce a sequence of feature vectors $\mathcal{X} = \{\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_t\}$. Then, this sequence was divided into overlapping segments of T feature vectors each as

$$\overbrace{\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_T}^{\text{Segment 1}}, \vec{x}_{T+1}, \ldots, \vec{x}_t$$

$$\vec{x}_1, \overbrace{\vec{x}_2, \ldots, \vec{x}_T, \vec{x}_{T+1}}^{\text{Segment 2}}, \ldots, \vec{x}_t$$

$$\vdots$$

$$\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_{t-T}, \overbrace{\vec{x}_{t-T+1}, \ldots, \vec{x}_t}^{\text{Segment } t-T+1}$$

The identified speaker of each test segment was compared to the actual speaker of the test utterance and the number of correctly classified segments were tabulated. This was done for each of the test speaker. The final performance evaluation was then computed as the percent of correctly identified T-length segments over all test utterances.

$$\text{Accuracy} = \frac{\text{\# of correctly identified segments}}{\text{total \# of segments}} \times 100 \qquad (5.1)$$

5.2 Database Description

The experiment was done using two sets of speaker databases. The first one was the database of speech samples recorded in ISI, Kolkata while the other one was recorded in IISc, Bangalore. The features of the databases are summarized in Table 5.1.

	ISI Database	*IISc Database*
Number of Speakers	7	21
Male-Female ratio	0 : 7	11 : 10
Mother Tongue	Bengali	Hindi, Punjabi, Bengali and Oriya
Recording Language	~~English~~ Bengali	English + Mother Tongue
Sessions per Speaker	6	4
Description of Sessions	3 utterances of one fixed sentence + 3 utterances of another fixed sentence	1 utterance of one fixed sentence + 3 utterances of arbitrary sequences of words
Sampling Frequency	44100 Hz	16000 Hz
Speech Duration	29 sec. approx for every speaker	varied from 40 - 120 sec.

Table 5.1: Summary of the speaker databases

5.3 Results

In both the data sets, the frame length for extracting the features was taken 20 msec with 10 msec overlap. We used 13 coefficients per frame as a feature vector extracted using a filter bank of size 20.

5.3.1 Results for 7 Speaker Database

The reference model was created with one utterance from each speaker. Testing was done using utterances other than those used for training. The duration of each segment of the test speech used was 5 sec. The accuracy obtained is shown in Figure 5.1.

The summary of the results obtained is tabulated in Table 5.2 which shows the accuracy in % and the graphical representation is shown in Figure 5.3.

	8 Gaussian Mixtures		16 Gaussian Mixtures		32 Gaussian Mixtures	
	Text Dep.	Text Indep.	Text Dep.	Text Indep.	Text Dep.	Text Indep.
Min.	99.81	100.00	100.00	100.00	100.00	100.00
Max.	100.00	100.00	100.00	100.00	100.00	100.00
Avg.	99.97	100.00	100.00	100.00	100.00	100.00

Table 5.2: Comparison of results obtained for the 7 speaker database.

5.3.2 Results for 21 Speaker Database

The reference model was created with sentential utterance from each speaker. The training speech was 40-75 sec. long. Testing was done using utterances other than those used for training, i.e., the random sequences of words. The testing speech was 60-120 sec. long. The duration of each segment of the test speech used was 10 sec. The accuracy obtained is shown in Figure 5.2.

The summary of the results obtained is tabulated in Table 5.3 which shows the accuracy in % and the graphical representation is shown in Figure 5.4.

	8 Gaussian Mixtures		16 Gaussian Mixtures		32 Gaussian Mixtures	
	Text 1	Text 2	Text 1	Text 2	Text 1	Text 2
Min.	43.61	49.75	51.08	68.22	72.02	79.93
Max.	100.00	100.00	100.00	100.00	100.00	100.00
Avg.	87.62	91.75	93.18	95.77	96.26	97.31

Table 5.3: Comparison of results obtained for the 21 speaker database.

5.4 Analysis of Results

From the results obtained, we can see (Figure 5.3 and Figure 5.4) that as we increase the number of Gaussians from 8 to 16 and 32 in the mixture, the classification accuracy increases in both the cases which is expected. But the average accuracy falls from 100% to 97.13% on increasing the population size from 7 to 21, which shows that the accuracy decreases with the increasing population size. Nevertheless, in both the cases, the accuracy is more than 90%.

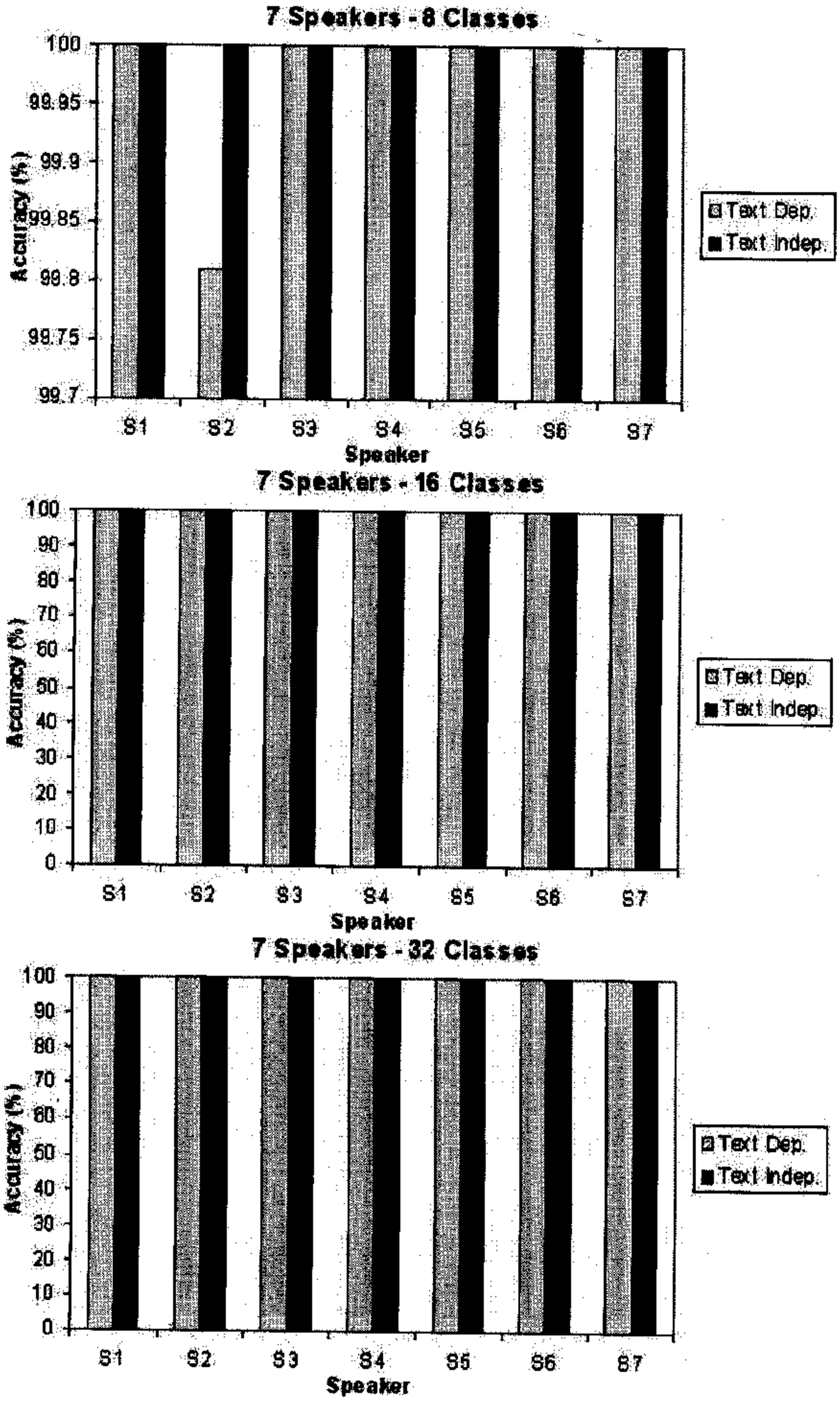

Figure 5.1: Accuracy for individual speakers using 8/16/32 Gaussians in the mixture using 7 speaker data

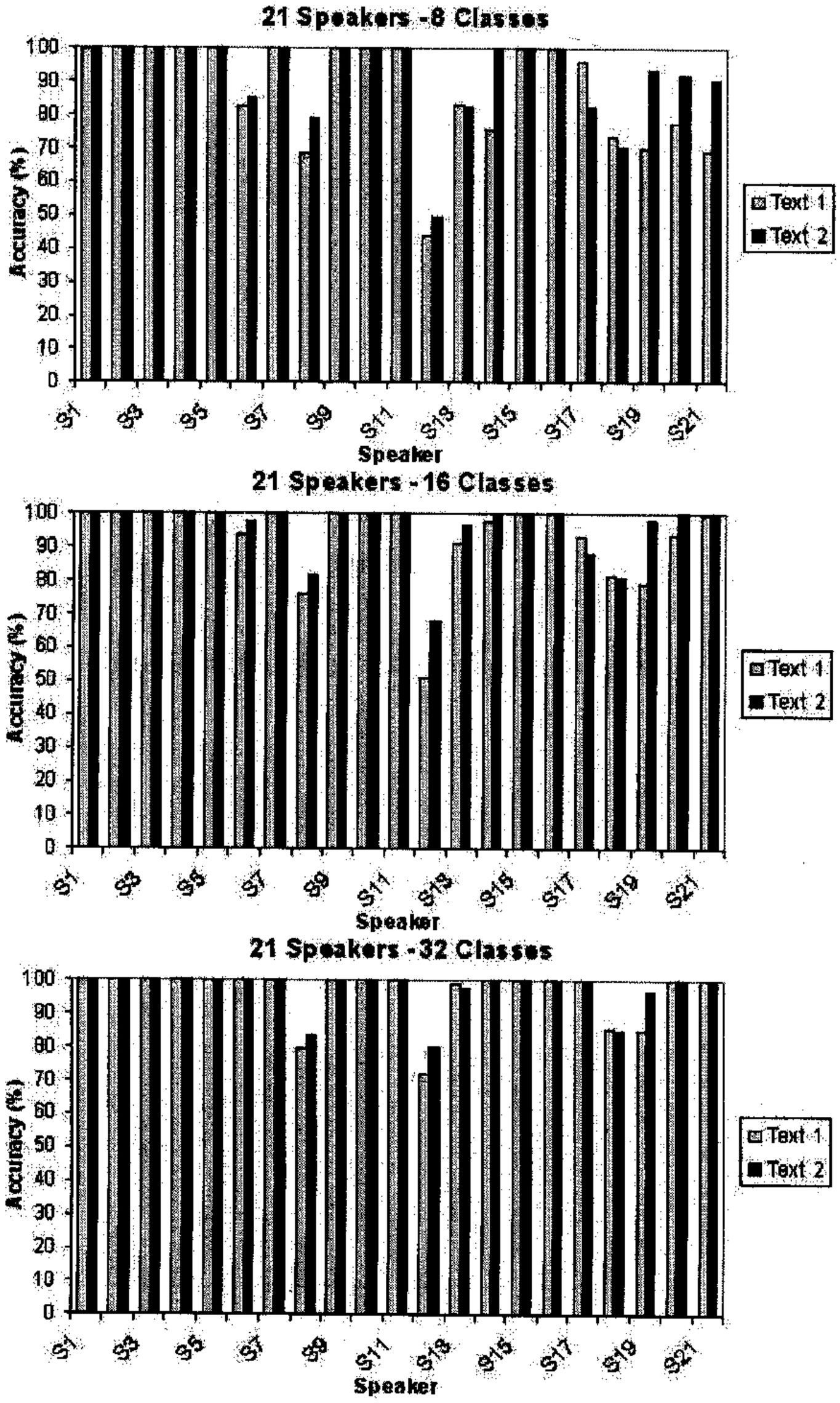

Figure 5.2: Accuracy for individual speakers using 8/16/32 Gaussians in the mixture using 21 speaker data

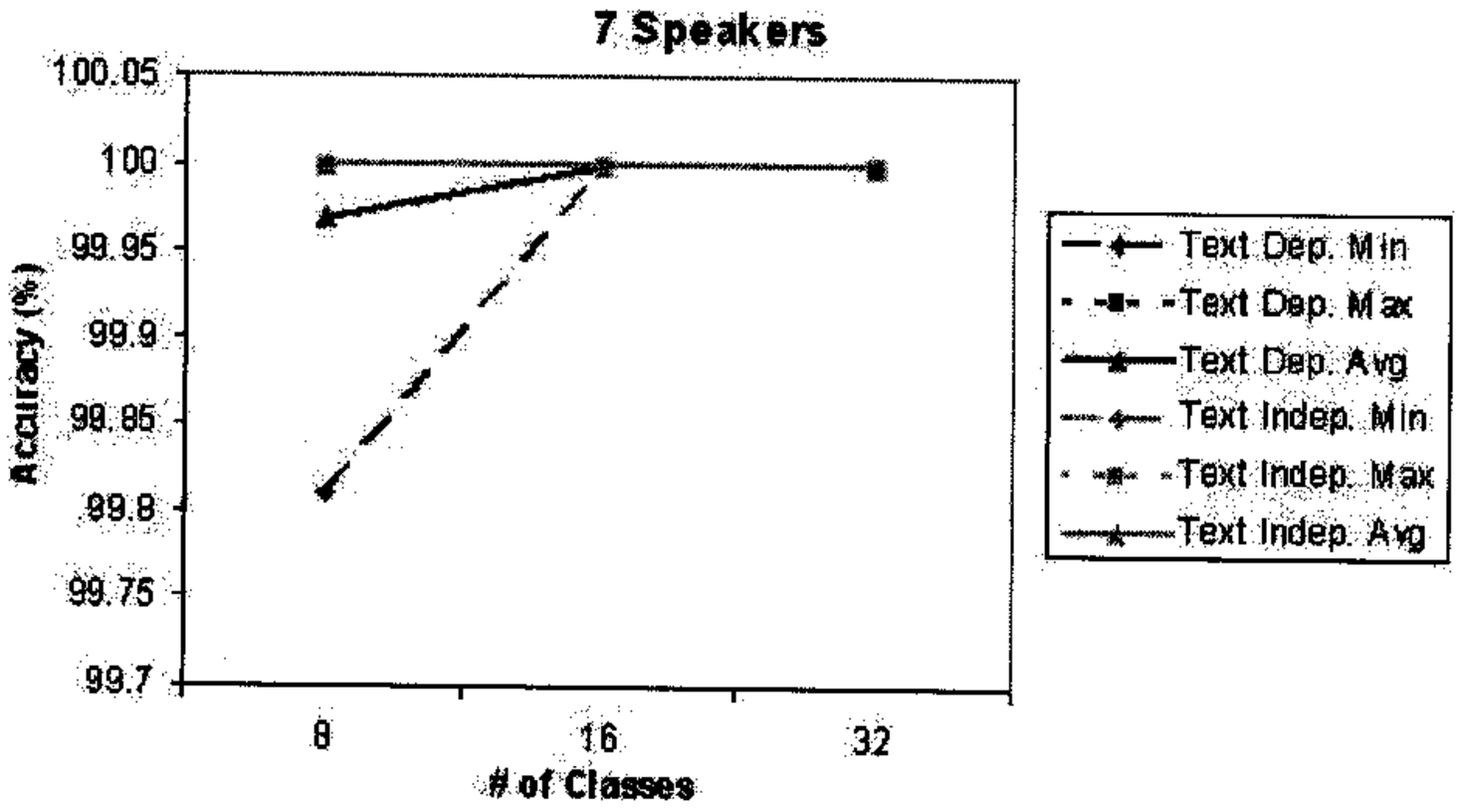

Figure 5.3: Accuracy obtained using 7 speaker data

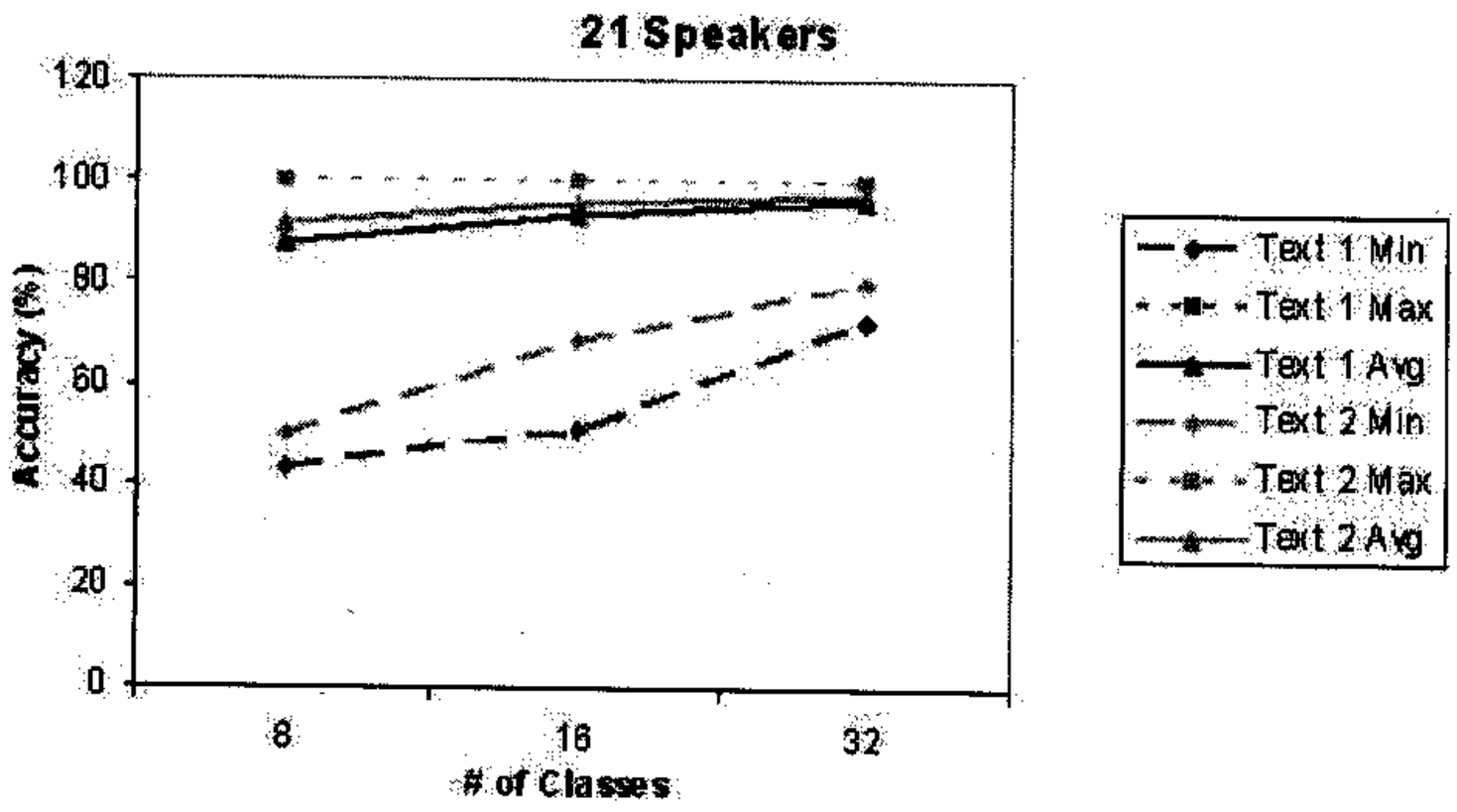

Figure 5.4: Accuracy obtained using 21 speaker data

Chapter 6

Conclusion

We have used MFCC to extract feature vectors as it models the non-linear auditory response of the human ear which resolves frequencies on a log scale. GMMs are motivated for modelling speaker identity based on two interpretations. The component Gaussians represent characteristic spectral shapes from the phonetic sounds and are able to model the short-term variations of a person's voice, allowing high identification performance for short utterances. The GMM is also capable of modelling arbitrary feature distributions. As is observed that on increasing the number of mixtures in a model increases the identification performance, a minimum model order is needed to adequately model speakers and achieve good identification performance.

The performance figures obtained are high because of small data sets and idealized conditions under which recording was done. More testing is necessary on a variety of channels like telephone and cellphone (noisy environment), and a greater number of speakers. It is expected that under such realistic conditions, performance will drop to less than 75% as shown by T.F. Quatieri [QUA04] using the NTIMIT database. But, with high quality speech sounds (TIMIT database), the Gaussian mixture model maintains an accuracy of nearly 100% with large population also. A comparative study is being done with iterative feature selection and classification using statistical networks.

Bibliography

[BIL97] Jeff A. Bilmes, *A Gentle Tutorial on the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov models.* Technical report, University of Berkeley, 1997.

[CAM74] J.P. Campbell, "Speaker Recognition: A Tutorial," *Proc. IEEE,* vol. 85, no. 9, pp. 1437-1462, 1997.

[COM] H.P. Combrinck and E.C. Botha, "On the Mel-Scaled Cepstrum," University of Pretoria, Pretoria.

[DEM77] A. Dempster, N. Laird and D. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," *J. Royal Stat. Soc.*, vol. 39, pp. 1-38, 1977.

[DUD01] R.O. Duda, P.E. Hart and D.G. Stork, *Pattern Classification.* 2nd ed., John Wiley and Sons, New York, 2001.

[FIG02] M.A.T. Figueiredo and A.K. Jain, "Unsupervised learning of finite mixture models," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 24, no. 3, pp. 381-396, Mar 2002.

[FLA72] J. Flanagan, *Speech Analysis Synthesis and Perception,* 2nd ed., Springer-Verlag, New York and Berlin, 1972.

[JAI99] Anil K. Jain, Ruud Bolle and Sharath Pankanti, *Biometrics: Personal Identification in Networked Society.* The Kluwer International Series in Engineering and Computer Science, vol. 479, Springer, Jan 1999.

[LAC88] G. McLachlan, *Mixture Models.* Marcel Dekker, New York, 1988.

[OPP98] Alan V. Oppenheim and Ronald W. Schafer, *Discrete-Time Signal Processing.* Prentice-Hall of India, New Delhi, 1998.

[PHA] Sujay Phadke, Rhishikesh Limaye, Siddharth Verma and Kavitha Subramanian, "On Design and Implementation of an Embedded Automatic Speech Recognition System," Indian Institute of Technology, Bombay, Mumbai.

[QUA04] Thomas F. Quatieri, *Discrete-Time Speech Signal Processing.* Pearson Education, New Delhi, 2004.

[RAB03] Lawerence R. Rabiner and Biing-Hwang Juang, *Fundamentals of Speech Recognition.* Pearson Education, New Delhi, 2003.

[RAB04] Lawerence R. Rabiner and Ronald W. Schafer, *Digital Processing of Speech Signals.* Pearson Education, New Delhi, 2004.

[REY92] Douglas A. Reynolds, *A Gaussian Mixture Modeling Approach to Text-Independent Speaker Identification*, Ph.D. Thesis, Georgia Institute of Technology, Atlanta, 1992.

[REY95] Douglas A. Reynolds and Richard C. Rose, "Robust Text-Independent Speaker Identification using Gaussian Mixture Speaker Models," *IEEE Trans. Speech Audio Processing*, vol. 3, no. 1, Jan. 1995.

[STE40] S.S. Stevens and J. Volkmann, "The Relation of Pitch to Frequency: A Revised Scale," *American Jour. of Psychology*, vol. 53, pp. 329-353, July 1940.

[VOI64] W.D. Voiers, "Perceptual Bases of Speaker Identity," *J. Acoust. Society of America*, vol. 36, pp. 1065-1073, 1964.