A comparison of features for large population speaker identification

Norman Tinyiko Baloyi

Submitted to the faculty of Engineering and the Built Environment, University of Cape Town, in fulfilment of the requirement for the

Master of Science in Electrical Engineering

in the research field of

Communications Engineering

Electrical Engineering, UCT

Cape Town, South Africa

September 2000

To be a Champion, one needs to imitate the Champion, do whatever the champion does - Unknown

To be the Great Champion, one should start where the Champion finished - Tinyiko

To the One who Loved us,

with Lovely people around the Globe

and

Educators educating

We must not cease from exploration. And the end of our exploring will be to arrive where we began and to know the place for the first time.
T. S. Elliot

DECLARATION

I declare that this dissertation is my own work. It is being submitted for the Master of Science in Communications Engineering at the University of Cape Town. It has not been submitted before for any degree or examination at this or any other university.

Signature: Signed by candidate

Baloyi N.T

ACKNOWLEDGMENTS

Whatever is worthwhile is worthy of mention. Anything worthwhile is usually difficult and normally involves a group of people for its success.

First and foremost I am thankful to the Great Counsellor, the Omnipotent God, for His courage, strength and wisdom.

It was a good learning experience to be under the Supervisor of super vision, Dr Daniel Johannes Mashao. I really want to thank him for his close supervision, guidance, patience, constant help, all-time availability and encouragement. The codes he provided were of great importance. With my whole heart I say, **Thank You**.

Mom, Mthavini Mihlava Baloyi and Dad, Risimati John Baloyi, in the midst of difficulties, have shown unfailing love and patience so that I could fulfil my ambition to gain real research skills. Their support and encouragement will long be cherished in my life. I love you mom and dad, live and see many more years.

Without the Foundation for Research and Development (FRD), now NRF, for their financial support, my studies would not have been possible.

Friends, more especially Nelson Manganye, Master's colleagues at the University of Cape Town, Wits University and University of Pretoria and every Beloved, who gave me the courage that activated the Spark inside me to come-out shining strongly, thank you. May your encouragements be a blessing to everyone.

You are all real-shining stars in my life and I love you.

Abstract

Speech recognition systems all have one criterion in common; they perform better in a controlled environment using clean speech. Though performance can be excellent, even exceeding human capabilities for clean speech, systems fail when presented with speech data from more realistic environments such as telephone channels. The differences using a recognizer in clean and noisy environments are extreme, and this causes one of the major obstacles in producing commercial recognition systems to be used in normal environments. It is the lack of performance of speaker recognition systems with telephone channels that this work addresses.

The human auditory system is a speech recognizer with excellent performance, especially in noisy environments. Since humans perform well at ignoring noise more than any machine, auditory-based methods are the promising approaches since they attempt to model the working of the human auditory system. These methods have been shown to outperform more conventional signal processing schemes for speech recognition, speech coding, word-recognition and phone classification tasks. Since speaker identification has received lot of attention in speech processing because of its waiting real-world applications, it is attractive to evaluate the performance using auditory models as features.

Firstly, this study aims at improving the results for speaker identification. The improvements were made through the use of parameterized feature-sets together with the application of cepstral mean removal for channel equalization. The study is further extended to compare an auditory-based model, the Ensemble Interval Histogram, with mel-scale features, which was shown to perform almost error-free in clean speech. The previous studies of EIH to be more robust to noise were conducted on speaker dependent, small population, isolated words and now are extended to *speaker independent, larger population, continuous speech.* This study investigates whether the EIH representation is more resistant to telephone noise than mel-cepstrum as was shown in the previous studies, when now for the first time, it is applied for speaker identification task using the state-of-the-art Gaussian mixture model system.

LIST OF ACRONYMS

ACW	Adaptive Component Weighting
ASR	Automatic Speech Recognition
CF	Characteristic frequency
DFT	Discrete Fourier transform
DNN	Dynamic neural network
DTW	Dynamic time warping
EIH	Ensemble Interval Histogram
FFT	Fast Fourier transform
FIR	Finite impulse response
GMM	Gaussian mixture Model
HMM	Hidden markov model
IHC	Inner hair cells
LPC	Linear Predictive Coding
MEL	Mel-cepstrum
MFB	mel filter bank
NC	Number of Coefficients
PLP	Perceptual Linear Predictive
PFS	Parameterized feature-set
RASTA	RelaAtive SpectrTral
SID	Speaker Identification
SMC	Short-time Modified Coherence
SNR	Signal to Noise Ratio
SUBCOR	Subband-autocorrelation
VQ	Vector Quantization
ZCPA	Zero-crossings with peak amplitudes

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

Introduction

"Countless human babies have made the discovery---that simple sounds, joined together, can make something new. To an adult it is just babbling, yet it marks the start of a mental and physical struggle that is long, arduous and mercifully always forgotten ---a struggle to master the subtle, highly structured and pivotally potent ability called ***speech****, an ability that becomes effortless, and so integral to our being that we accept its enormous power without a thought. Yet its rich complexity makes us unique among the living things, lying at the start of how we become that we grow to be. Speech made in much less than a minute can penetrate our deepest emotions or fling our thoughts to the uttermost star. Truly, speech is the cradle of our humanity. Speech is more information-rich and revealing than is generally appreciated." [126].*

Speech is one of the most important modes of communication between human beings. Speaking is a skill usually learned in infancy and used almost effortlessly from then onwards. As machines become even more predominant in our lives, it is natural to desire man-machine interfaces other than keyboards, mice, and monitors. Automatic Speech Recognition (ASR) is speech technology by machine. Speech technology is a field that is technologically dynamic as well as exciting in its scope, encompassing a multi-disciplinary blend of skills from hard and soft sciences - ranging from mathematics, engineering, and computer science to human perception, linguistic and phonetics [126].

The naturalness associated with speaking and hearing gives little indication of the complexity of the problems of speech processing. Several decades of research in different avenues of speech processing, such as the production, transmission and perception of speech, have yielded remarkable progress, but many fundamental questions still lack definite answers. Part of the problem lies in the unique nature of speech as a continuous acoustic signal carrying a very large amount of information.

For humans and machines to communicate using speech, the capability to automatically extract information from speech waveforms such as the message spoken, the identity of the speaker, or the language spoken is necessary. Technology to extract information from speech has progressed slower than initially expected; it has proven far more difficult than originally envisioned to devise algorithms for automatically performing these tasks that humans take for granted [55].

The work in this thesis focuses on one of the tasks mentioned above: extracting the identity of a speaker from his or her speech. This problem can be broken into two more concrete tasks: *speaker identification* and *speaker verification*. The work of this study concentrates on speaker identification.

1.1 The current state of speaker recognition systems

The ultimate goal of all speaker recognition studies is to devise an automatic time-independent, unbiased system that duplicates the human ability to perform fast, accurate, and text-independent recognition of speakers. Though this ability seems

common and natural to us, the use of machines to do the same task is nontrivial. We have heard of the future of the recognition system, the current state is discussed next.

As recently as five years ago, speaker recognition had reached very high levels of performance, with word-error rates dropping by a factor of five. This performance is largely due to improvements in four main areas. Firstly, the use of common speech corpora allows an easy use of large training sets and an easy way of comparing the results of new recognition systems. Secondly, new ways in acoustic modelling such as Hidden Markov models (HMM), the modelling of cross-word effects, changes in feature vectors over time etc., are a few of the techniques that have helped to reduce word-error rates by up to a factor of two. Thirdly, the use of language modelling with statistical n-gram grammars using probabilities improved recognition of large vocabulary corpora. Finally, the improvements in search algorithms and workstation speed and memory have allowed for a shorter experimentation cycle.

1.2 Problem

Speaker recognition systems reviewed above all have one criterion in common, they are designed to work in a controlled environment using clean speech. If these systems are exposed to speech taken from noisy environments then their performance degrades rapidly.

Speaker recognition (identification and verification) systems do not yet perform well enough to be in widespread use. As will be seen shortly, though performance can be excellent, even exceeding human capabilities for clean speech, systems fail when presented with speech data from more realistic environments such as telephone channels. The differences using a recognizer in clean and noisy environments are extreme, and this causes one of the major obstacles in producing a commercial recognition system to be used in normal environments. The telephone channel introduces many effects that severely distort the speech signal, such as additive noise (such as car engine noise, background babble, white noise etc.), linear filtering, and nonlinearities due to transducers. It is the lack of performance of speaker

identification and verification systems with telephone channels that this work addresses.

1.3 Applications and Benefits of Speech Technology

This section introduces the attractive potential areas for speech applications and the benefits that come out as a result. As an ambitious technology, speech is promising many things that are yet to be done.

1.3.1 Applications

Automatic speech recognition (ASR) is largely aimed at facilitating and expediting human-machine interaction, e.g. replacing keyboards and control knobs with spoken commands interpreted through a voice interface. ASR is an area of immense commercial relevance and is central to the explosive growth of multimedia systems. Commercial applications for ASR include voice dialling on mobile wireless phones, dictation, data entry onto forms, database access (e.g., flight reservation), database management and control, eyes-free and hands-free machine control in factories and laboratories for quality control [97]. Many financial institutions, as well as companies furnishing limited access to computer databases, would like to provide automatic customer service by telephone. Since personal number codes (keyed on a telephone pad) can be lost, stolen, or forgotten, ASR (if sufficiently reliable) can provide a viable alternative. ASR is capable of linking a person to a voice in police work [122].

Speech-recognition systems can usefully employ speaker-recognition technology. One of the first applications of speaker recognition was forensics to investigate the death of Charles I. Current applications include secure control to information, banking, computer networks, PBX and work areas [126]. Gender recognition, based on a variant of speaker-recognition techniques, is already in use in many speaker-independent speech recognizers to improve performance. Practical applications for automatic speaker recognition include criminal investigations, surveillance, and access control. A more suitable application is access control, since it can be expected that the speakers are co-operative; that is they respond to instructions and try to

control variability. The speaker recognition systems can be applied whenever the identities of speakers are known, even if these speakers are not known. In meetings, conferences, or conversations, the technology makes machine identification of participants possible. If used in conjunction with continuous speech recognizers, automatic transcriptions could be produced containing a record of who said what [40]. This capability can serve as the basis for information retrieval technologies from the vast quantities of audio information produced daily. Voice identification has the convenience of easy data collection over the telephone. The other applications of speaker recognition include banking over the telephone, identity verification for banking transactions over the phone, identification of specific speakers in multimedia recordings, voice recognition for smart voice mail systems, voice identification for access to buildings, computer security, as well as access to secure documents over the internet. Speaker recognition systems can be used to help identify suspects. Services of security are plentiful. Some existing applications use voice in conjunction with other security measures, perhaps a codeword, to provide an extra level of security. Systems exist that place telephone calls to check that a speaker is where he or she is supposed to be [40].

Currently, telecommunications provides one of the largest market sectors for speech recognition, with applications as diverse as voice dialling, credit card validation, and access by voice to other information services [126].

Research into the transmission of speech has yielded applications in wireless telecommunications and audio-video teleconferencing. Transmission applications include the wired telephone network where tight specifications are imposed in terms of quality, delay and complexity. Applications of speech coding in storage are voice messaging and voice mail such as those used in telephone answering machines, and voice response systems used as telephone query processors in many large businesses. A growing area in speech transmission is the digital coding of wideband speech for applications like Digital Audio Broadcasting of compact disk audio over Frequency modulation channels, and surround sound for High-Definition Television. Automation of operator-assisted services, credit card sales validation, call distribution by voice commands, inbound and outbound telemarketing, expanded utility of a rotary phone,

repertory dialling, and catalogue ordering are some other telecommunications speech applications [97].

In the medical field, voice creation and editing of specialized reports uses speech recognition. In public service, postal services are faster when operators can individually sort postal parcels by reading, then speaking the labels, and the parcels get routed to the respective bags for delivery. In the military field, speech recognition is applied in warships to control weapon systems. In transport, uses include training air-traffic controllers in the avionics field and booking and timetable inquiries [2].

Speech recognition is useful for the military and the police where they communicate with a central computer radio [9]. In the criminal investigation, speech recognition technology helped police to automate radio inquiries. The usage allows the response time to be reduced and efficiently enhanced for a police officer on the beat [21].

Ainsworth [2] explained that the use of home appliances via voice input is effective for the handicapped too. Speaker independent acoustic helped assists in image coding and animation [11].

The application for text-to-speech synthesis is a telephone-based catalogue ordering service, where the system can respond with a full description of the item as well as saying the name and address of where the item is to be sent. Access to news is another example. The ability to speak e-mail messages over the telephone is another important application, enabling people to check their messages when away from a computer terminal [126].

1.3.2 Benefits

Since the use of speech technology can be an effective alternate to manual methods like the keyboard, mouse and touch screen, this technology allows users to utilize a large range of commands compared to keyboards. Any of these commands can be given at any point in time and this reduces mental encoding. Where users are unable

to use their hands and eyes at the same time while engaged in other areas of work, speech technology is very useful. Since speech is a faster and more natural mode of communication, verbal and manual tasks can accomplish multiple jobs to save a lot of time and increases productivity.

Speech technologies can reduce operating costs by the automation of operator services and call centres; allow systems control by voice to make them easier to use and hence improving customer satisfaction; extend system usage to areas where other digit entry systems fail, where users require easily remembered access codes; provide a simple means of getting large amounts of variable text data to users over the telephone; creating new speech services [126].

Speech recognition can serve as an economic purpose by reducing the amount of labour required. The above list is by no means complete, but provides an indication of the types and variety of applications. Speech recognition is nowadays regarded by market projections as one of the more promising technologies of the future. Speech communication is, and will continue to be, key to the use of the network, currently accounting for well over 90% of revenue. Having the right applications and service is the key to revenue generations [126].

1.4 The Challenges of speech

Human beings are able to understand speech under various conditions such as noisy environment, different speaker accents, speaking rates or speaking levels and various modes of speech (i.e. continuous speech or isolated utterances). The possibilities of automatic speech recognition could go far beyond the ones mentioned in section 1.3 if speech recognition systems could achieve recognition rates comparable with those of human beings. Human beings are referred to be more robust to noise and therefore robust speech recognition systems are necessary in order for speech recognition systems to be applied anywhere, any time.

Although signal processing strategies show promise in leading to a robust system (Section 1.6), the fundamental method for improving robustness is to understand

better the many sources of variability in the speech signal. Figure 1-1 shows some of the many sources of variability in the speech signal from the viewpoint of a machine recognizer. Variability is typically due to the talker and the nature of the task, the physical environment, and the communication channel between the user and the machine. Changes in the speaker's voice are caused by modifications of articulation to increase intelligibility where auditory feedback is affected by excess levels of noise. This is known as the *Lombard Effect*, causing highly variable intra and inter-speaker distortion highly damaging to recognition performances. The main variances can be found in increases in the speaker's pitch, amplitude, vowel duration, and spectral tilt, as well as formant frequencies F1 and F2. Even more variability arises if the speech originates from more than one speaker, because of differences in vocal tract, accent and speaking style. Some of these sources of variability are clearly irrelevant to the task and should be treated as noise. Other sources of variability, such as speaking rate and fundamental frequency, should be treated as knowledge rather than noise sources.

Perhaps the most important technical challenge for applications is dealing with the effects of the communications channel through which speech is received. In many applications this is a telephone channel. The difficulties arise since the channel may vary from utterance to utterance. Some of the most significant factors are band limiting effects; spectral distortion due to a non-flat frequency response in the passband; and additive or nonadditive channel or environmental noise.

The changing channel effectively creates variability in a speaker's acoustics that far exceeds his/her normal variability. The variability essentially distorts their patterns in the feature space, increasing confusion. Variability also manifests itself as the occurrence of artifacts such as crosstalk and noise events.

The most significant factor affecting automatic speaker recognition performance is variations in signal characteristics from trial to trial. There is a high intra-speaker variability over time due to health (respiratory illness, laryngitis, etc.), stress, emotional factors, speech effort and speaking rate, aging, gender, etc. It is also possible for speakers to change their pronunciation, affecting spectral characteristics

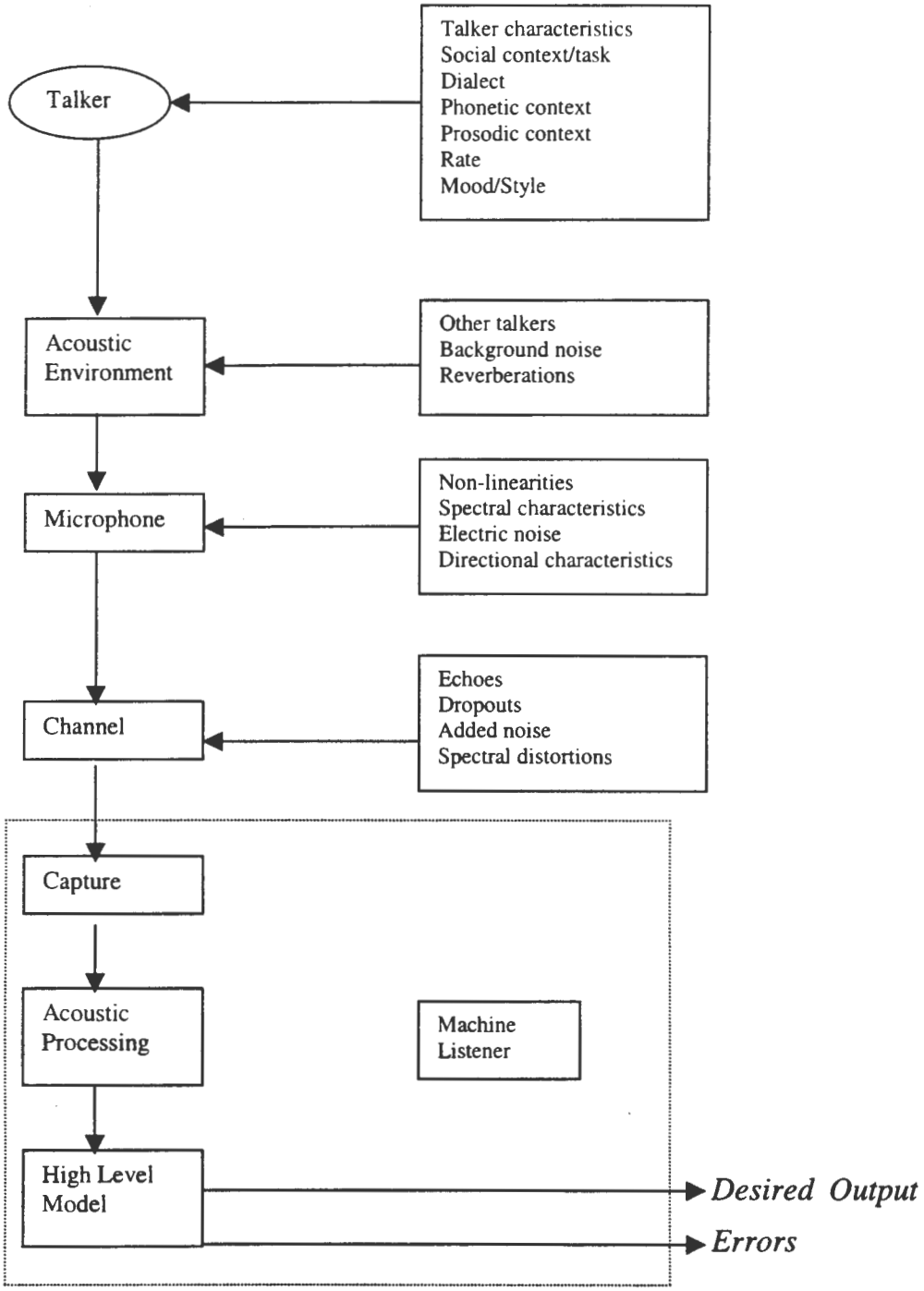

Figure 1-1: Sources of variability in the speech signal [13].

of individual sounds. Involuntary variations and voluntary variations are of equal concern. Speakers cannot repeat an utterance precisely the same way from trial to trial. Speakers also produce non-speech sounds, such as breath noises and lip smacks, which must be detected and tolerated. More difficult to deal with are spectral and pitch changes arising from physical or emotional changes or from certain types of speech pathologies. Speech pathologies, such as stammering or stuttering, which

produce large variations in utterances from trial to trial, are especially difficult to deal with.

1.5 Continuous Speech Recognition Systems

Work on the recognition problem was started four decades ago by clean (relatively free of noise and distortion) isolated word-speech from a single speaker and bounding the problem with constraints such as a small vocabulary and simple grammar. After a series of breakthroughs in processing algorithms, automatic speech processing (ASR) technology today has advanced to speaker-independent, large-vocabulary, continuous-speech recognition being developed on several systems around the world. These ASR systems include state-of-the-art speaker-independent speech recognition system SPHINX system at CMU [67, 68], the HTK at Cambridge University [127], the speech recognition system at LMSI, France [64], the SUMMIT system at MIT [129], the TANGORA system at IBM [56], the Tied-Mixture system at Lincoln Labs [PAU90], and speech recognition system at AT&T Bell Labs [69].

Most of these systems adopt the pattern matching approach to speech recognition, which involves the transformation of speech into appropriate representations with signal processing techniques, creation of speech models via statistical methods, and the testing of unknown speech segments using pattern recognition techniques. A pattern recognition system basically consists of a front-end feature extractor and a classifier, as shown in Figure 1-2. The first module, feature extractor, serves to extract information of interest from the large quantity of input speech data, in the form of a sequence of feature vectors. This parametric representation generally compresses speech while retaining relevant acoustic information such as location and energy of formants.

The classifier takes the features computed by the feature extractor and performs either template matching or probabilistic-likelihood computation on the features, depending on the type of algorithm employed. Before it can be used for classification, the classifier has to be trained so that a mapping from the feature label of a particular class is established. Since an object is characterized in the classifier by a module or a

part of an integrated model, training is also the stage of part of enrolment. Such an approach has been demonstrated to be efficient in performing a pattern recognition task. However, the implicit assumption of this approach is that the training and testing conditions are comparable. Problems arise when there is a mismatch between the environments for training and testing, which is generally true in most applications.

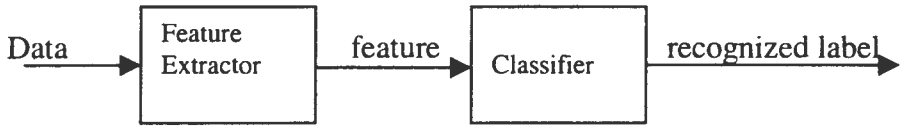

Figure 1-2: A structure of a pattern recognition system.

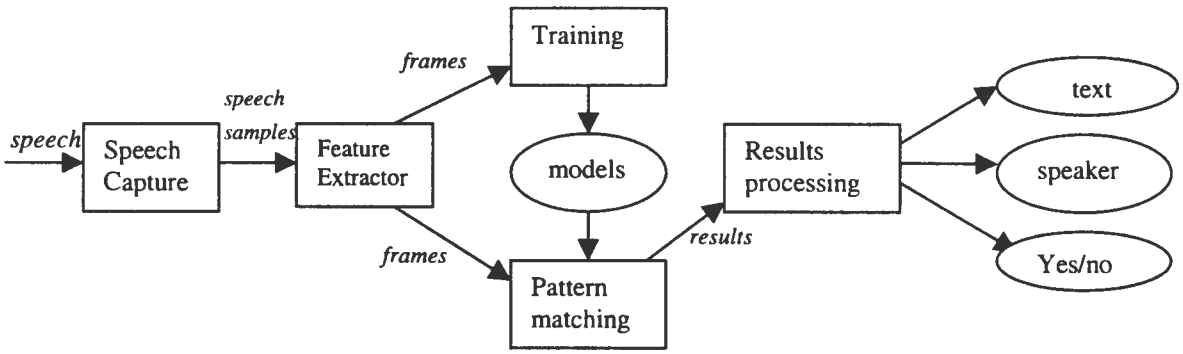

Figure 1-3: Pattern recognition system for speech and speaker recognizers [126].

1.6 Robust Speech Techniques

Many real-life applications for speaker and speech recognition technology are impeded by the large degradation in system performance due to environmental differences between training and testing. This is known as the *mismatched condition.* It is the effect of types of noise that produce serious mismatches (Section 1.4). Studies have shown that most contemporary systems achieve good recognition performance if the conditions during training are similar to those during operation (matched condition) [110]. Robustness improving techniques explicitly based on matched training and testing conditions, noise level, SNR or the particular signal distortion, are clearly not desirable for robustness to multiple adverse acoustic conditions. Users will

naturally be reluctant to rely on automatic speech recognition if they have to talk in a highly constrained way, if it fails on a day when they have a cold, or if performance drops severely when there is a reasonable level of background noise. It is impractical to train for diverse signal conditions, therefore it is advisable to make the ASR more robust.

Robustness in speech recognition can be defined as minimal, graceful degradation in performance due to changes in input conditions caused by different microphones, room acoustics, background or channel noise, different speakers, or other small (insofar as human listeners are concerned) systematic changes in the acoustic signal [13]. At present, speech recognition systems are not very robust. Systems trained in the laboratory fail when exposed to operating conditions in the field [12985]. Their performance degrades suddenly and significantly with modifications as a minor change in microphone or telecommunication channel. For example, LPC features work well under similar acoustic training and testing conditions, but the system performance deteriorates significantly under adverse signal conditions [15]. In [1], for an alphanumeric recognition task, the performance of the SPHINX system falls from 77-85% to 19-37% accuracy on cross conditions.

Several methods have been proposed to improve robustness of speech recognition systems [1, 12, 22, 39, HER91, 92]. These include training the system in a noisy environment, signal enhancement pre-processing [92, EPH87, 1], special transducer arrangements [VIS86], robust distance measures [SOO87], and alternative speech representations [75, 114, 34], using noise masking techniques, modifying the parameter selection and computation process, improving segmentation techniques, using different types of microphones, and using multiple microphones with beam-forming algorithms.

In the case of distortion by additive noise, well-established speech enhancement techniques can be used to suppress the noise [LIM83]. A variety of strategies have been suggested to overcome the channel effects and noise on speech recognition systems. These include feature classification, feature selection, signal processing, and signal acquisition. One method is to enhance the speech by spectral subtraction [BOL79] so that the features are more representative of clean speech if those noise

effects are suppressed. Also, if the features are the cepstral vectors, mean removal [27] attempts to remove the transmission channel effects. Efforts are also made to find new features that are robust to noise and channel effects. Some examples of these are short-time modified coherence [75], ear-based features [34], fine structure features [55], etc.

The classifier can be made more robust by compensating for the distortions at the classification stage. Statistical approaches are usually adopted to obtain the probabilistic modelling of features so that a robust mapping from the testing data to the training data can be created. Methods such as optimum filtering [84], Gaussian mixture model (GMM) [104], and hidden Markov model (HMM) adaptation [86] all fall into this category. Also, robust distance metrics such as the Itakura spectral distortion measure [52, 118] and the projection measure [75] will lead to more accurate labelling of the test data.

Gish [39] has demonstrated that by simply prefiltering different recordings of telephone speech with a fixed filter, text-independent speaker recognition performance over telephone lines can be improved significantly. Signals thus filtered are less mismatched in their channel characteristics. For text-dependent speaker recognition, blind equalization method [27] and channel-invariant acoustic features [117] can be used to create a spectral difference between test and training material. A more effective way to construct a channel-resistant speaker recognition system is to model the channel statistically. Gish [37, 38] used multivariate Gaussian probability density functions to model channels.

However, more recently techniques have been developed which take advantage of speech production and perception knowledge. Since the human system is the most efficient recognition system known, then there is hope that by using auditory, physiological, and production knowledge in the design of recognition systems that some of the successes of human recognition can be duplicated. A brief explanation of both signal processing and knowledge based approaches will be covered in chapter 3.

Approaches to robust recognition can be summarized under the following research areas: better training methods, improved front-end processing, and improved back-

end processing or robust recognition measures. The recognition approaches have in turn been used to address improved recognition of speech in a noisy environments, Lombard effect, work load task stress or speaker stress, microphone or channel mismatch [45].

1.7 Recognition Performance in Noise

Techniques mentioned in section 1.5 generally use some estimate of the noise such as noise power of SNR to obtain better spectral models of speech from noise-corrupted signals. The enhancement techniques have been directly applied to speech recognition [EPH87], [92]. In [92], the optimal least squares estimator of short-time spectral components is computed directly from the speech data. The use of the optimal estimator increased the accuracy for a speaker- dependent connected digit recognition task using a 10 dB SNR database from 58% to 90%. In [EPH87], the short-time noise level as well as the short-time spectral model of the clean speech are iteratively estimated to minimize the Itakura-Saito distortion [51], between the noisy spectrum and a composite model spectrum. For a speaker-dependent isolated word recognition task using 10 dB SNR test speech and clean training speech, the performance without speech pre-processing was found to be 42% and improved to 70% after pre-processing the noisy speech.

Methods for making the recognition system microphone-independent based on additive correction in the spectral domain were presented in [1]. Cross-microphone evaluation was performed on an alphanumeric database in which utterances were recorded simultaneously using two different microphones with average SNR's of 25 dB and 12 dB. In the first, SNR-dependent cepstral normalization, a correction vector depending on the instantaneous SNR is added to the cepstral vector. This method improved recognition from 19%-37% baseline to 67%-76%. The second method, codeword-dependent cepstral normalization, computes a maximum likelihood estimate for both the noise and cepstral tilt and then a minimum mean squared error estimate for the speech cepstrum. The performance of this method ranged from 75% - 74%.

In [VIS86], several single-sensor and two-sensor configuration of speech transducers were evaluated for isolated word recognition of speech corrupted by 95 dB and 115 dB sound pressure level broad-band acoustic noise similar to that present in a fighter aircraft cockpit. Performance improvements were reported with various multisensor arrangements as compared to each single sensor alone, but the task was limited since the testing and training conditions were matched.

Robust distance measures aim to emphasize those regions of the spectrum that are less corrupted by noise. In [SOO87], a weighted Itakura spectral distortion measure which weights the spectral distortion more at the peaks than at the valleys spectrum is proposed. The dynamic time warping (DTW) based speech recognizer on an isolated digit database for a speaker-independent speech recognition task, using additive white Gaussian noise to simulate different SNR conditions was used to test this measure. At high SNR, the performance was the same as the original unweighted Itakura distortion measure. At medium to low (5 dB) SNR's, this measure improved from the accuracy of 72% to 88%.

A technique for robust spectral representation of all-pole sequences, called the Short-Time Modified Coherence (SMC) representation, is proposed in [75]. A square root operator in the frequency domain compensates for the inherent spectral distortion introduced by the autocorrelation operation on the signal robustness to additive white noise. For 10 dB SNR spoken digit database, the SMC maintained an accuracy of 98%.

Rasta (RelAtive SpecTral) [50], methodology suppresses constant or slowly-varying components in each frequency channel of the speech spectrum by high-pass filtering each channel with a filter that has a sharp spectral zero at zero frequency [50]. For isolated digits, with training on clean speech and testing on speech corrupted by simulated convolutional noise, Rasta-PLP (Perceptual Linear Predictive) technique yielded 95% accuracy while LPC yielded 39% accuracy.

1.8 Related Theory

The human auditory system exhibits better robustness than any machine processor under different adverse acoustic conditions. For this reason, models based on the human auditory system have been proposed. SENEFF [114] is one of auditory models and has yielded good results for isolated word recognition [JAN92, 70]. The Ensemble Interval Histogram (EIH) [34] is another auditory models and it was evaluated with a DTW based recognizer on a 39-word alpha-digit speaker-dependent task in the presence of additive white noise. In high SNR the EIH performed similarly to the DFT front-end, whereas at low SNR it outperformed the DFT front-end markedly.

Jankowski [54] compared mel cepstrum and three auditory models – Seneff's mean rate response and synchrony response [113], and the EIH. An isolated word task in a T1-105 database recorded from 5 male and 3 female speakers was used to train and test the systems, with the use of continuous density mixture Gaussian Hidden Markov Models (HMMs) as classifier. The experiment was trained on clean speech and tested on speech distorted by additive noise or spectral variability. Auditory models slightly outperformed the mel-cepstrum.

Sandhu [111] compared mel-cepstra with EIH for phone classification under adverse conditions with Mixture Gaussian HMMs performed as a classifier on a TIMIT database [3] on 3 male speakers. MEL outperformed EIH in clean speech, whereas EIH outperformed MEL for the speech passed through the telephone channel.

Kim [62] [60] conducted EIH in Korean word recognition experiments made by 20 male speakers, and also reported another word recognition experiments on the speech data of 50 Korean words [61]. Ghitza performed EIH for speech processing in a database of 96 pairs of confusable words spoken in isolation by several male and female speakers [35], speech coding and speech processing experiments on the sentences spoken by four male and female speakers [36], and conducted EIH on the database of two males and two females for speech recognition [33]. In these studies, an auditory model, EIH, was shown to outperform MEL.

1.9 Contribution of this Study

This study differs from the previous evaluations in the following ways:

- The size of the database is much larger than those used earlier. The database contains 630 male and female speakers.
- The use of noisy continuous speech database instead of isolated or connected word databases.
- Speaker identification is performed on EIH, instead of word recognition, phone classification, or speech coding.
- *Gaussian mixture models (GMM)* are used instead of Gaussian HMMs or dynamic time warping (DTW) base recognizer. *EIH is applied for the first time on GMM.*
- *And to our knowledge, EIH is evaluated for the first time for the text-independent closed-set speaker ID.*

The speech front-ends representations are passed to the implemented start-of-the-art speaker recognizer, Gaussian mixture model. Use of GMM classifier is justified by its being an established, general classifier, which acts as a hybrid between standard parametric classifier, which assumed predetermined distributions and non-parametric classifiers, which typically are computationally expensive. The telephone-channel NTIMIT database is used because real-world applications will operate in noisier conditions and because it is standard, phonetically rich and hand segmented. This database provides us with an excellent source for the study of continuous training and evaluation. Text-independent speaker identification is chosen since there is a great interest of real-world applications over communications channels.

Although good results have been achieved using Ensemble Interval Histogram (EIH) for phonemic patterns, word recognition, speech coding and speech processing, the question remains whether this technology can be used effectively for speaker identification. This thesis attempts to investigate the application of EIH for the text-independent closed-set speaker identification.

1.10 Outline of the thesis

The organization of the thesis is as follows:

- Chapter 2 sketches the process of sound production, and contains a short description of different sound classes. This chapter is motivated by the fact that the success of auditory models lies in our knowledge in both psychological and physiological studies of the auditory system. As auditory function and speech perception are better understood, new parameters important to the auditory models will be uncovered, and the recognition rate will improve. For these reasons, this chapter serves as a background to what is currently known about the auditory system and speech perception.

- Chapter 3 describes different signal representations including mel-cepstrum, parameterized feature-set, and Ensemble Interval Histogram, compared in this study. A brief description of part of the human auditory mechanism is given as background for EIH. The details of implementation for these representations are provided.

- Chapter 4 gives speaker identification background, speaker identification problem (SID), methods to solve the problem and their advantages and limitations. This chapter also describes a state-of-the-art SID system that will be used later in the work. SID systems can be split into two parts, a front-end and a pattern classifier; this chapter focuses on the second one part.

- Having developed the techniques, and then applied them, Chapter 5 revisits the speaker identification problem, evaluating the features described in chapter 3. This chapter first describes the test conditions, revisits the SID system, outlines the database used, and discusses the performance/results.

- Chapter 6 summarises the work done in this thesis and the conclusions drawn from the results. This chapter attempts to make a convincing argument that there

is indeed room for performance improvement in SID systems. Possible directions for future research are provided.

1.11 Summary

This chapter introduced the problem of speaker recognition and some of the issues in current speech research. The problem of recognizing noisy speech and different approaches taken to attain robustness in speaker recognition were described.

Since more recent techniques take advantage of speech production and perception knowledge, the next chapter contains speech production and speech sounds that serve as a background.

CHAPTER 2

Speech generation

This chapter is mainly referenced from [97, 87, 126]. The description of the speech organs and the mechanics of producing and perceiving speech in human beings are discussed. The study of the manner of sound production in the vocal tract and the physical properties of the sounds thus produced is called *phonetics*. It is by the understanding of these processes that leads to several approaches to speech recognition by machine. The ideas of acoustic-phonetic characterization of sounds lead naturally to straightforward implementation of a speech recognition algorithm based on sequential detection of sounds and sound classes. This chapter is outlined as follows, section 2.1 contains organs of speech, section 2.2 contains physiology of speech production, and the description of the different sounds in Section 2.3.

2.1 Organs of Speech

Vocal organs are divided into three main subsystems: lungs and trachea, larynx, and vocal tract. The properties of these subsystems are briefly described next.

A. Lungs and trachea

The lungs compress air and deliver it to the system though the trachea. These two organs are the power supply of the system. They also control the loudness of the resulting speech, but they make an audible contribution to speech.

B. Larynx

Larynx contains the principal sound generating mechanism. Its parts are cricoid cartilage, thyroid cartilage, arytenoid cartilage, and vocal chords. The cartilages are mostly frameworks. The acoustical function of vocal chords is to provide the principal excitation source for speech.

C. Vocal tract

The vocal function of the vocal tract is the colouring and articulation of the voice. The vocal tract also contains the principal points from which the speech sounds are radiated.

2.2 Speech Production

The production process begins when a talker formulates a message (in his mind/brain) that he/she wants to transmit to the listener through speech. This message is conveyed via spoken sounds and then uttered aloud with the execution of a series of neuromuscular commands. Once the speech signal is generated and propagated to the listener, the speech perception process begins.

Figure 2-1 is a sketch of the mid-sagittal cross-section of the human apparatus [97]. Air enters the lungs through the normal breathing mechanism. Air is then expelled from the lungs through the trachea, causing the vocal chords to vibrate, and finally shaped by motion of the articulators (i.e. the lips, jaw, tongue, and velum) to produce different sounds.

The vocal tract begins at the opening of the vocal chords, and ends at the lips. Vocal tract consists of two parts, the *pharynx* and the *mouth*. The *nasal tract* is the cavity between the velum and the nostrils. The nasal tract is acoustically coupled to the vocal tract to produce nasal sounds of speech.

Depending on the pressure gradient across the glottis, the mass and tension of vocal folds, two kinds of air flow are generated upwards through the pharynx, quasi-periodic and noise-like. Quasi-periodic pulses of air are produced when the vocal folds vibrate in a relaxation oscillation at a fundamental frequency. These excite the resonant cavities in the vocal tract resulting in voiced speech sounds. Broad-spectrum noise-like airflow is generated by forcing air at a high enough velocity through a constriction in the vocal tract, so as to cause turbulence. This produces unvoiced sounds by exciting the vocal tract with a noise-like waveform, and plosive sounds as when air pressure is built up behind a complete closure in the oral cavity and then abruptly released.

The nasal tract remains disconnected from the vocal tract for most sounds and gets acoustically coupled to it when the velum is lowered to produce nasal sounds. The resonant frequencies of the vocal tract are called *formants*. Different sounds are formed by varying the shape of the tract to change its frequency selectivity. The rate of change of vocal tract configuration categorizes sounds as *continuant* and *noncontinuant*. The former are produced when a non time-varying vocal tract configuration is appropriately excited, e.g., vowels and nasals, and the latter are produced by a time-varying vocal tract. Vowels, which are continuant sounds, can be differentiated into close, open, high, low and rounded based on the positions of the articulators. The consonants can also be alternatively classified by their place-of-articulators, e.g., labial, alveolar, velar, dental, palatal or glottal, along with the

manner-of-articulation (plosive, fricative, nasals etc.). There are different ways of characterizing sounds based on the physical mechanism of production.

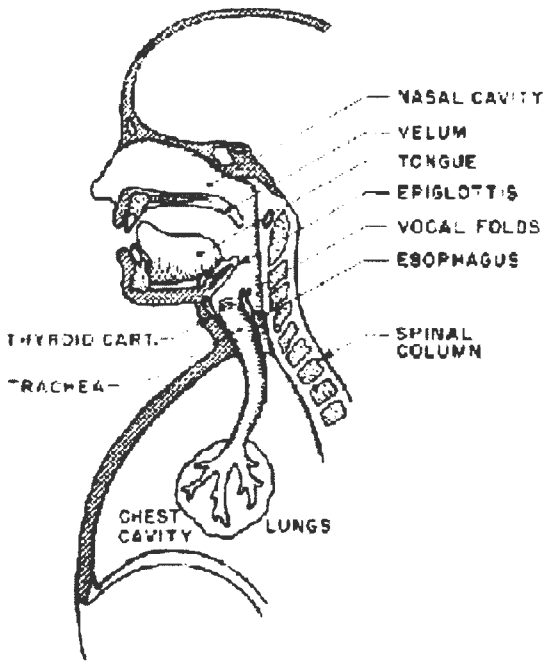

Figure 2-1: Schematic view of the human vocal mechanism.

Speech is produced as a sequence of sounds. Sound is produced during a speech either as a result of the vibration of the vocal folds, or as a result of acoustic noise resulting from turbulent airflow, or a combination of these efforts. Hence the state of the vocal cords, as well as the positions, shapes, and sizes of the various articulators, changes over time to reflect the sound being produced.

The general sounds of English and the relevant acoustic and phonetic features of the sounds are discussed next.

2.3 Speech sounds and features

Linguistically, the sound can be broken down into a large number of phonetic units called *phonemes*. Phonemes are the sound segments that carry meaning distinction, identified by minimal pairs of words that differ only in part.

The phonemes are categorized into different groups (Figure 2-2). The classification is generally done by defining some common behaviour of the vocal tract for each particular group of sound.

Vowels

Vowels are produced by exciting an essentially fixed vocal shape with quasi-periodic air pulses caused by the vibration of the vocal cords, and are the most stable set of sounds. They are generally longer in duration as compared to consonants, and are spectrally well defined. All vowels are voiced sounds. The articulatory configuration of the vocal tract gives an indication of the type of vowel that has been uttered. Furthermore, vowels display a striking periodicity in their temporal representation. This periodicity reflects formant frequencies and is useful in classifying the vowels.

Diphthongs

Diphthongs are vowels during which the sound quality changes, due to tongue and/or jaw movement. A diphthong is produced by varying a vocal tract that starts at or near the articulatory position for one vowel and moves to or toward the position for another.

Semivowels

Liquids and glides are called *semivowels* since they are vowel-like in nature. They are generally characterized by a gliding transition in vocal tract area function between adjacent phonemes. They are voiced sounds, but their overall acoustic characteristics depend strongly on the context.

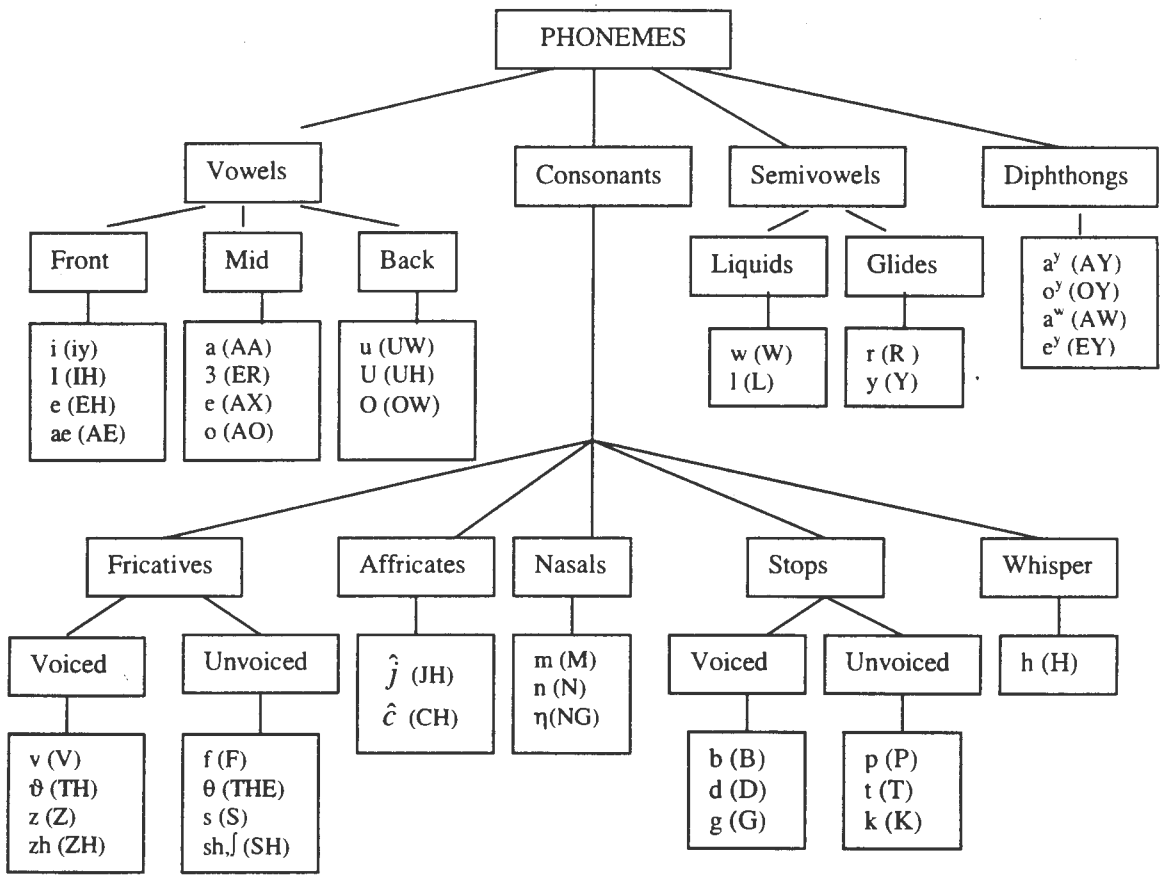

Figure 2-2: Chart of the classification of the standard phonemes of American English

Nasals

The nasal consonants are produced by lowering the velum, so that the air flows through the nasal tract, in order to acoustically couple the nasal tract to the vocal tract, with glottal excitation and a complete constriction of the vocal tract at some point in the oral activity. Sound is radiated through the nostrils and the mouth acts as a resonant cavity. The resonances of the spoken nasal consonants and nasalized vowels are spectrally broader than those of vowels. All three nasals have a prominent low frequency first formant, called the *nasal formant*.

Fricatives

Fricatives are produced by forming a constriction in the vocal tract and forcing air through the constriction so that turbulence is created. The back cavity below the constriction traps the oral cavity in the case of nasals and introduces anti-resonances in the sound radiated from the lips. There are two kinds of fricatives, voice and unvoiced. The voiced fricatives differ from their unvoiced counterparts in that two exitation sources are involved in their production.

Stops

Stops are noncontinuant, plosive sounds. They are generally of short duration, and are more difficult to identify from the spectral information alone. Their properties are greatly influenced by the context. There are two kinds of stops, voiced and unvoiced.

Affricates

An affricate is an initial closure of vocal tract followed by gradual release producing turbulence. Affricates also display a closure-burst in the spectrogram, followed by the fricative region.

Whisper sounds

The phoneme /h/ is produced by exciting the vocal tract with a steady airflow without vibration of the vocal cords. Turbulence is produced at the glottis. This is also the mode excitation of whispered speech. The spectral characteristics of /h/ depend on the vowel following it.

2.4 Summary

This chapter briefly discussed the organs of speech, the process of speech production, and described some characteristics of sounds classified by the manner of production. The next chapter introduces us to the way these sounds are used to produce features.

CHAPTER 3

Feature Extraction

Feature extraction is the process by which speech data is reduced to much smaller amount of data which represent the important characteristics of the speech. This reduced speech data can later be used to generate a representation of each speaker.

Broadly speaking, speaker identity is correlated with physiological and behaviour characteristics of the speaker. Behavioural correlates of speaker identity in the speech signal are more difficult to specify. Generally speaking, the characteristics of speech that have to do with individual speech sounds are referred to as *segmental*, while those that pertain to speech phenomena involving consecutive phones are referred to as *suprasegmental*.

Although the exact factors in a speech signal that are responsible for speaker characteristics are not all known, it is a fact that humans are able to distinguish among speakers based on their voices. Studies on interspeaker variations and factors affecting voice quality have revealed that there are various parameters at both the segmental and suprasegmental levels that contribute to speaker variability [18]. Explicit measurements of speaker characteristics are often difficult to carry out. For example, such characteristics as nasality and breathiness are difficult to measure because of the difficulty in labelling, segmenting, and measuring nasalized speech events. Voice source measurements have sometimes been mentioned as good candidates for characterizing speakers. However, other than pitch, many voice source characteristics are not easy to extract from the speech signal. Even though many voice characteristics are difficult to measure explicitly, we know that many characteristics are captured implicitly in the kinds of signal measurements that can be performed relatively easy. Such signal measurements as short-term and long-term spectral energy, overall energy and fundamental frequency are relatively easy to obtain. Moreover, these measurements can resolve differences in speaker characteristics that often surpass the ability of human discrimination. Thus, voice characteristics, whether physiological or behavioural in origin, which can be measured automatically and easily from the speech signal are the bases for most automatic recognition systems.

Despite the fact that one cannot exactly quantify interspeaker variability in terms of features, current speaker identification systems perform very well with clean speech. However the performance of these systems can decrease significantly under certain acoustic conditions, such as noisy telephones lines [101]. This poor performance is mainly due to how badly the features are affected by noise.

All speech recognizers include an initial signal processing front-end that converts a speech waveform into features useful for further processing. The front-end is required to extract important features from the speech waveform that are relatively insensitive to talker and channel variability unrelated to speech message content. This first stage also reduces the data rate into later stages of the speech recognizer and attempts to decrease redundancy inherent in the speech waveform.

It has been shown experimentally that speaker recognition performance strongly depends on the front-ends unit that pre-processes the speech signal [83]. Speech information is primarily conveyed by the short-time spectrum, the spectral information contained in an interval of 10-30 ms. Types of features to characterize a short-time spectrum are discussed in section 3.1, followed by the representations of the signals, mel-scale in section 3.2, parameterized feature-set in section 3.3, and EIH in section 3.4. Data-reduction technique of the front-ends in section 3.5, a short comparison of the two front-ends, MEL and EIH, in section 3.6, the summary of the chapter is given at the end in section 3.7.

3.1 Types of Features

For a text-independent system to perform well, the proper choice of features to be used is necessary. Ideally features should [93]:

- Occur naturally and frequently in normal speech,
- Be easily measurable,
- Vary as much as possible among speakers, but be as consistent as possible for each speaker,
- Not change over time or affected by speaker's health,
- Not be affected by reasonable background noise, nor be dependent on specific transmission characteristics,
- Not be modifiable by conscious effort of the speaker, or at least, be unlikely to be affected by attempts to disguise the voice.

Several kinds of features have been developed; some of them are discussed next.

Linear predictive coding (LPC)[97, 68, 96, 73]

A speech sample in a signal can be modelled as a linear combination of previous speech samples in the signal. The coefficients of the linear combination are features and are compressed further by using codebook methods, such as vector quantization (VQ). LPC spectral representations have

been used extensively for speaker recognition; however they can be severely affected by noise.

Filter bank response analysis [97, 10, 23]

The energy output of each filter in a filter bank serves as an element in a feature vector. The total energy in a small or large interval of speech can be used in conjunction with other features.

Adaptive Component Weighting (ACW) [4]

ACW scheme is motivated by the characteristics of the parameters of the parallel form of the all-pole model. ACW cepstral features coefficients emphasize the formant structure of the speech spectrum while attenuating the broad-bandwidth spectral components. The attenuated components correspond to the variations in spectral tilt of transmission and recording environment, and other characteristics that are irrelevant to speaker identification. In the cepstral domain, the ACW scheme can be viewed as an intraframe cepstral processing technique by which frame-dependent weights are applied to the LP cepstra.

Zero and level crossings [97, 106]

The number of zero crossings at a present level for an interval of a speech signal is taken as feature for the interval. Methods are researched to reduce noise inherent to the combination of zero crossing detection and discrete signals [99].

Formants [97, 10]

Formants refer to the natural or resonant frequencies of the vocal tract, which varies with time as speech is uttered. Formants are difficult to resolve if the speech levels are low, which require estimation methods to be researched.

Warped frequency scale methods

One of three different scales is typically used to construct a filter bank to extract spectral energies, which may be processed further to obtain cepstrum coefficients. Recent studies have found directly computed filterbank features to be more robust for noisy speech recognition [104, 103]. Mel-frequency filterbank [97, 54, 124] is detailed in this study since it was used for comparison with the EIH model.

These features are sensitive to noise in the signal, as well as to the type of microphone used. There are two approaches to improving features. The first is to modify the LPCs or cepstal coefficients to minimize the effect of noise. Introducing robust distortion measures can help to automatically emphasize those features that have less distortion, explained in section 1.6.

The human auditory system is a speech recognizer with excellent performance, especially in noisy environments. However, the human auditory system is known to extract noise-robust features, which is one of the essential ingredients for successful applications in the real world. Since humans perform better at ignoring noise than any machine, auditory-based methods are the promising approaches, since they attempt to model the working of the human auditory system. These front-ends have been shown to outperform more conventional signal processing schemes for speech recognition tasks [34, 113, 120]. These methods include:

Seneff auditory models [54, 113]

In this model, first –stage spectral analysis is performed with a bank of critical-band filters, followed by a model of nonlinear transduction in the cochlear that accounts for observed auditory features such as adaptation and forward masking [SMI46, 47]. The output is delivered to the average rate of

neural discharge, called the *mean rate response*, the other is a *synchrony response* which yields enhanced spectral contrast showing spectral prominence for formants, fricatives and stops.

Subband-autocorrelation (SUBCOR) [59]

SUBCOR analysis technique extracts periodicities present in speech signals by computing autocorrelation coefficients of subband signals at specific time-lags, and was shown to outperform the smoothed group delay spectrum for speech recognition tasks under noisy environments.

Perceptual linear prediction (PLP) [57, 14, 58, 106]

PLP analysis method is a perception based technique in which a short-term power spectrum of a speech signal is transformed into a more auditory-like spectrum, typically by means of a Bark-scaled spectral analysis [58], before performing conventional linear prediction (LP) analysis. After which an autoregressive all-pole model approximates the spectrum. The robustness of the PLP analysis to additive noise was reported in [57].

Zero-crossings with peak amplitudes (ZCPA) [60, 62]

ZCPA is a simple and efficient auditory model, proposed as a robust front-end for speech recognition systems in noisy environments. The ZCPA model consists of a bank of bandpass cochlear filters and nonlinear stages at the output of each cochlear filter. The cochlear filterbank represents frequency selectivity at various locations along a basilar membrane in the cochlear. Neural firings are simulated as the upward-going zero crossing events of the signal at the output of each bandpass filter. Each peak-amplitude between the successive zero-crossings is used as a nonlinear weighting factor to a frequency bin to simulate the relationship between the firing rate and stimulus intensity. The histograms across all filter channels are combined to represent the output of the ZCPA model. Unlike the EIH model, developed in this study,

ZCPA model is free from unknown parameters associated with the level and level values.

Ensemble interval histogram [97, 34, 36]

EIH employs a coherence measure as opposed to the direct energy measurement used in conventional spectral analysis. It is effectively a measure of the spatial extent of coherent neural activity across a simulated auditory nerve. The EIH is computed in three stages – bandpass filtering of speech to simulate basilar membrane response, processing of the output of each filter by level-crossing detectors to simulate inner hair cell firings, and the accumulation of an ensemble histogram as a heuristic for information extracted by the central nervous system.

3.2 Mel-Cepstrum

The mel-warped cepstrum is calculated using a filter-bank approach in which the sets of filters are of equal bandwidths with respect to the mel-scale of frequencies. For *speaker recognition*, feature vectors such as predictor coefficients, line spectral pairs, log area ratios, vocal-tract functions, and cepstral coefficients were compared and the cepstral coefficients were found to provide best results [5]. The mel filter bank based cepstral transformation, when performed for *word recognition*, has been shown to outperform other conventional signal processing methods [16]. Another study found cepstral coefficients to outperform log area ratios (LAR) for speaker verification [27]. Mel-cepstrum demonstrated to perform almost 100% in the large population of 630 speakers for speaker identification task in a clean speech. For these reasons, mel-cepstrum is chosen for comparison in this study.

3.2.1 Mel Scale

Psychophysical studies have shown that human perception of the frequency content of sounds for either pure tones or speech signals, does not correspond to a linear scale. The human ear is more sensitive to lower frequencies than the higher one [121]. For

each tone with a certain actual frequency in hertz, a subjective pitch is measured on the "mel" scale. The pitch of a 1 kHz tone at 40 dB higher than the perceptual hearing threshold is defined as 1000Hz mels. The subjective pitch is essentially linear with the logarithmic frequency above 1000Hz.

Mel accounts for this frequency sensitivity by first filtering speech with a filterbank that consists of filters that have increasing bandwidth and center-frequency with increasing frequency [16].

3.2.2 Computation of mel-cepstrum

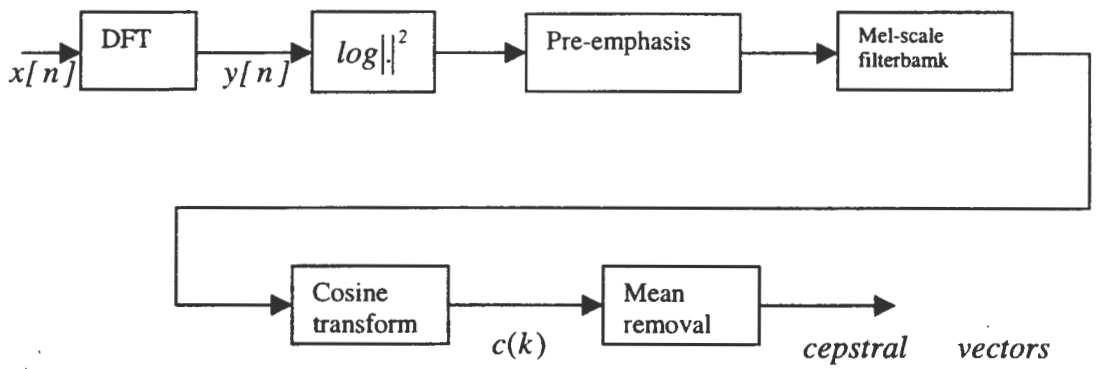

Figure 3-1: Mel-scale cepstral feature analysis

Mel-frequency spaced filterbank cepstral coefficients (MEL) are filter bank features derived from the DFT spectrum. MEL is computed in a standard manner [16, ROS93], as shown in Figure 3-1. The input speech at 16 kHz is first multiplied by a 32-ms-long Hamming window every 16 ms. A discrete Fourier transform (DFT) is computed for each window waveform segment. In the frequency domain, a vector of log energies is computed from each waveform segment by weighting the DFT coefficients by the magnitude frequency response of a filter bank. The magnitude spectrum of short-term speech is pre-emphasized (see section 5.4). The center frequencies of the filters are spaced equally on a linear scale from 0 to 1 kHz and above 1 kHz are spaced logarithmically, each center frequency is 1.1 times the center frequency of the previous filter, giving more detail to the low frequencies [128]. Each

filter's magnitude frequency response has a triangular shape that is equal to unity at the center frequency and linearly decreasing to zero at the center frequency of the two adjacent filters [123]. This is the traditional mel-cepstrum.

3.3 Parameterized feature-set

The parameterized feature-set (PFS) is similar to the mel-cepstrum feature-set described above, except that it does not use a mel-scale filterbank. This method removes the pitch contribution in the spectral domain before conversion to the cepstral domain [77]. This process is called *liftering process*. Liftering process removes the pitch effects by liftering the spectrum with a 41-point optimal FIR filter. Liftering process is important since the recognizer is used in a speaker-identification task. The parameterized feature-set has been used for the speech recognition task [78]. This set of features incorporates spectral compression and filtering. The spectrum for a particular set of two parameters α and β are sampled nonlinearly, such that

$$A\sum_{i=1}^{\alpha}\beta^{i-1} = N/2 \qquad (3.1)$$

where N is the size of a DFT spectrum, A is a constant greater than 1, and the parameters α and β characterize the spectral compression of the feature-set. Different combination of α and β gives different performance of the system. Values of α and β can be defined such that α = [4, 6, 8, 10] and β = [1.0, 1.2, 1.5, 1.7, 2.0] etc. In the experiments the chosen values are α = 4 and β = 1.7. The parameterized feature-set was researched by Mashao [77] and is used in this thesis for comparison.

The spectral representations are transformed to cepstral coefficients (see section 3.4). This is done because of the (near) orthogonalizing property of the cepstral transformation. Recognizing that the cepstral feature vectors can be influenced by the response of the communication channel, cepstral mean removal is applied for channel equalization (see section 5.5).

3.4 Motivation of auditory models

The motivation for investigating spectral analysis methods that are physiologically based is to gain an understanding of how the human auditory system processes speech, in order to be able to design and implement robust, efficient methods of analyzing and representing speech. It is generally assumed that the better we understand the signal processing in the human auditory system, the closer we will come to being able to design a system that can truly understand meaning as well as content of speech.

Auditory-based methods try to model the ear and study the effects of applying speech signals to that model. They are motivated by the facts that human beings are able to understand speech under a variety of conditions, including noisy environments. By modeling the stages of auditory processing in the human, better ways of presenting speech may be obtained. Many models have been built for this purpose, including Ensemble Interval Histogram (EIH) [33].

3.4.1 The human auditory system

The auditory system is systematically organized around a midline between the left and the right sides. The most peripheral part at each side consists of the external and middle ears, the cochlea, and the auditory nerve. The auditory nervous system, which receives its inputs from both the left and right auditory nerves, consists of several neural nuclei that can be grouped into three major parts-the auditory brainstem, the auditory midbrain, and the auditory cortex. Each nucleus can be divided into different zones, characterized by their morphological structure, neurophysiological response, and their input/output mappings. The main information flow is along an ascending pathway that begins at the external ears and ends in the auditory cortex. In parallel, there is an information flow in a descending pathway that begins at the auditory cortex and can be traced all the way down to the middle ears[97].

3.4.2 Physiological basis for the EIH Model

The auditory periphery comprises three distinct parts: the outer ear, the middle ear, and the inner ear. The outer ear consists of the pinna (the ear surface surrounding the canal in which sound is funneled) and the external canal. Sound waves are guided through the outer ear to the middle ear, which consists of the eardrum (which moves due to the sound pressure) and a mechanical transducer comprises the hammer, the incus and the stapes (which conveys the motion of the eardrum into mechanical vibrations along the inner ear). The inner ear consists of the cochlear, which is a fluid-filled chamber partitioned by the basilar membrane, and the auditory nerve. The mechanical vibrations at the entrance of the cochlear (a 2 ½ turn, snail-like tube) excites the fluid inside the cochlear and cause the basilar membrane to vibrate at places associated with the frequencies of the input acoustic wave. Distributed along the basilar membrane (in a dense but discrete manner) are sensors called inner hair cells (IHC) that are innervated by the auditory nerve fibers and transform the mechanical displacement of the basilar membrane into a firing activity at the nerve fibers [36].

The mechanical displacement of the basilar membrane, at any given place, can be viewed as the output signal of a band-pass filter whose frequency response has a resonance peak at a frequency which is characteristic of the place. This resonance frequency is called *characteristic frequency* (CF) [35]. The log of CF is approximately proportional to distance along the membrane, and the distribution of the inner hair cells (IHC's) along the cochlear partition is essentially uniform. There are some 4000 IHC's along the basilar membrane. The displacement of the basilar membrane is reflected in the AC component of the IHC receptor potential. The transformation from mechanical motion to receptor potential involves several nonlinearities, the most relevant to the present discussion being the half-wave rectification which is a consequence of the unidirectional depolarization of the IHC. Each IHC is innervated by approximately ten auditory-nerve fibers, whose spontaneous activity ranges between 0 and 100 discharges per second. The spontaneous rate is highly correlated with the fiber diameter and the size of the synaptic region between the fiber and IHC. The spontaneous discharge rate is also

correlated with the threshold of response [34]. For any given CF region, fibers with spontaneous rate tend to have between 5 and 20 dB lower threshold than units with low rates of background activity. Occasionally, low-spontaneous units may have as much as 40-60 dB higher threshold than high spontaneous units of comparable CF [71], [72].

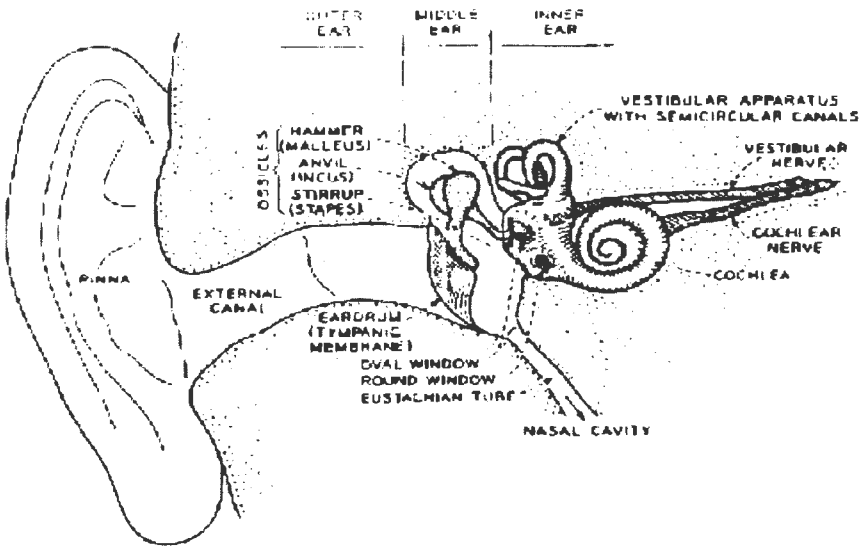

Figure 3-2: Physiological model of the human ear.

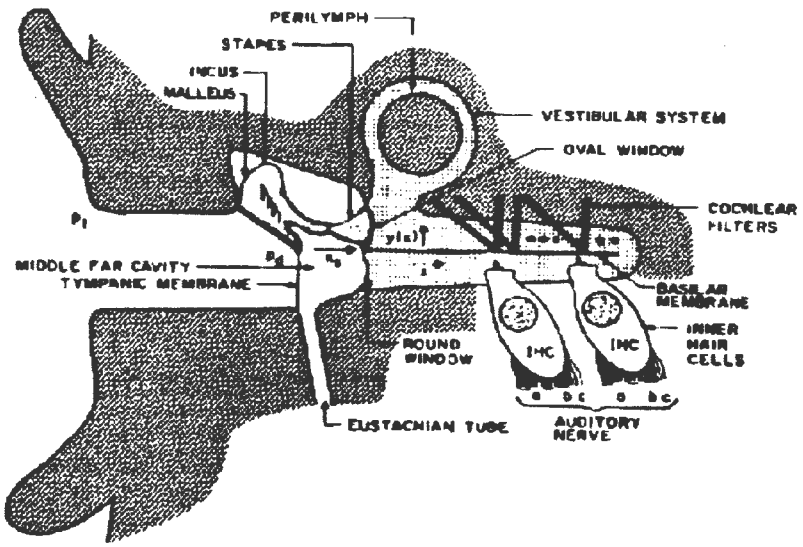

Figure 3-3: Expanded view of the middle and inner ear mechanics.

3.4.3 Computation of EIH

Ensemble interval histogram (EIH) [97], [34], [36] employs a coherence measure of the spatial extent of coherent neural activity across a simulated auditory nerve. The processing approach is motivated by conclusions drawn from the gross characteristics of the cat's auditory-nerve firing patterns. The EIH is computed in three stages – bandpass filtering of speech that models the frequency selectivity at various points to simulate basilar membrane response, non-linear processing of the output of each filter by level-crossing detectors to simulate inner hair cell firings, and the accumulation of an ensemble histogram as a heuristic for information extracted by the central nervous system [97] as shown in Figure 3-4.

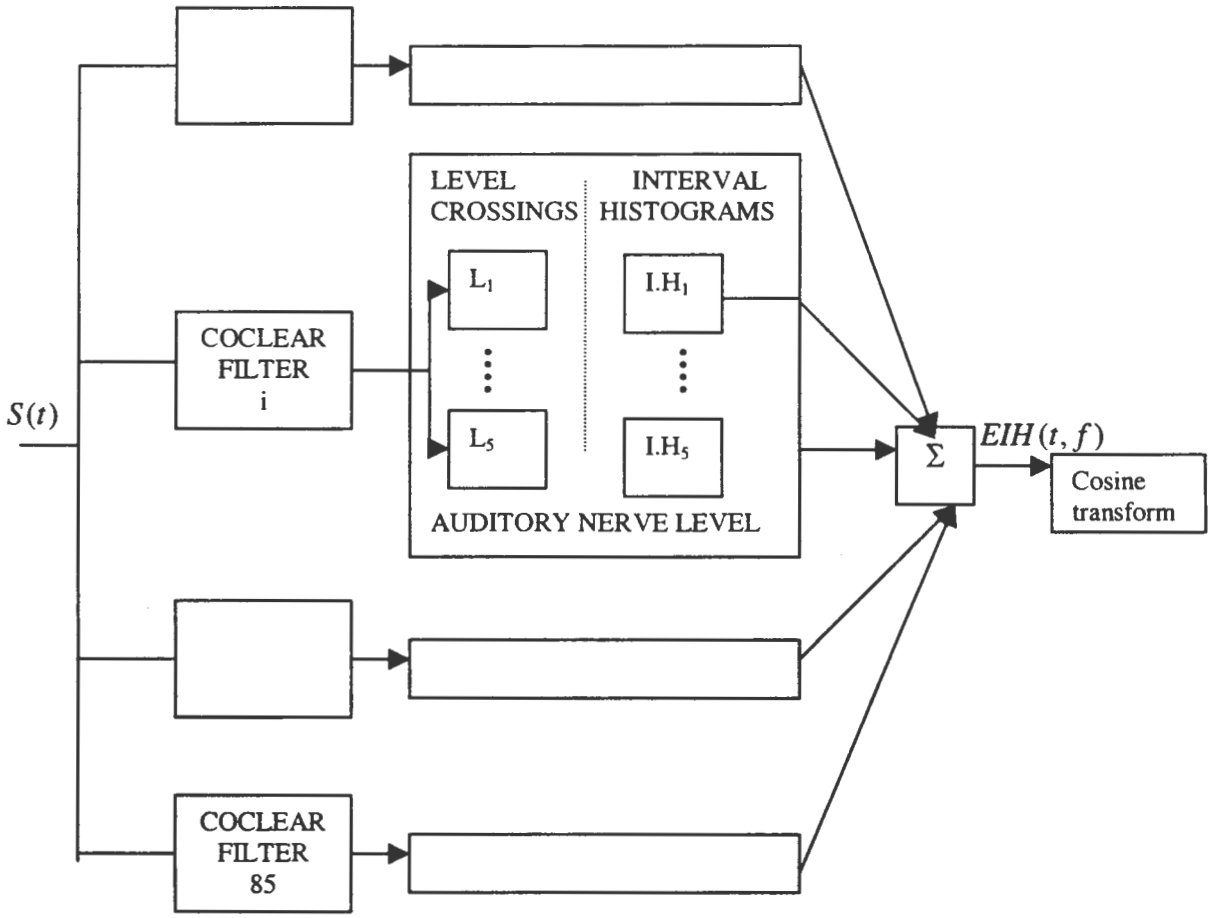

Figure 3-4: Block diagram of the EIH model.

The three stages of EIH model are discussed in the next section.

A. Filter generation

In this model, 85 cochlear filters, equally spaced, on a log-frequency scale, between 150 and 7000 Hz, are used to sample the mechanical motion of the basilar membrane. These filters are based on actual neural tuning curves for cats. The amplitude responses of 28 filters (one every 6) of these filters are shown in Figure 3.5.

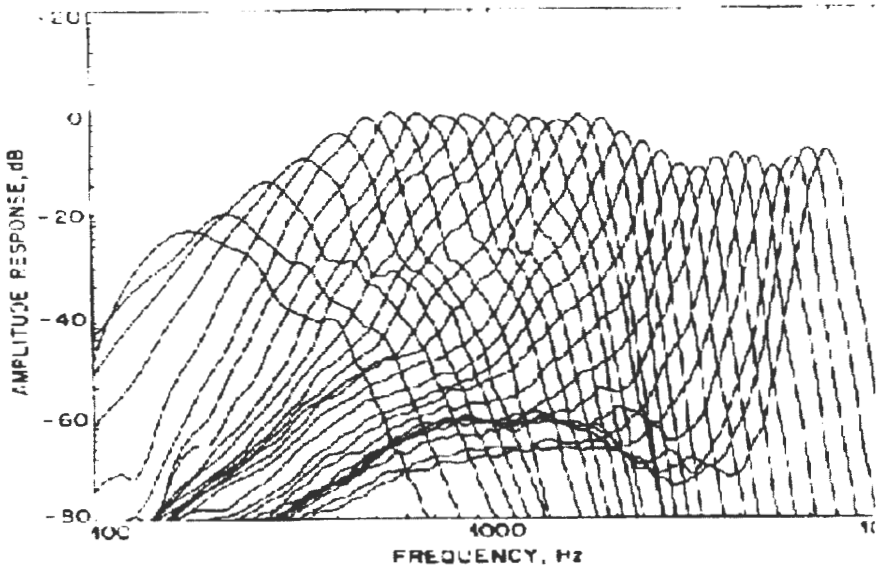

Figure 3-5: The amplitude responses of 28 (one every eight) simulated cochlear filters in a log frequency/decibel scale.

The corresponding cochlear filters have been simulated in detail, using actual turning curves for cats [71]. The phase characteristics of these filters is the minimum phase and the relative gain, measured at the center frequency of the filter, reflects the corresponding value of the cat's middle ear transfer function [34].

B. level-crossing detectors

Moreover, the non-linear processor, which converts the filter bank output to neural activity, is an array of level-crossing detectors, placed after the output of each cochlear filter. The number of levels is chosen to be five. The detection levels of each detector are pseudo-randomly distributed, based on measured distributions of level firings, thereby simulating the variability of fiber diameter and their synaptic connections. The level crossings are measured at positive threshold levels which are

uniformly distributed in a log scale over the dynamic range of the signal. Only intervals between successive upward-going level crossing are considered. The time intervals between the levels crossing reflect the periodicity of the stimulus.

The counter is set for each level detection, and this counter is similar to a AN fiber, which emits a neural impulse whenever it is activated. The output of the level-crossing detectors represents the discharge activity of the auditory nerve fibers which, in turn, is the input to a more central stage of neural processing, which gives the overall EIH.

C. Histogram accumulation

An estimate of the interval probability density function of a given level can be obtained by computing a histogram of the intervals from the point process data produced by the level-crossing detector. This is accomplished by distributing the reciprocal of the intervals in a histogram consisting of successive bins. The similarity between all individual density functions is measured by summing the corresponding histogram bins across all levels and all channels to find one EIH. Hence the output of the EIH models the neural spikes produced in the human brain and which is believed to contain the relevant information for speech recognition.

3.4.4 Characteristics of the EIH model

Several factors affect the properties of the interval histogram, and they are briefly discussed. A *bin allocation* and a *window length* that are matched to the bandwidth characteristic of the cochlear filter provide a unified representation that exhibits fine resolution at low characteristic frequencies (CF's), and fine temporal resolution at high CF's.

Bin allocation can either be according to ERB-rate scale [82], which is motivated by the tonotopic organization along the auditory pathway or linear. Error distribution of the EIH with linear bin-allocation better predicts the error distribution of the ERB-scale [36]. For this reason, *linear bin allocation* is chosen in this study. Using a linear-

frequency scale turns out to be advantageous for recognition since only quarter of the number of bins is used for the first 1000 Hz. The improvement of the system in performance is due to the difference between the feature vectors, which is mainly dictated by the nature of bin allocation.

Another parameter that affects the properties of the interval histogram is the size of the observation window. Motivated by the tonotopic organization along the auditory pathway, the window length is set to be inversely proportional to the center window [33]. In this study the window length is 40/CF to capture the about 40 periods of the signal at each channel, CF.

Proper determination of the number of levels and level values is very important for reliable performance of the EIH model especially in noisy environments. However, there is no method available to determine these values, except by trial-and-error (see section 5.4 A).

Another feature of the EIH representation is its inherent dominance property. Because of the shape of the cochlear filters, a high-amplitude spectral component f_1 is capable of dominating the output of a cochlear filter whose CF is appreciably greater than f_1, even in the presence of a second less intense component, f_2, equal to the filter CF. When such synchrony occurs, the output of this channel counts to the EIH at bin f_1 rather than bin f_2. The precise number of counts contributed by any given channel to the f_1 bin of the EIH depends on the degree to which its output is dominated by that frequency [34].

Conceptually, the EIH is a measure of the spatial extent of coherent neural activity across the simulated auditory nerve. Mathematically, EIH is the short-term probability density function of the reciprocal of the intervals, measured over the time-frequency domain [97].

3.5 Data Reduction and Smoothing

The output of the MEL, PFS, EIH front-ends is considered as "pseudo-spectra". Therefore these pseudo-spectra are then processed to generate feature vectors for the speech recognizer with the goal of reducing recognition error rate across test conditions. The dimensionality of the feature vectors produced by the front-ends is reduced by the *discrete cosine transform*. The three-dimensionality feature-reducing techniques are outlined next.

A. Principal Components Analysis

Principal components analysis (PCA) [54] is the linear transformation on an input feature space, producing modified feature spaces, $\vec{x}_i'$

where

$$\vec{x}'_i = \hat{A}\vec{x}_i, \qquad (3.2)$$

$\vec{x}_i$ is ith input feature vector

$\hat{A}$ is the transformation matrix

The rows of $\hat{A}$ are the eigenvectors of the covariance matrix for x_i and $\hat{A}$ is determined such that individual elements of $\vec{x}_i'$ are uncorrelated. Only the eigenvectors corresponding to the largest N eigenvalues of the covariance matrix, where N is the largest number of the output features desired, is used to reduce the dimensionality of the feature vector.

B. Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) [26] is a linear transformation on the input feature space. It uses the same method as PCA to reduce the dimensionality of the feature vectors, but maximizes some measure of class separability. Transformation matrix is applied to all training and testing speech frames as with PCA.

C. *Discrete Cosine Transform*

Cepstral coefficients show a high degree of statistical independence and a good performance has been achieved with both speaker identification/verification systems [102]. Cepstral coefficients, *C(k),* are generated by

$$C(k) = \sum_{i=1}^{N} f(i)\cos(k(i+0.5)\frac{\pi}{N}) \quad 1 \le k \le N \quad (3.3)$$

Where N is the desired length of the cepstrum. The lowest cepstral coefficient (C_0) is a measure of the average log energy in the speech frame. C_0 is used as an estimate of the loudness at time t_0. Since we are not considering the effect of the signal intensity, C_0, is not used for all front-ends in the cepstral feature vector, instead an added extra coefficient replaces it. The cosine transform is applied to obtain the autocorrelation function. The degree of smoothing depends on the number of autocorrelation function. Thus, cosine transform compresses the spectral information into lower-order coefficients and also decorrelates them to allow the subsequent statistical modelling to use diagonal covariances matrices [128].

Cepstral-order

Determining the number of coefficients needed to model a speaker adequately is an important but difficult problem. There is no theoretical way to estimate the number of coefficients *a prior* [101]. Thus, the cepstral representations of *30 coefficients* are used for both PFS and EIH. In the preliminary experiments, a different number of cepstral coefficients is conducted in table 5.1, and the order of 30 was shown to give the better results. This order of fit was required because of the larger dynamic range of both front-ends. The cepstral coefficients are then used as input features to the speech recognizer.

3.6 Comparison of MEL and EIH

MEL and EIH front-ends both begin with linear filter banks; by plotting each channel's output versus the center frequency of the channel's first stage linear filter, a spectrum-like representation can be achieved. In the case of MEL, the filters are not used to filter the incoming speech as such as in the case of EIH, they are used as filters in the conventional sense. The filters of both MEL and EIH differ. MEL uses identically shaped filters, equally spaced in a frequency band determined by the sampling frequency. The filter banks of EIH are equally spaced on a logarithmic scale and the shape of the filters is very different. The primary difference between these two front-ends is the nonlinear processing that occurs after the filter bank, namely the energy based approach versus the coherence-based approach, Figure 3-6. In the second stage of MEL, the power spectrum estimate is achieved by computing the short-term power at the output of each filter. Whereas in the second stage of EIH, timing synchrony analysis is based on measuring coherence across a multidimensional point access produced by passing the filter bank output signals through the array of multidimensional detectors.

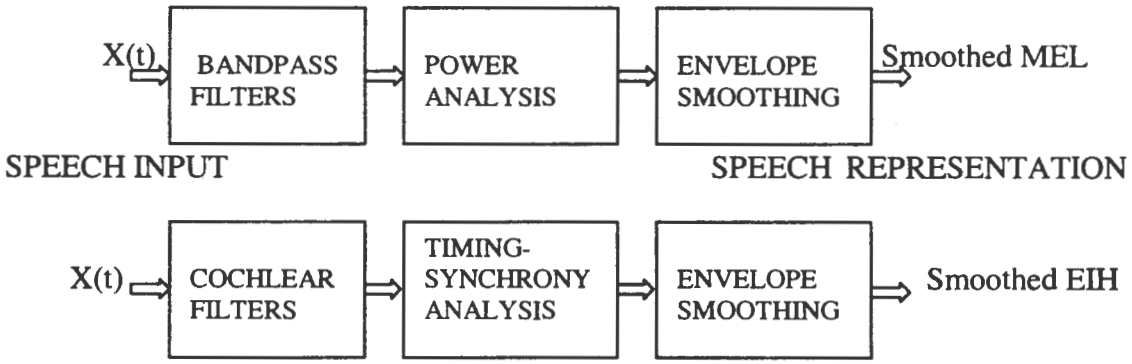

Figure 3-6: Comparison of MEL and EIH.

Ghitza [34] showed that the robustness of the EIH in the presence of noise is due mainly to the properties of *the timing-synchrony analysis* rather than the unique shape of the cochlear filters. The outputs of the EIH and MEL are pseudo-spectra and they are then processed to generate feature vectors for the speech recognizer. This is the third stage called *envelope smoothing.*

3.7 Summary

This chapter discussed the characteristic features of the speech representations, the Mel cepstrum analysis, parameterized feature-set, and the Ensemble Interval Histogram. The motives for choosing these two front-ends have been briefly discussed and their comparison have been made. The procedures for computing their respective parameters have been discussed. The next chapter discusses the Speaker identification system that processes the features.

CHAPTER 4

Speaker Identification

The main purpose of speech communication is to convey messages. It follows, therefore, that the message is the most important information embedded in the speech signal. But it is not the only information, other kinds of information embedded in the speech signal include the *identity of the speaker, the language spoken, the presence and type of speech pathologies, and the physical and emotional state of the speaker.* Of these other kinds of information, *the identity of the speaker* has received most of the attention. This is because speaker characteristics impose striking and comprehensive effects on spoken utterances and because there are many practical applications awaiting practical and effective methods of automatically extracting speaker identity information from speech signals.

The identity of people can manifest itself through many ways; among them one can find body aspect, personality traits, fingerprints, etc. The aim of speaker recognition is to use voice as the only basis for guessing the identity of the talker.

Automatic speaker recognition comes traditionally in two flavours and two colours. The two flavours are *speaker identification* and *speaker verification*, and the two colours *are text-dependent* and *text-independent*. These terms are explained as a background in section 4.1. The SID problem is introduced in section 4.2, the methods to solve the problem in Section 4.3, the advantage and disadvantage of these methods in section 4.4. Finally, the recognition system implemented for this study is discussed in section 4.5.

4.1 SID Background

This section provides the speaker recognition background that consists of speaker verification, identification, speaker-dependent, speaker-independent.

4.1.1 Speaker-Recognition

Speaker recognition systems can operate in either an identification or a verification decision mode. Automatic speaker verification (ASV) [108] and automatic speaker identification (ASI) [6] are related but different areas of voice recognition. Speaker verification is concerned with "*Is this the person who he/she claims to be*?", whereas speaker identification asks the question "*Who spoke this*?" Both ASI and ASV use a stored database of reference patterns, for N known speakers, and similar analysis and decision techniques are employed.

ASI can be subdivided into two categories, *closed-set* and *open-set problems*. The closed-set problem requires choosing which of the N voices known to the system best matches a test voice. Naturally, the larger N is, the more difficult the task is. Alternatively, one may want to decide whether the speaker of a test utterance belongs to a group of N known speakers. This is called the open-set problem, since the speaker to be identified may not be one of the N known speakers. ASV is a special case of an

open-set problem and requires comparing the test pattern against one reference pattern and involves a binary decision whether the test speech matches the template of the claimed speaker. If a speaker scores well enough on the basis of a test utterance, then the target speaker is accepted as being known.

The fundamental difference between the identification and verification modes is the number of decision alternatives. In the identification mode the number of decision alternatives is equal to the size of the population, whereas in the verification mode there are two decision alternatives, *accept* or *reject* the identity claim, regardless of the size of the population [18]. Since N comparisons and decisions are necessary, the error rate can be much higher for ASI than for ASV [115].

Mathematical definitions

Since mentioned earlier, this study concentrates on speaker identification problem. Speaker identification is determining who spoke a recorded utterance through speech analysis. We now make mathematical definitions of the above terms.

Let $S = \{S_1 \ldots S_N\}$ be a set of N speakers

Let $U = \{U_1 \ldots U_N\}$ be a set of utterance set of speakers

Let U_x be an utterance by speaker S_x

then

Closed set SID determines the value for x in $[1 \ldots N]$.

Open set SID determines the value for x in $[0 \ldots N]$,

where $x=0$ means S_x is not in S.

Text independent SID means U_x is not necessary in U.

This thesis deals with closed-set ASI problem, i.e. the problem of identifying one speaker among a known set of speakers.

4.1.1 Speaker-dependence Recognition

The most reliable way to tackle the individuality problem and achieve high recognition performance is to train the system using the new speaker's voice. This method is called the *speaker dependent method.* A speaker-dependent system is created especially to operate for a single speaker. Several techniques [6, 108, DOD73] are available and take advantage of speaker co-operation by making use of a password [76]. In the training phase, the speaker teaches the machine his very own way of pronouncing a fixed sentence; in the recognition phase, the actual pronunciation of the password is matched against the learned one, and a score is established which gives the similarity of the two sentences. A threshold decides if they are issued from the same speaker or not. This system is usually easier to create, develop, cheaper to buy and more accurate. The system is trained comprehending pronunciations, inflections, and accents. It can be customized efficiently and accurately for any particular speaker.

Speaker dependent systems are desirable in voice command applications such as telephone dialing. Another application is voice verification, where the unique physical characteristics of a particular voice are used to authenticate an individual. This is highly relevant to banks and is the subject of another financial futures web page.

This system has limitations and is not as flexible as the speaker-independent system. These problems include [67]:

1. The training is inconvenient for users.
2. A large amount of processing is necessary before the system becomes available for use. Sometimes users must wait for several hours after uttering training speech for the processing to be completed.
3. There are several applications in which training cannot be performed, such as telephone directory assistance and banking services.
4. In some applications, such as the dictation of conferences and trials, more than one speaker is present.
5. Storing the parameters of all speakers individually requires an extremely large amount of storage in some cases.
6. Each speaker's voice is essentially variable under the influence of stress, tiredness, illness, differences in location and characteristics of telephone sets and

microphones, variations in background noise, room acoustic characteristics, and frequency band limitation, and so on. Voice also change over time.

To reveal its full power and flexibility, speech recognition systems should be speaker-independent and should accept natural language and unrestricted lexicon.

4.1.2 Speaker-Independence Recognition

A speaker independent system is created for any speaker, with any accent or dialect, to communicate with a computer using continuous speech, a large vocabulary, and increasingly natural language patterns. These systems are more difficult to develop, more expensive and the extent of accuracy is greater, but more flexible than speaker dependent systems.

The important point about this technology is that it can be used by anyone (or at least anyone with access to a telephone), anywhere, without any prior training or special equipment. Good dialogue design is critical, and many other issues need to be carefully considered, such as customer authentication, interaction with other delivery channels, and links to host systems.

Methods for recognizing speech in a speaker independent fashion have therefore been investigated by many researchers. By using a reference template obtained by averaging the spectral patterns of many speakers for each word, high recognition accuracy can be achieved with only a small vocabulary and no similar word pairs in the spectral domain. If it is necessary to recognize accurately a large vocabulary uttered by unrestricted speakers, the recognition system must incorporate more sophisticated functions for coping with voice individually.

4.2 The SID problem

Speaker identification systems answer the question "*who spoke this utterance*?" Such a system has as input a speech waveform $x[n]$, as well as a *model* M for each candidate speaker, in order for the system to have some representation of the possible

speakers. Figure 4-1 shows a block diagram of a generic SID system. The system is correct if the chosen speaker S_j is the speaker S_i that was the source of $x[n]$.

The SID process

- Extract acoustic features sets $\{F_1 \ldots F_N\}$ from the utterance $\{U_1 \ldots U_N\}$ and make a model.
- In the same manner, extract a feature set F_x from the unknown utterance U_x.
- Apply a matching algorithm to determine which, in any model of $\{F_1 \ldots F_N\}$ matches F_x.

The performance of SID systems is typically determined by presenting a large number of test utterances $x_{test}[n]$, each being produced by a speaker S_i, and evaluating the percentage of $x_{test}[n]$ for which the SID system identified S_i correctly; that is; the percentage of test waveforms for which $S_i = S_j$.

Figure 4.1 can be expanded by breaking it into a front-end F and a pattern classifier P, as shown in Figure 4.2. The output of F is a vector $\vec{f}$ of features that P will use to make the speaker discrimination.

In practice, the front-end F and pattern classifier P are related to the extent that the structure of the models M_i and therefore the pattern classification machinery is dependent on the behaviour of the features $\vec{f}$ for the speakers to be identified. If $\vec{f}$ clusters very well by speaker identity, the pattern classification can be relatively simple.

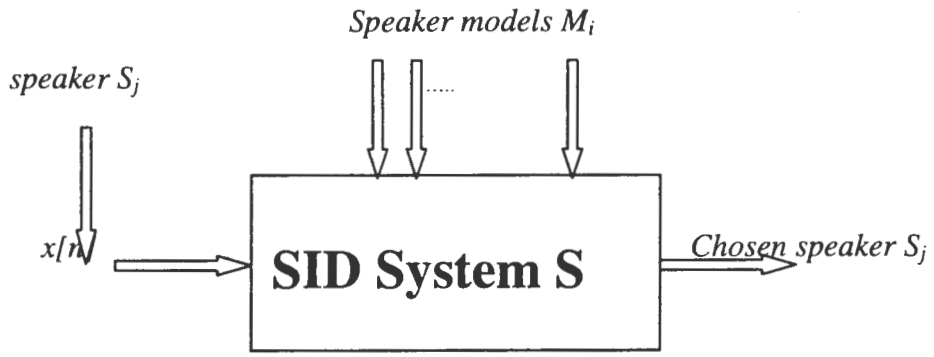

Figure 4-1: Block diagram of SID system

This thesis focuses only on changes to the front-end *F;* the pattern *P* is viewed as fixed. Although almost every current work with SID systems involves optimizing *P*, there is a good reason to believe that improving *F* can improve overall SID performance.

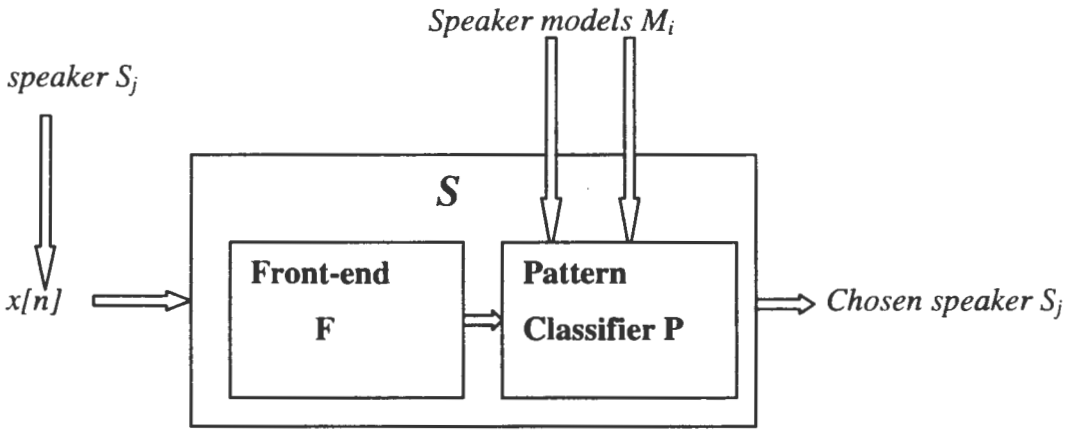

Figure 4-2: Expanded Block diagram of SID system

4.3 Speaker-independent Recognition Methods

A number of approaches have been attempted in an effort to build speaker-independent recognition systems. These include the signal processing approach, information theoretic approach, pattern recognition approach, artificial intelligence approach, statistical approach, and neural network approach. Each of these is briefly outlined next.

Signal processing Approach-Invariant features

The most basic and desirable approach to speaker independence is to find feature parameters that are invariant between speakers, that is, commonly extracted from the same phonemes or words uttered by all speakers and effective for separating different phonemes and words. This approach is very ambitious, since there is no evidence that common physical features exist in the same words uttered by different speakers even if they can be clearly recognized by humans [30].

Information theoretic approach – Discriminant methods

Information theoretical approach gave rise to discriminant features and distant measures that are effective for discriminating phonemes uttered by unspecified speakers. Recently this approach has been actively investigated in the framework of HMM recognition systems. In this approach, the mutual information methods and various discrimination methods, including corrective training and sophisticated decision boundary-adjustment techniques have been tried [76]. One of the problems associated with these methods is that they are very sensitive to the difference between training and testing samples, since the recognition strategies are elaborately tuned to the training samples.

Pattern recognition approach-Multiple template method

The multiple-template method, in which multiple templates representing individual variables are designed by clustering techniques to cover the range of individual variation, is one of the successful methods for speaker-independent recognition. Each word in the vocabulary is uttered by many speakers, and these utterances are then divided into clusters. Several algorithms are then used in combination to achieve clustering. The speech sample at the center of each cluster or the mean value for the speech data associated with each cluster is stored as a reference template.

In the recognition phase, distances between an input utterance and all multiple templates of all vocabulary are calculated, and the word with the smallest distance is selected as the word spoken. When the vocabulary size is large, the multiple-method approach has the problems of requiring huge storage for reference templates and extensive computation, since the multiple templates equivalently increase the vocabulary size. To reduce the amount of computation, VQ-based pre-processing has been combined with the SPLIT method [29, 31].

Artificial intelligence approach-Knowledge engineering

Phonetic features that are robust against speaker-to-speaker variation have been investigated using knowledge engineering approaches based on spectrogram-reading experiments. This approach attempts to express the speech knowledge within a formal framework using well-defined mathematical tools. In the SUMMIT system [129], features and decision strategies are discovered and trained automatically, using a large database.

Statistical approach – HMM models

Hidden Markov models can probabilistically represent efficiency and flexibility of the variations of phonetic features and transitions between them, and these feature are trained using utterances by many speakers. The HMM model has another advantage in that it uses a relatively small amount of computation for recognition, since each word or phoneme is represented by a sequence of a small number of states. One of the most successful examples of HMM-based speaker-independent continuous speech recognition system is the SPHINIX large-vocabulary system [66, 67]. This system achieved speaker independent word accuracy of 95.8% on the DARPA resource management task with grammars of perplexity, 20.

Neural networks approach

The dynamic programming neural network (DNN), based on the integration of dynamic programming (DTW) and multilayer neural networks, has been proposed as a method for SID word recognition [SOK89]. A high recognition accuracy of 99.3% was obtained for isolated Japanese digit recognition using this method.

4.4 Advantages and Limits of SID Methods

The SID approach has the advantage that a large database for building the recognition system can easily be obtained in comparison with the speaker-dependent approach. One of the disadvantages of the SID approach is that it neglects various useful characteristics of the speaker in spite of the fact that they can be learned after the recognition of several words or sentences. If these characteristics can be properly used, the recognition process is expected to be accelerated due to the narrowing of the search space. Another disadvantage is that when the distributions of feature parameters are very broad or multimodal, as in the cases of a combination of male and female voices as well as various dialects, it is difficult to separate phonemes using SID methods. Another practical difficulty in generating speaker-independent recognition systems is that model specificity and model trainability are two compatible goals when only a fixed amount of training data is given. More specificity usually reduces trainability, and increased trainability usually results in overgenerality.

The VQ approach has demonstrated good performance on limited vocabulary tasks, but it is limited in its ability to model the possible variabilities encountered in an unconstrained speech task. HMM has been used as probabilistic model for both text-independent and text-dependent speaker recognition [79]. HMM models the temporal sequencing among speech sounds which is advantageous for text-dependent task but for text-independent tasks the sequencing of sounds found in training data does not necessarily reflect the sound sequences found in the testing data and contains

little speaker-dependent information. Several different neural networks (NN), such as multilayer perceptrons, time-delay NN, and radial basis functions, have been applied to various speaker-recognition tasks. NN's have produced good speaker recognition performance comparable to that of VQ systems [8]. Their major drawback is that the complete network must be retrained when a new speaker is added to the system.

Gaussian mixture speaker model provides a probabilistic model of the underlying sounds of a person's voice. The probabilistic framework allows the application of newly developed noise and channel robustness techniques from the speech recognition area. Furthermore, this model is computationally efficient and can easily be implemented on a real-time digital signal processor. The Gaussian mixture speaker model outperformed the vector quantization (VQ), unimodal Gaussian classifier (GC), radial basis function (RBF) [104]. These results indicate that Gaussian mixture models (GMM) provide robust speaker representation for the difficult task of speaker-identification using corrupted, unconstrained speech. Next is the detailed discussion of Gaussian mixture model [100] since it has been chosen as the recognition system.

4.5 Gaussian mixture model SID

The use of Gaussian mixture models (GMM's) for speaker-identification was shown to provide superior performance compared with several techniques [104].

Speaker-recognition is a subset of pattern recognition. Three stages are generally involved in building a pattern-recognition system; training, testing, and implementation. In the training stage, a set of parameters of the model is estimated so that in some sense the model learns the correspondence between the features and the labels of the objects. One such learning criterion is to minimize the overall estimation error. In the testing stage, the parameters of the model are then adjusted using a set of cross-validation data to achieve a good generalization of the performance of the system. The cross-validation data usually consists of a set of features and labels that are different from the training data. The task of recognition is carried out in the implementation stage, where the feature with an unknown label is passed through the system and assigned a label at the output.

Two stages of evaluating a SID system; training speaker models and testing the systems utterances of known speakers are discussed next for the specific system used in this thesis.

4.5.1 Training Speaker Models

The task of automatic speaker identification is to determine the identity of a speaker by machine. In order for humans to recognize voices, they must be familiar with the voices. It is the same case with the machines. The process of *getting to know* speakers is referred to as *training* and consists of collecting data from utterances of people to be identified.

Before testing, a model M_i representing each candidate speaker S_i must be generated. The SID system used in this work incorporates the following probabilistic model:

$$P(\vec{f}|S_j) = \sum_{i=1}^{M} c_i^{S_j} N(\mu^{S_j}, \Lambda^{S_j}), \tag{4.1}$$

$$N(\vec{\mu} \mid \Lambda) = \frac{1}{(2\pi)^{\frac{n}{2}} |\Lambda|^{\frac{1}{2}}} e^{-\frac{1}{2}(\vec{f}-\vec{\mu})^T \Lambda^{-1}(\vec{f}-\vec{\mu})} \tag{4.2}$$

Where $\vec{f}$ is the feature vector computed by the front-end F and $N(\mu,\Lambda)$ is a *multivariate Gaussian distribution*, also referred to as a *multivariate normal distribution* [20].

$|\Lambda|$ and Λ^{-1} indicate the determinant and inverse, respectively, of *covariance matrix* Λ. μ is the *mean of the distribution* $N(\mu,\Lambda)$. Since $P(\vec{f}\,|S_j)$ is the sum of M over multivariate Gaussian distribution, the model in equation 4.1 is known as a *Gaussian mixture model (GMM)*. c_i are the *mixture weights*, which must sum to unity for $P(\vec{f}\,|S_j)$ to be normalized.

Given training-speech from a speaker, the GMM parameters ($c_i^{S_j}$, μ^{S_j} and Λ^{S_j}) needs to be estimated. The most popular and well-established method is *maximum-likelihood* (ML) estimation [80]. However, ML parameter estimate can be obtained iteratively using *expectation-modification (EM) algorithm* [17] to maximize the probability of observing the features from the training data for speaker S_i with model M_i. For preliminary experimental tests, EM algorithm was used and slight improvement of results of 1-3% was obtained. But due to the computational expensiveness, EM algorithm was not used.

In practice, Λ^{S_j} is typically assumed to be diagonal. This property is necessary due to the notorious difficulty of training the non-diagonal elements of covariance matrices. In actuality, the elements of $\vec{f}$ are decidedly not uncorrelated; the use of multiple mixtures, though, allows modelling correlated data as a sum of M Gaussian distributions, each one with uncorrelated features. Figure 4-3 shows how a correlated distribution can be modelled by a number (in this case, 5) of uncorrelated mixtures, in two dimensions. The training of the experiment to find model parameters estimates was conducted through the use of k-means clustering algorithm. More details of training and testing procedures can be found in [100].

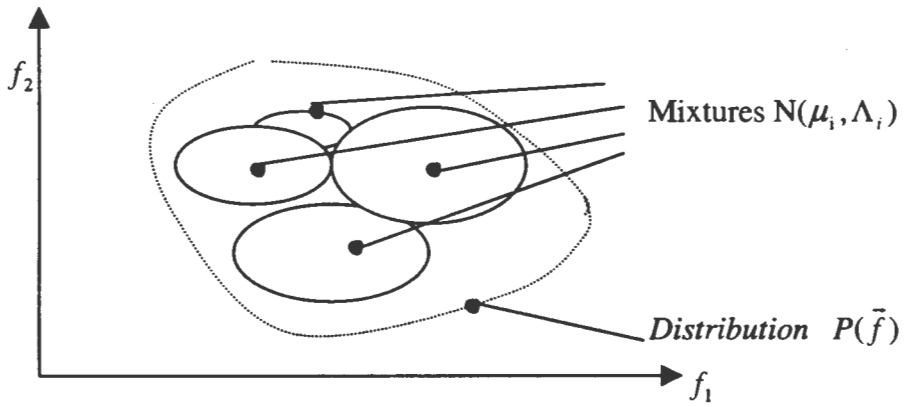

Figure 4-3: Modelling a complex distribution with Gaussian mixtures, each with a diagonal covariance matrix.

4.5.2 Testing the System

Testing is the task of comparing an unidentified utterance to the training data and making the identification. The speaker of a test utterance is referred to as the *target speaker*. Given speaker-models M_i, the system can now be tested. For each testing utterance $x[n]$, features $\vec{f}_{test}$, are calculated, the probability of each speaker-model given the features is evaluated, and the speaker is chosen with the highest such probability:

$$\max_i P\left(M_i \mid \vec{f}_{test}\right) \tag{4.3}$$

This technique is called *maximum a posterior probability (MAP)* classification, since $P(M_i|\vec{f}_{test})$ is known as the a *posterior probability*. $P(M_i|\vec{f}_{test})$ cannot be directly evaluated; by Bayes rule [19], is equal to

$$\frac{P\left(\vec{f}_{test} \mid M_i\right)P(M_i)}{P\left(\vec{f}_{test}\right)} \tag{4.4}$$

Since $P(\vec{f}_{test})$ is identical for all M_i, it is sufficient to minimize the quantity

$$P\left(\vec{f}_{test} \mid M_i\right)P(M_i) \tag{4.5}$$

Where $P(M_i)$ is the prior probability of speaker M_i, being the source of $\vec{f}_{test}$. If $P(M_i)$ are assumed equal, which is customary, the problem becomes finding M_i, that minimizes

$$P\left(\vec{f}_{test} \mid M_i\right) \tag{4.6}$$

This quantity is available; it is simply the GMM speaker model derived in the training procedure. It is important to note that in practice, there is a stream of $\vec{f}_{test}[n]$ for given testing utterance; typically the front-end generates feature vectors at a fixed rate. The frames are assumed to be independent; this implies that both the speaker-models M_i and the likelihoods calculated above do not depend on the order of the feature vector.

4.6 Summary

This chapter has briefly described the speaker identification SID problem, SID methods, SID advantages and limitations, and also described a state-of-the-art Gaussian mixture model (GMM) SID system that will be used in this work for SID evaluations. The next chapter describes the results of the front-ends for speaker-identification using GMM system as a classifier.

CHAPTER 5

SID Experimental Work

This chapter evaluates the features mentioned in chapter 3, mel-cepstrum (MEL), Parameterized feature-set (PFS), and Ensemble Interval Histogram (EIH) for closed-set text-independent speaker-identification (SID) task. The chapter first introduces the SID test conditions in section 5.1, SID recognition system in section 5.1.1, and the database in section 5.1.2. Section 5.2 discusses the performance of SID on clean speech and noisy speech. A series of preliminary experiments performed has four parts. The first set of experiments shows the large impact that the values of front-end parameters have on the classifier performance. The second set of experiments examines the effects of speaker population size on speaker-identification performance. The third set of experiments evaluates robustness techniques for improving performance using telephone speech. Finally, the last set of experiment compares the performance of the features. This last experiment investigates whether

the EIH representation is more resistant to noise than the MEL-cepstrum for ***speaker identification*** *as it was demonstrated for other areas of speech processing.* Therefore the comparison of these front-ends, EIH and MEL, is conducted. Then the performance of each of these front-ends is compared with parameterized feature-set (PFS). PFS has been used before for a speech recognition task [77, 78] where it improved the performance of the system. In this thesis, too, *PFS is investigated if it can improve the performance of speaker identification as it did in speech recognition.* To answer these questions, features are used as an input to the same state-of-the-art speech recognition system and compared in the same database.

5.1 SID Test Conditions

This section discusses the evaluation conditions where the speaker identification experiments are conducted. The database and classifier are discussed. Databases play an integral role for speaker recognition systems in training and testing phases. The few databases that have been developed are discussed in the next section and the motivation of the one chosen in this study is given. These databases are:

A. *Switchboard corpus [41]*

This database consists of conversational telephone speech and was used as a benchmark for the NIST'95 (National Institute of Standards and Technology) Evaluation [85]. It consists of 543 speakers and is used for robust text-independent speaker identification.

B. *King Database [42]*

The king database contains 10 short conversations of approximately 45 s, each recorded during 10 separate sessions from 51 males. It was recorded both locally using a high-quality microphone and after transduction by a carbon-button microphone and transmission over long-distance telephone line. Thus, provides a high-quality (clean) version and a telephone quality version of the

speech. This database was designed for text-independent speaker identification and verification.

C. *Stereo-ATIS Database*

The stereo-ATIS database was recorded over a local telephone line and contains read sentences concerning flight-related issues. This database contains the voices of 13 male speakers. Sentences are on average 4 s long.

D. *TI-105 [98]*

This database is an isolated word database, which has a vocabulary of 105 aircraft command words. It contains 8 speakers, five males and three females.

E. *SRI Digits Database*

SRI –digits database contains the voices of 10 male speakers. The text of the data that was collected from 20 to 23 telephone handsets, consists of spoken digits only.

F. *TIMIT [24], NTIMIT [53] etc.*

The TIMIT/NTIMIT databases have a rich collection of sentences spoken by speakers from the dialect regions of the United States and supplies labels of acoustic-phonetic units for all the sentences. These databases provide us with excellent source for the study of continuous training and evaluation. Text-independent closed-set speaker-identification performs almost error-free in clean speech (i.e. TIMIT database) and poorly in the noisy speech (i.e. NTIMIT database, explained in section 5.1.2) that has been passed through the telephone channel [101]. Text-independent speaker identification can overcome problems that may arise if the speaker is uncooperative, and there is a great interest for speaker-identification over communications channels, which have no linguistic constraints [7]. These databases also have the

advantage of a larger population of speakers (630 people) in non-conversational speech. For these reasons and for real-world applications, speech recognition systems most certainly operate in noisier conditions than is typical of clean/quiet laboratory recording conditions. The telephone-channel NTIMIT database is chosen for system training and evaluation.

GMM, as discussed in the next section, is performed as a recognition system.

5.1.1 SID System

All SID experiments use the Gaussian mixture model (GMM) system described in more detail in Chapter 3. Use of GMM classifier is justified by its being an established, general classifier which acts as a hybrid between standard parametric classifiers, which assumed predetermined distributions, and non-parametric classifiers which typically are computationally expensive, such as *k-nearest neighbors* [26].

5.1.2 Database

The TIMIT speech database [24] has the following properties:

- All utterances recorded in quiet room
- 16kHz sampling rate with 16 bit Pulse Code Modulation samples
- 630 speakers (438 males and 192 females)
- 6225 words and 6300 sentences
- 10 utterances per speaker, each 3 seconds long on average
- Continuous, read (not conversational) speech
- Balanced coverage of speech phonemes
- Full phonetic labelling

The NTIMIT database [53] has the same properties of TIMIT except the first; NTIMIT was created by transmitting all TIMIT utterances over actual telephone channels. Some interesting properties of the transmission process:

- NTIMIT utterances time-aligned with TIMIT utterances
- 50% local / 50% long distance channels

- Various distances for long distance channels
- Effects of telephone handset and microphone included

NTIMIT thus attempts to simulate the effect of TIMIT speakers recording original utterance over telephone channels. The TIMIT database was recorded with a high-quality, essentially distortionless Sennheiser microphone, while the NTMIT database was recorded with a carbon-button microphone. The TIMIT/NTIMIT pair is particularly interesting since all NTIMIT utterances are identical to TIMIT utterances except for the effect of the telephone channel. This fact allows a comparison of any speech-processing algorithm between very "clean" speech and telephone speech (Section 5.2). Time-alignment also means that the full phonetic labelling property of TIMIT thus applies to NTIMIT as well.

The NTIMIT database consists of 10 sentences of approximately 3 *s* each for each speaker. These sentences are denoted as *sa* sentences (two), *si* sentences (three), *sx* sentences (four). The sentences of each speaker are numbered from 1 to 10. Eight sentences of each speaker (approximately 24 *s*) are used for training. The remaining two sentences are used for testing and are adjacent to each other (that is, 1-2; 3-4; 5-6; 7-8; 9-10). The average result is taken as the recognition rate. In a large population, where the comparison of results is conducted, the testing on different two leave-out sentences, and averaging the result, is necessary. But in the first three experiments, 1) evaluating front-end parameters, 2) the effect of population size, and 3) the use of pre-emphasis, only the last two sentences (*sx* sentences), 9 and 10, are used for testing. The reason is that these experiments are conducted in a small population of speakers.

5.2 GMM SID performance

Reynolds [100, 101] has evaluated the performance of GMM models for SID, using MEL-features in several speech databases. SID results with two databases (TIMIT and NTIMIT) are particularly interesting. Reynolds [101] showed the performance of a GMM SID system on both TIMIT and NTIMIT, as a function of the number of speakers in the evaluation; all evaluations were closed-set.

Interesting results appear. First, the SID performance for clean speech is near 100%, up to a population size of 630 speakers. The results suggests that the speakers are not inherently confusable; given features generated from speech produced in the best possible acoustic environment, a GMM classifier can perform almost error-free.

Second, SID performance with telephone speech drops off considerably as the population size increases, even at 100 speakers. The TIMIT speakers were found not to be inherently confusable; clearly the significant performance loss is caused by telephone channel characteristics introduced by NTIMIT, and not by any fundamental inadequacies in the pattern classification process.

5.3 Scoring speaker utterances

As described in Section 4.5.1, the models of each speaker trained can now be used for identification in the testing phase, using a scoring method. Before discussing this scoring method, let us first review the recognition process. The recognition process can be divided mainly into three parts: After pre-processing, which comprises the preparation of the speech signals and the computation of suitable features, there is a training part in which the characteristics of a speaker in the form of reference vectors is determined. Finally, there is a test part, in which new templates, not contained in the reference set, are classified automatically. Figure 5-1 shows these three parts.

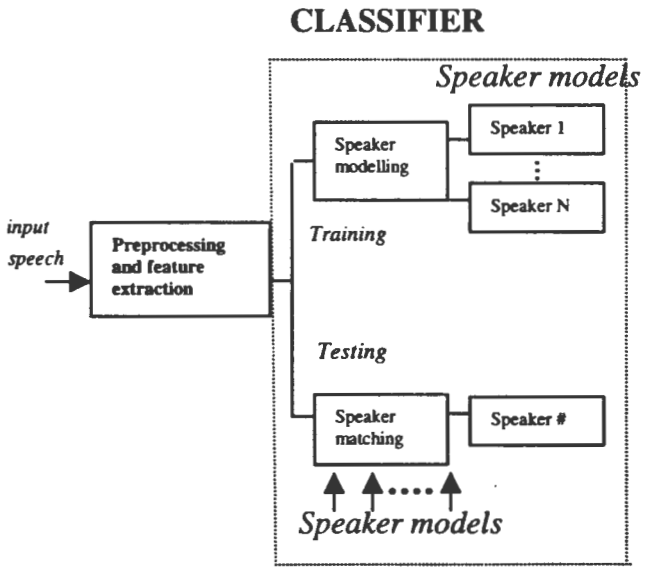

Figure 5-1: Generic speaker recognition system

During the identification, the system is presented with a sequence of observations **X**, produced by one of the S modelled speakers. The identity of the speaker producing **X** is determined by finding the speaker model which maximizes the *posteriori* probability across the speaker set [89].

For each test speech utterance, scores are generated which determine our belief that the models of the speakers of interest produced the utterance. The identity of the speaker is assigned to the model that produced the highest score. The review of the steps involved is described below [40], Figure 5-2.

- Generate a normalized score, for each segment, for each statistic, and for each speaker,
- Robustly combine the normalized score for each statistic, for each speaker,
- Combine the different statistic score.

However, as the population size and the length of test material increases, the computational cost of performing the identification can increase substantially.

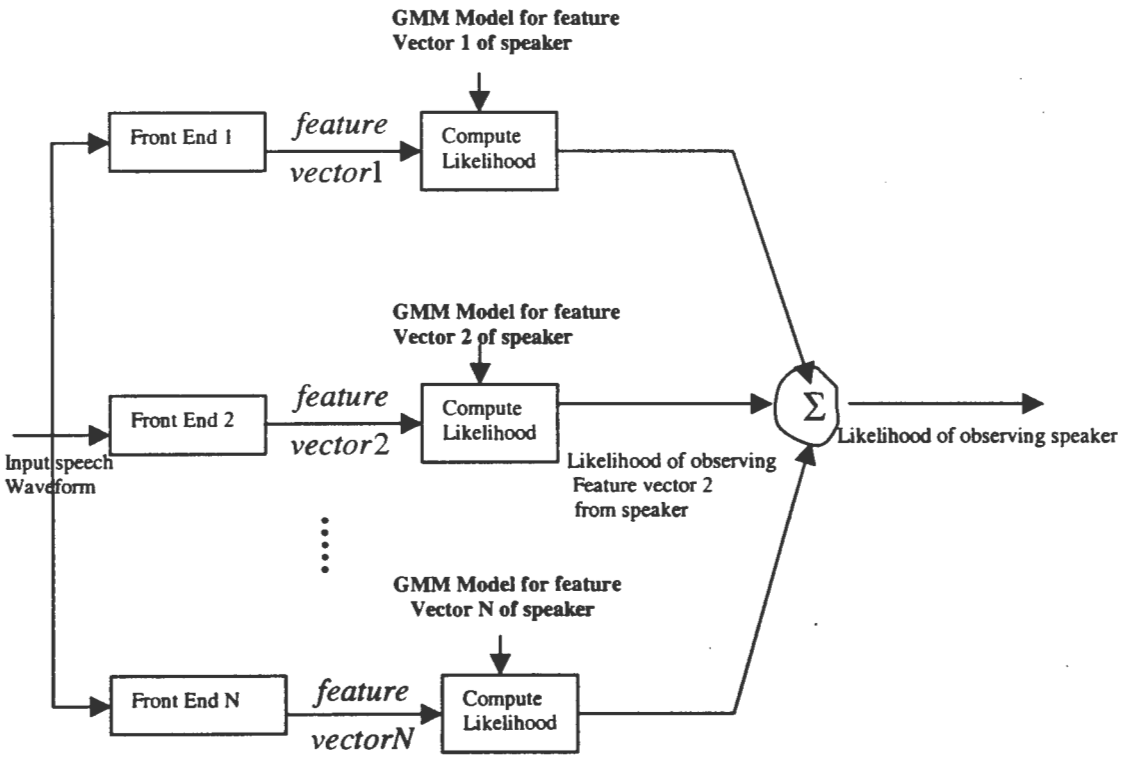

Figure 5-2: Schematic diagram of method for generating a single SID from multiple feature streams.

5.4 Experimental Results

In speaker-independent experiments of EIH, a subset of 168 speakers is used. The male subset contains 112 speakers, while the female subset contains 56 speakers of the telephone-channel NTIMIT database. These subsets were chosen due to the computational expensiveness of the EIH. During training, a model of each speaker is created at approximately 4 minutes in a Pentium II machine. PFS takes approximately 1 minute to create models for 15 speakers. That is, for 630 speakers, EIH takes approx. 630*4 minutes=42hours whereas PFS takes approx. 42 minutes. Another reason for the chosen subsets is to compare with results of other researchers where the same subset of the same NTIMIT database was used. PFS and MEL are compared in the full database of 630 speakers. The preliminary experimentation shows the effect of front-end parameter in system performance, the performance as the population size increases, and the use of pre-emphasis at the front-end. Then the comparison of the front-ends and the analysis of the results conclude the experiments of the study.

A. PARAMETER SETTINGS

The number of parameters affects the computation of the cepstrum, such as

Number of cepstral coefficients
Number of filters in the filterbank
Effective voice bandwidth
Frequency warping constant: -1.0 to +1.0

A series of preliminary experiments are performed to show that the values of front-end parameters have a large impact on the classifier performance. Varying the filter shape, while keeping the other parameters constant, made the error rate vary between 26%-29 % [83] on the preliminary experiments of speaker-identification task. Other researchers [15] reported similar results in their experimental study on the influence of filterbank design and parameters on isolated word recognition. The experiment conducted in this study is the evaluation of the number of cepstral coefficients (NC) using features of the EIH model. The results are in Table 5.1.

POPULATION SIZE OF 38 SPEAKERS	NC	PERFORMANCE (%)
	12	65.8
	20	71.0
	24	73.7
	30	**84.2**
	36	76.3
	40	73.7

Table 5.1: Optimization of the front-end parameters for the EIH-cepstrum feature, when varying number of cepstral coefficients.

The overall performance of the system uniformly improves as the number of cepstral coefficients increases and then decreases after the coefficients of size 30. These results agree with the results for Reynolds's for SID study where the order of cepstral coefficients of 24 outperformed 12 for mel-cepstrum features [103]. Similar results [83] were reported where the increase in the number of cepstral coefficients improved the results of the system, and the number of cepstral coefficients of 17 was used. Using low cepstral coefficients hurt speaker ID performance compared with using high order coefficients. It is likely that for low orders of cepstral coefficients, speaker information dominates in the representation, but, as the number of the cepstral coefficients increases, the channel information begins to dominate. The experiments indicate that the performance of the speaker-identification (SID) system fluctuates significantly with the choice of the front-end parameters. This fluctuation could be due to one of the two reasons: 1) the features are very sensitive to the front-end parameters and 2) the models generated are sensitive to small changes in parameter values.

Multiple levels-crossing detectors with different values are used for frequency and intensity information in the EIH. These multi-level crossing detectors preserve the loudness/intensity information as well as the frequency information of the EIH

representation. Several parameters such as the number of levels and levels values are extremely critical for reliable performance. However, no elegant method exists for determining these values, except by trial and error. Qualitatively speaking, if the level values are near zero, the intensity information of the signal will not be well represented. Some of level-crossing detectors with high values may be useless if the level value captures the crossing points at the high amplitude range of the signal, since timing information at the higher level becomes incorrect in noisy environments. The higher level values are more sensitive to additive noise. Thus, proper determination of the number of levels and level values is very important for reliable performance of the EIH model especially in noisy environments. Table 5.2 gives the performance of level values with which two different methods have been used. The first method, MEAN, level values are uniformly distributed on a log scale over amplitude range of each filtered signal by using a mean L_j and a standard deviation, 0.2 L_j. The mean/level values are therefore, $\{L_j\}_{j=1}^{5}$. The second method, MAX, takes the highest value of the possible maxima of each filtered speech and uniformly distributes the level values. The utilization of multiple level-crossing detectors provides the intensity information of the signal, which may be one of the useful cues for automatic speech recognition. Since auditory models need more computation time, five levels were chosen in the experiments for faster computation as compared to seven [62] or twelve [54].

PERFORMANCE ON 38 SPEAKERS	EIH-MEAN %	EIH-MAX %
Best performance	78.9	84.2
Worst performance	73.7	76.3

Table 5.2: Comparison of two different kinds of level values of EIH (MEAN and MAX).

The other factors that affect the performance of EIH are *bin allocation* and the choice of the *window length.* Linear bin allocation was used since it was shown to outperform a bin allocation according to the bark scale [130] for recognition task [36]. The length of the window is set to 40/CF to capture about 40 periods of the signal at

each channel, provided that the signal is a sinusoid with a frequency equal to the characteristic frequency of the channel, CF. Thus, the window length becomes long for low frequencies, and short for high frequencies, see Figure 5-3. As a result, frequency resolutions are finer, while time resolutions are poorer at low frequencies, and vice-versa at higher frequencies. This property is consistent with psycho-acoustic observations. An EIH observation was computed once every 10 ms. The EIH spectrum contains the spectral envelope information as well as spectral fine-structure information. Only part of this information is used for recognition (features) as it is generally accepted. The current approach is to retain only the spectral envelope and the frame energy information.

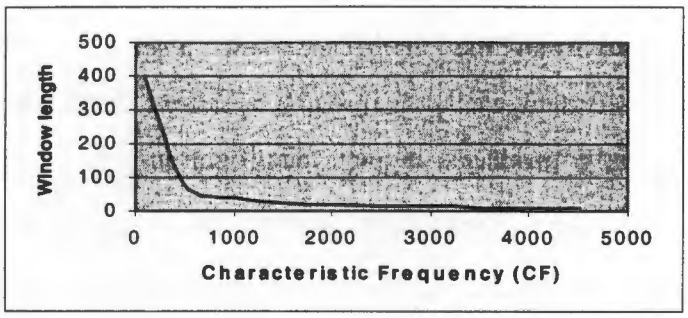

Figure 5-3: EIH time window with frequency

The front-end parameters that resulted in the best performance for the speaker-ID system were chosen. The performance of the system varies significantly with the choice of front-end parameters; this was demonstrated experimentally that a large performance improvement could be obtained by optimizing the front-end parameters.

The number of levels used in all the coming experimentation of the EIH is 5 whereby the level values are distributed on a large amplitude range of each filtered signal by using its highest value of the possible maxima. The cepstral coefficients of size 30 will be used for all front-ends.

The main difference in features is how severely the features are affected by channel and noise effects and how easily these effects can be decoupled from the underlying speech spectral information. All features use *30 cepstral coefficients,* as this was an appropriate envelope fit. EIH needed higher number of coefficients since the pseudo spectra exhibit sharper peaks. The next test that must be performed is to determine the performance of the system as the population size increases.

B. POPULATION SIZE

The size of the speaker population is the other factor that defines the difficulty of the speaker identification task. The increase in the number of speakers in the system increases the probability of an incorrect classification. The similarity of the speakers in the population also affects the performance since a set of speakers with dissimilar voice characteristics cannot easily distinguish speakers. The population with males and females generally produces higher identification performance than a more homogeneous set of speaker (i.e. all males, or all females).

Table 5-3 examines the performance of the speaker identification system as a function of population size of male and female, all male, and all female with respect to population size using the telephone speech. Once more, the EIH features are evaluated.

POPULATION SIZE	EIH	
	No. correct/Total	Score in %
7 (Male and Female)	7/7	**100**
14 (Female)	13/14	**92.9**
23 (Male)	21/23	**91.3**

Table 5.3: Speaker identification results for EIH model, percentage of utterances correctly identified.

In the population size of 7 speaker (male and female), both PFS and EIH score without an error. When the population size increases, the system performance

decreases. The closed set speaker identification (SID) requires that the decision of who spoke the utterance be compared with the N voices of N speakers and chooses the best score that matches a test voice. Therefore the larger N is, the more difficult the identification task is. But in clean database, TIMIT, Reynolds [104] showed that the performance is barely affected by increasing population sizes, which shows that the limiting factor is not a crowding of the feature space.

C. PRE-PROCESSING

Pre-processing is now explained with the purpose of removing the noise. The results of PFS and EIH with and without the application of pre-emphasis filter are shown in Table 5.4.

1. Noise removal

Since speech is segmented into frames, speech/noise/silence frames are estimated by constructing the histogram of the frame energies and those with low-energy are discarded using a speech activity detector [100]. Only frames of energies higher than the decided thresholds are kept for further processing. These speech/silence frames are used in all the experiments so that the identical speech frames are used in training and testing for all features. Sentence-to-sentence noise variation from transmission over difference telephone lines is a source of mismatch between training and testing data, that is why it is very important to eliminate noise frames from the NTIMIT data for good performance.

2. Preemphasis

Due to physiological characteristics of the speech production system (chapter 2), voiced section of the speech signal have an attenuation of approximately 20 dB per decade. To counterbalance this negative slope, a pre-emphasis filter is generally used before spectral analysis. First-order finite impulse response (FIR) filter of the form

$$H(z) = 1 - 0.95z^{-1} \qquad (5.1)$$

is then applied to lessen the susceptibility of the speech signal, *x(n)*. That is the pre-emphasized signal, *y(n)*, is given by

$$y(n) = x(n) - \alpha x(n-1) \qquad (5.2)$$

The constant α=0.95, determines the cut-off frequency of the single-zero filter through which *x(n)* passes. The idea of pre-emphasis is to flatten the spectral peaks of the signal. Many speech recognition systems have eliminated the pre-emphasis stage and compensate for the spectral slope as part of the speech recognition statistical model. Table 5.4 gives the results of PFS and EIH with and without pre-emphasis in the population size of 38 speakers.

38 SPEAKERS	PRE-EMPHASIS USED?	PFS %	EIH %
	NO	97.4	76.3
	YES	94.7	84.2

Table 5.4: PFS and EIH performance with and without pre-emphasis

Pre-emphasis filter increased the performance of EIH, due to the crossings of many levels, which captured more information of each speaker. PFS results decreased a bit with the use of pre-emphasis filter and therefore it is not applied in the experiments.

D. ROBUSTNESS TECHNIQUES

Filtering effect with band limits is the major spectral degradation found in speech collected from the telephone network that imposes some spectral shaping on the speech spectrum [94]. This degradation which is left uncompensated can produce severe reductions in identification performance due to mismatch between training and recognition data. Since NTIMIT is a telephoned-channel database, robustness

techniques are used to compensate for spectral variability introduced by the telephone channel and handset. Some spectral variability compensation techniques such as long-term mean removal, difference coefficients, and frequency warping, are discussed.

1. Frequency Warping

Frequency warping is applied to the magnitude DFT spectrum to avoid any differences in channel bandwidth and using any spurious out-of-band spectral components. This is done by

$$f' = \frac{f - f_{\min}}{f_{\max} - f_{\min}} f_N \qquad 5.3$$

where f_N is the original Nyquist frequency. The linear warping both eliminates spectral components outside the specified frequency range [f_{min}, f_{max}] and expands the spectrum bandwidth for subsequent processing.

3. Mean normalization

The method of mean normalization have been used in many speaker recognition systems [5, 27, 63, 101]. This method consists of removing the bias component by subtracting off the global average vector from each feature vector. From each channel from which speech was collected, the global mean is subtracted off each vector before training a speaker model or scoring for recognition. All feature vectors then have the same global mean and different channel bias does not affect speaker discrimination. The above assumes a time-invariant channel filter.

If the channel filter is time-varying, an adaptive bias removal method, such as RASTA processing [50], can be used. This compensation not only removes the channel filter bias, but also removes the global mean of the speech feature vectors.

Quadratic Trend Removal [MIS98] assumes that the channel filter is relatively smooth across frequencies compared with the speech spectrum

and can be modeled by a quadratic polynomial in the log spectrum domain. Quadratic Trend Removal is best suited for modeling the channel filter over the passband.

Using mean normalization for clean speech improves identification performance by minimizing intersession variability. When used on telephone speech, removal of global average minimizes both intersection variability and removes the spectral shaping imposed by different telephone channels. Since the database is passed through the telephone channel, mean removal is used in this study. To minimize channel filter effects, another method that can be used is cepstrum different coefficients [109], discussed next.

4. *Time Difference coefficients*

Difference coefficients have been shown to contain speaker specific information and to be fairly uncorrelated with the static cepstral frame vectors; however, when used by themselves, they do not perform as well as the static feature vectors [117]. To combine the two feature sets, the difference coefficients are appended to the cepstral feature vectors. The new feature vectors not only contain channel invariant features but also spectral transitional information along with the instantaneous cepstral coefficients.

The most straightforward representation is the first difference. However, the first difference is susceptible to noise since it amplifies the high frequency components of the temporal trajectories of the cepstral coefficients. Therefore, the time derivative is approximated by a polynomial approximation [50], which has the effect of bandpass filtering the temporal trajectories. These filtered coefficients are known as the delta-cepstral coefficients.

Differential cepstral parameters are not used in this study since there is no variability between recording microphones in NTIMIT database and

little variability between the telephone lines. Kim [62] showed that delta features have improved the performance for noisy data in MEL and performance improvements were smaller in EIH in the word recognition task. Sandhu [111] reported the same results for the phone classification task. One reason why dynamic features have a deleterious effect on the EIH representation could be the frequency-dependent time window used in EIH analysis, Figure 5.3. An appropriate method for calculating delta coefficients must be devised for EIH. Augmenting dynamics such as delta and delta-delta features to static features improved recognition task. Computing delta features is equivalent to a FIR filtering.

Cepstral mean removal was found to be the best channel compensation technique for all features [103]. Therefore, it is the chosen channel compensation technique in this study. The next section compares the results of features.

E. COMPARATIVE RESULTS ON SMALL POPULATION

POPULATION SIZE	EIH	PFS
7 (Male and Female)	100	100
14 (Female)	92.9	92.9
14 (Male)	92.9	85.7
24 (Male)	87.5	95.8
38 (Male and Female)	79.8	94.7
76 (Male and Female)	68.4	82.8

Table 5.5: Comparative results of PFS and EIH features on a small database

As can be seen from Table 5.5, for 7 speakers, both PFS and EIH perform without an error. This result may be due to the selected speakers with different speaking styles. EIH and PFS performs the same for 14 female speakers, but EIH performs slightly better than PFS for 14 male speakers. This again may be caused by the variability on

the selected speakers, i.e. set of speakers with dissimilar voice characteristics. As the population increases, PFS outperforms EIH.

F. COMPARATIVE RESULTS ON LARGE POPULATION

A comparative study using the same 168-speaker subset of the NTIMIT database, comprising of 56 females and 112 males, with that of the other researchers is now conducted. The comparison is based on no-cross sex, Figure 5-4.

Jankowski [55] conducted standard mel-cepstra for speaker identification task using GMM as a classifier and reported the accuracy of 77.2 % for male speakers and 73.6 % for female speakers. After combining mel-cepstra with the format modulation features added to formant information and formant difference information, the performance improved to 81.2 % for males and 81.8 % for females.

TESTING-SENTENCES	FEMALE (56)		MALE (112)	
	PFS	EIH	PFS	EIH
1,2	71.4	67.9	70.5	60.7
3,4	71.4	69.6	77.7	64.3
5,6	80.4	71.4	74.1	67.8
7,8	80.4	73.2	83.0	69.6
9,10	83.0	75.0	83.9	70.0
AVERAGE	**77.3**	**71.4**	**77.84**	**66.5**

Table 5.6: The average performance of PFS and EIH in a large population

Comparing the results of Jankowski using mel-cepstra and the ones in this study using parameterized features (PFS), where the same classifier, GMM, is used in the same equal subsets of NTIMIT database, the results of PFS in this study are 83.0% for the males and 83.9% for the females. *This indicates that parameterized feature-set has better performance.* Even when the combined features (MEL and Formant modulation features) of Jankowski improved the results to 81.1% for males and

81.8% for females, they are still below the results of this study using parameterized features.

Another interesting comparison of results is with the ones for Plumpe *et al.* [90] where the same NTIMIT database was used and GMM as a classifier. The 168-speaker subset comprising of 112 males and 56 females was conducted once more. Their results for MEL-cepstrum are 56.7% for males and 66.3% for females. Combining MEL-cepstrum features and modeled Glottal Flow Derivative (GFD) source features improved the results to 59.8% for males and 69.0% for females. Obviously, these results are lower than MEL results in this study. They are also lower than the EIH results obtained in this study. EIH results for females (75%) are also higher than that the MEL results for females for Jankowski (73.6%). *This indicates that EIH outperforms MEL for other researchers.* In this study, EIH was outperformed by the parameterized feature-sets, which is an improved version of traditional mel-cepstrum. No wonder it was outperformed, but it is not bad either. The parameterized feature-set (PFS) uses liftering process instead of the mel-scale filterbank in the traditional mel-scale cepstral coefficients (section 3.2.2).

G. *COMPARATIVE RESULTS ON FULL NTIMIT DATABASE*

This section contains the most interesting results for comparison. This is so since this study was mainly dedicated on improving Reynolds' results [104]. The large population of all 630 speakers, 438 males and 192 females, was conducted on the same database, NTIMIT. What makes this study more exciting is because in the same population, Reynolds gets almost 100% performance on a clean speech. EIH was not conducted in this large population because of computation time. In the following experiments, 3 different kinds of test are conducted, no cross-sex, cross-sex, and all database (combined sexes). These terms are discussed next.

1. **No cross-sex**

   This set of experiments has separate sexes, male and female, whereby training is conducted on each sex and tested on that particular sex. The identification/recognition of each speaker is on that particular sex only.

This set-up of experiment is in Figure 5.4. Table 5.7 contains the results of this experimentation.

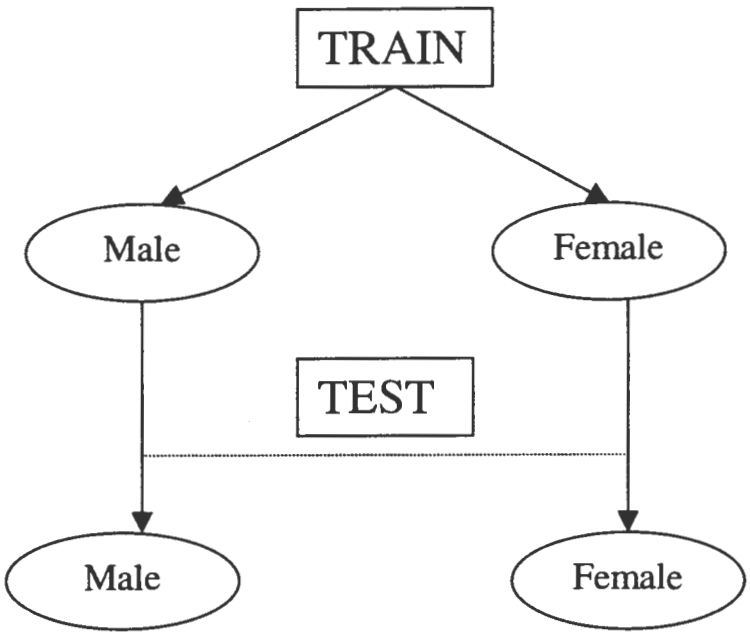

Figure 5-4: A representation of no cross-sex.

SEX	RESULTS
438 (Male)	74.4
192 (Female)	63.8

Table 5.7: Results without cross-sex.

2. Cross-sex

The cross-sex experiment is conducted on the combined set of male and female speakers. Training is done on a full database of 630 speakers and the identification of each sex, males and females, is done on the combined set, Figure 5.5. The identification in cross-sex is expected to be lower than the one without cross-sex, due to the population size of 630 as compared to 438 of males and 192 of females. In this mode of experiment, a male speaker can be identified instead of a female speaker. Table 5.8 contains the results of the cross-sex experiment, which are compared with the previous results of

Reynold's [104] that were conducted on cross-sex experiments on the same database.

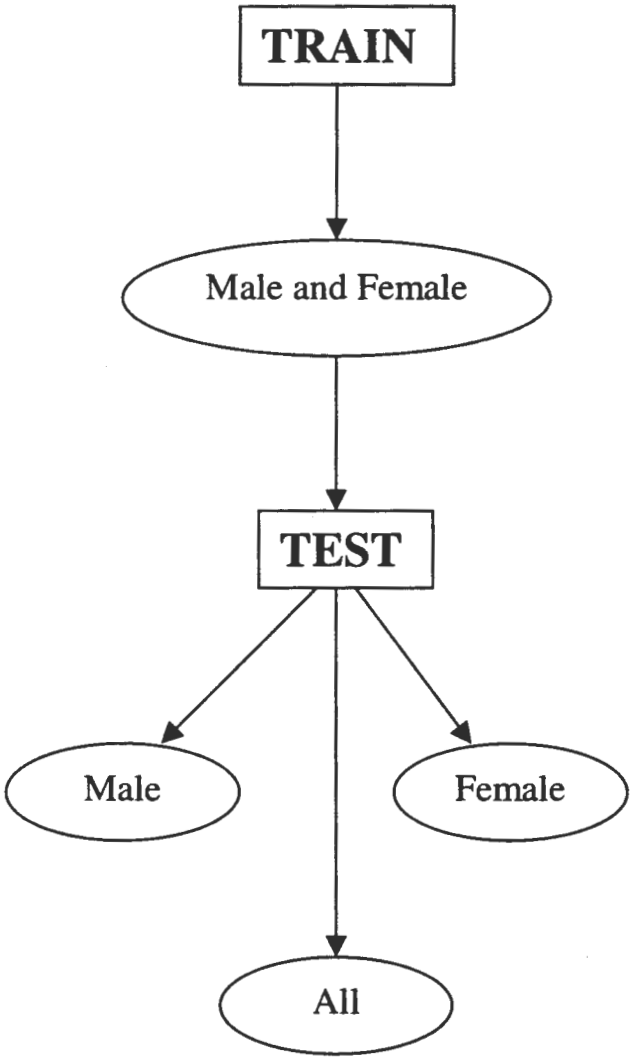

Figure 5-5: A representation of cross-sex.

SEX	REYNOLDS (MEL)	THIS STUDY (PFS)
438 (Male)	62.5	73.5
192 (Female)	56.5	62.7

Table 5.8: Comparison of previous and current results with cross-sex on a full database, respectively.

3. All-sex

On the full database the performance of male and female speakers are not separate, but combined. The first experiment is conduced on all 630 speakers, where different adjacent sentences as described in section 5.1.2 are used for testing. The second set of experiment is conducted on the increasing population of 10, 100, 200, 300, 400, 500, 600, 630 speakers. Figure 5.6 contains the results of this experiment.

TESTING-SENTENCES	630 SPEAKERS
1,2	64.8
3,4	65.3
5,6	69.0
7,8	72.7
9,10	70.8
AVERAGE	**68.5**

Table 5.9: The performance of PFS on a full database

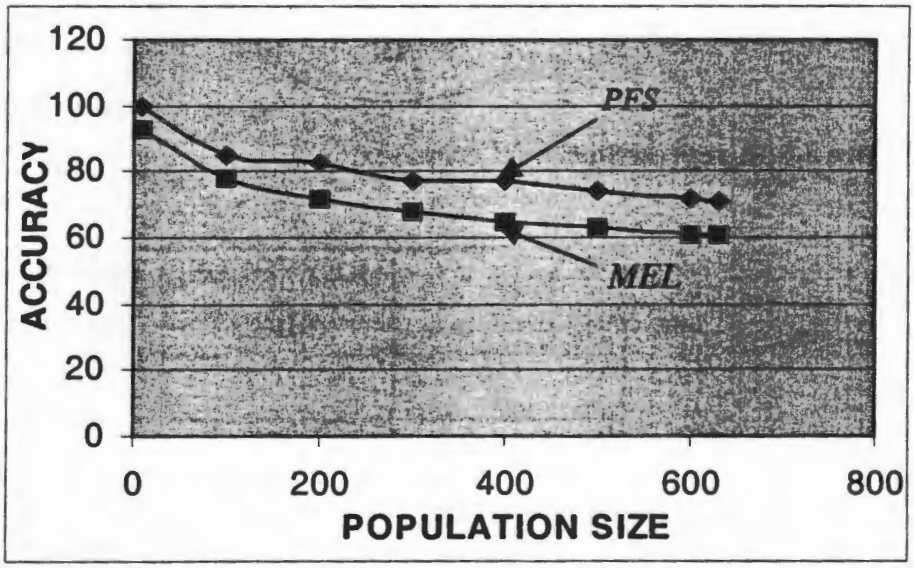

Figure 5-6: Comparative results of MEL and PFS as a function of population size.

COMBINED-SEX	REYNOLDS (MEL)	THIS STUDY (PFS)
630 (Male and Female)	**60.7**	**72.7**

Table 5.10: Comparison of previous and current results of combined sex on a full database.

As can be seen, all the results of male and females speakers are improved by a larger margin. The relative difference, in percentage, of MEL and PFS, is given by

$$\frac{\frac{PFS - MEL}{PFS + MEL}}{2} * 100 \qquad (5.4)$$

Using equation 5.4, on the cross-sex (Table 5.8) (i.e. results of male/female speakers on database of mixed-sex), the results improved by 16.2% for males and 10.4% for females. Without the cross-sex, Table 5.7 (i.e. results of male on male only), the female and male results improved even more than those of cross-sex. This is the expected result as was mentioned in section 5.4B, that is, a population of separate sexes (males and females) produces higher identification performance than a homogenous set of speakers. The performance is also influenced by the fact that the population size of the cross-sex is much larger in cross-sex (630 speakers) as compared to separate sexes, which makes the identification more difficult.

On the full database, 630 males and females, Reynolds got 60.7% and in this study 72.7% was obtained. Undoubtedly, using equation 5.4 again, the speaker identification performance has been improved by 18%. Therefore the main goal of this study has been achieved.

Comparing the results of MEL and PFS as the function of population size in Figure 5.6, it can be seen that both front-ends are affected by the increasing population size. The telephone-channel database, NTIMIT, decreases the performance of the accuracy as the population size increases, but this is not the case with the clean

database (TIMIT) [104]. The superiority of PFS is again proven by Figure 5.6, where it outperforms MEL.

These studies have a lot in common; both use GMM as a classifier and NTIMIT as a database. Parameterized feature-set (PFS) increased the results of speaker identification (using MEL features) by a larger margin of 18%. Cepstral mean removal for channel equalization, was also used in this study, but was not used in Reynolds' paper. Reynolds obtained maximum likelihood (ML) estimates of the model parameters using the expectation-maximization (EM) algorithm, which is known to improve results. EM algorithm was not used in this study because of computation time, but it does improve results by the order of 1-3%.

The next section analyses the results of front-ends that have been used for comparison in this study.

5.5 Analysis of Results

Since the same recognizer, GMM, and the same database, NTIMIT, have been used in all experiments, the difference in performance is due to the difference between the feature vectors. In the experiment (see Table 5.5 and Table 5.6), it can been seen that PFS outperforms EIH for the speaker-identification problem. From the outset, we assume that EIH is an appropriate representation of an acoustic signal. This is so since the EIH is constructed from a detailed simulation of the human auditory-nerve firing patterns, using rules derived from general properties of observed auditory nerve activity in the cat. The deficiency, therefore, may be the result of inappropriate use of the EIH in the context of speech processing. As a reminder, EIH is performed for the first time for the speaker-identification task and also performed for the first time to a GMM classifier.

Other researchers have obtained large reduction in error rates with auditory models in noisy speech [34]. These previous studies show significantly larger reduction in error-rates with auditory models than was presented in this study. Ghitza showed the significant error rate reduction with an isolated word recognition task

using the EIH model [34]. Stern also reported significant improvement using both the mean-rate and synchrony branches of Seneff's auditory model with a continuous speech database and a recognizer [120]. There could be several reasons for the differences between their study and the one presented in this study.

First, this work used a DFT-based control front-end, whereas both Ghitza and Stern used an LPC based cepstral front-end as the control. LPC front-ends do not, however, typically perform well in noisy environments. These results suggest that the use of a poorer performing control front-end might have caused other studies to find more reduction error rate with auditory models than is have found in this study. Ghitza's experiment is a speaker-dependent, isolated word recognizer, unlike the one in this study, which is speaker-independent. Ghitza and Stern also added artificially generated noise, whereas a recorded speech babble was used in this study.

EIH was showed to outperform MEL in real world noisy environments such as *factory noise* and *military operation room noise* [60], and *white Gaussian noise* [61] [62]. The experiments were done on the speech data of 50 Korean words of 16 speakers for word recognition. The related results were provided by Ghitza [34], Stern *et al.* [120], and Jankowski *et al.* [54]. Jankowski used a *speech babble noise* as a background noise for automatic word recognition. Jankowski [54] also compared MEL cepstral front-end and LPC front-end. Performance for both front-ends was better with the *speech babble* than with the *white noise*. Ghitza [33] demonstrated EIH spectrum to be more robust to *additive white noise* than the conventional Fourier Transform (FFT)-based front-end to a particular Dynamic Time Warping (DTW) recognizer. The difference in recognition rate between the conventional front-ends, MEL and PFS, and auditory model, EIH, is dependent on the *type of noise*. The difference is maximum when white Gaussian noise is used, and decreases when real-world noises are used. The reason is not clear yet, and remains as a future work.

Sandhu [111] compared MEL and EIH for phone classification through a telephone channel using 3 male speakers using HMM as a classifier. The results are interesting for comparison since telephone speech is used in this study as well. For static features, EIH outperformed MEL with the performance of 44.6% against 30.2%, respectively. After applying dynamic features to both front-ends, the results of

EIH and MEL were both increased, with EIH still better. But in this study, differential cepstral parameters are not used since there is no variability between recording microphones in NTIMIT database. As with other previous research of speech processing, EIH outperforms MEL for the speech passed through the telephone channel for speaker identification task.

Parameterized feature-set (PFS) was shown to improve text-independent closed-set speaker identification on a large population. Similar results on improvement of the performance using these features were conducted on speech processing task [77, 78]. On the full NTIMIT database, PFS improved the previous results from 60.7% to 72.7%, outperforming mel-cepstra with an increase of 18%. When the subset of NTIMIT database was experimented on, once again, PFS outperformed EIH. In this study we have shown that the feature-set can improve recognition performance and system robustness.

Since other researchers have reported a larger error-rate while comparing EIH and MEL, and PFS outperforming EIH for SID task, certainly, there may exist a more efficient density representation of EIH features than that of a GMM for approximating the feature distribution for SID. In this study, EIH was evaluated for the first time on Gaussian mixture model (GMM) and also for SID task. The aim, however, was to use the same established classifier which is general enough for features to show that the EIH features convey speaker identity information.

5.6 Summary

This study has presented a review of some of the techniques used in robust speaker recognition with an emphasis on feature extraction. The existing coefficients are not very robust to a wide variety of environmental conditions, such as telephone noise. MEL and EIH features were evaluated and compared, where EIH outperformed MEL on the telephone-channel degraded speech database. The superiority of the EIH used in this study was shown when compared with the other researcher's results for MEL where the same classifier, and the same database subset were used. The larger improvement of SID performance was obtained using a parameterized feature-set which outperformed both EIH and MEL. This indicates that the performance of the system can be influenced by the well-tuned features.

We must not cease from exploration. And the end of our exploring will be to arrive where we began and to know the place for the first time.
T. S. Elliot

CHAPTER 6

Conclusion

This section summarizes the whole work of this study in section 6.1, make conclusions based on the results obtained in section 6.2 and details the future work in this field area of study in section 6.3.

6.1 Summary

Since the identity of a speaker by his/her voice is the most attractive field in speech processing, the first goal of this study was to improve the performance of the SID system on a telephone-channel database. Firstly, the improvement was obtained by using parameterized feature-set (PFS) that differs from traditional mel-cepstrum in

that it uses a liftering process instead of the usual mel-scale filterbank. Secondly, the use of mean removal as channel equalization also improved the performance of the system.

Since there is a great interest for speaker identification over communication channels for real-world applications, most speech recognition systems must certainly operate in noisier conditions than in a quiet environment. For this reason the telephone-channel NTIMIT database was chosen for system training and evaluation. The NTIMIT database has a rich collection of sentences spoken by speakers from the dialect regions of the United States and supplies labels of acoustic-phonetic units for all the sentences.

The use of GMM classifier is justified by its being an established, general classifier which acts as a hybrid between the standard parametric classifier, which assumed predetermined distributions, and non-parametric classifiers, which typically are computationally expensive and also capable of better modelling arbitrary feature distribution. The GMM models some underlying set of acoustic classes which reflect some general speaker-dependent vocal tract configurations that are useful for characterizing speaker identity. One of the powerful attributes of the GMM is its ability to form smooth approximations to arbitrary-shaped densities.

The second goal was to compare the performance of the speech representations: the traditional Fourier-based mel cepstrum (MEL), the auditory-based Ensemble Interval Histogram (EIH), and DFT-based Parameterized feature-set (PFS). The EIH and MEL front-ends were chosen since they were shown to outperform their counterparts, and PFS was chosen since it was shown to improve the recognition rate of the system for the speech processing task [77]. The comparison of these front-ends was interesting since MEL was performed almost 100% in clean speech [104], and the auditory models including EIH, were shown to outperform MEL in noisy conditions from the previous studies [54], of which all the experiments were not on speaker identification (SID). An auditory model, EIH in particular, was never evaluated on a large population, speaker-independent continuous speech recognition. In this study, it was also performed for the first on the state-of-the-art speech recognition system, Gaussian mixture model (GMM).

Front-ends, MEL, EIH, and PFS and the recognition system, GMM, have been performed to work on closed-set text-independent speaker-identification problem.

6.2 Conclusions

The study of speech representations, MEL, EIH, and PFS, using state-of-the-art recognition system, GMM, for text-independent speaker-identification on a telephone-degraded NTIMIT database, yielded the following comparative results:

- EIH outperformed MEL in a subset of NTIMIT database. This is in agreement with the previous studies [34] [54] [111] that have been performed for other areas of speech other than speaker identification, but they were based on a small population.

- On smaller population of 14 speakers, EIH performed slightly better than PFS, the reason may be the group of chosen speakers that happen to have close speaking styles. Whereas on the larger population, PFS outperformed EIH. Since it was indicated in section (5.5), the type of noise does influence the performance of the features. EIH was performed on text-independent closed-set speaker-identification (SID) for the first time. It is for the first time, once more, that EIH was conducted on GMM. Perhaps there may exist a more efficient density representation of EIH features than that of GMM for approximating the feature distribution for SID. The aim, however, was to use the same established classifier which is general enough for features to show that the EIH features convey speaker identity information. But the classifier performance is not likely to adversely affect the identification score since the system produced better performance in a smaller size of population.

- On the full database of NTIMIT, PFS outperformed MEL performed by Reynolds [104].

- Comparing the results obtained in this study with those of the other researchers, we could see that the developed system of this study produced better results. This

was the main purpose of this study, to improve the previous speaker-identification results. And it was done in fashion.

- The study continued with the auditory-based models since they demonstrated to be more robust to noise. It was for the first time, that the auditory model is performed for text-independent speaker identification.

Previous studies suggested that EIH performs worse than MEL in clean speech, but more robust in adverse conditions. These studies were conducted on a limited task, i.e. speaker dependent isolated words in small population size. This study extends the research to the task of speaker-independent, continuous speech in a large population.

The success of utilizing EIH depends on the amount of knowledge we have about the auditory system. This knowledge is achieved by combining data that has been collected in both psychological and physiological studies of the auditory system. Not all is known, at present, about the functional operations of the auditory nervous system.

6.3 Future work

Closed-set SID performance using the GMM SID system with the clean speech database is very high, i.e. almost 100% speaker identification accuracy up to a population size of 630 speakers. But the performance with the noisy speech database is poorer; this motivates the search for new features that are robust over the telephone network. A search for additional features is warranted by the fact that it is strictly the telephone channel that is causing the performance loss, since accuracy is near perfect with clean speech. More effort is needed in finding features for achieving very high recognition performance (especially under severe channel conditions and very low signal to noise ratios). There is a need to develop features that will be robust over the telephone network.

Although many recent advances and successes in speaker recognition have been achieved, there are still many problems for which good solutions remain to be found. Most of the problems arise from variability, including speaker-generated variability and variability in channel and recording conditions. These advances have not necessarily come about as an outgrowth of new or better understanding of speaker characteristics or how to extract them from the speech signal, but rather through improvements in techniques for making speaker –sensitive feature measurements and models. It is very important to investigate feature parameters that are stable over time, insensitive to the variation of the speaking manner, including the speaking rate and level, and robust against variations in voice quality due to causes such as voice disguise or colds.

The human auditory system is sufficiently complex that years of intensive research in the areas of auditory physiology and perception have left even some questions of auditory function unanswered. The methods used for converting auditory front-ends outputs into feature vectors were largely dictated by the structure of the speech recognition system. For a fair test of the effectiveness of the auditory models, more work is necessary to obtain improved ways of incorporating features from auditory models into speech recognizers. As auditory function and speech perception are better understood, new parameters important for speech recognition should be uncovered. It is expected that improvements will continue with this approach, but we hope that current and future understanding of speaker characteristics in the speech signal can be applied to provide even more effective speaker recognition system.

Future work in the area of front-end evaluation might focus on continuous speech since current speech recognition systems are primary targeted for continuous speech tasks.

References

[1] ACERO, A., AND STERN, R.M. (1990). Environmental robustness in automatic speech recognition. In *Proc. Int. Conf. Acoust., Speech, Signal Proc.*, pp. 849-852.

[2] AINSWORTH, W. A. (1988). *Speech recognition by machine*. Peters Peregrinus Ltd, London, UK.

[3] ARPA. (1990). *The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus, Training and Test Data*. NIST Speech Disc CD1-1.1.

[4] ASSEH, K. H., AND MAMMONE, R. J. (1994). New LP-derived features for speaker identification. *IEEE Trans. on Speech and Audio Processing*, pp. 630-638.

[5] ATAL, B. S. (1974). Effectiveness of linear prediction characteristics of the speech waver for automatic speaker identification and verification. *J. Acoust. Soc. Amer.*, 55: pp. 1304-1312.

[6] ATAL, B. (1976). Automatic recognition of speakers from their voices. *Proc. IEEE*, 64: pp. 460-275.

[7] BEEK, B., NEUBERG, E. P., AND HODGE, D. C. (1977). An assessment of the technology of automatic speech recognition for military applications. *IEEE Trans. Acoustic, Speech, Signal Processing*, 25: pp. 310-322.

[8] BENNANI, Y., AND GALLINARI, P. (1991). On the use of TDNN-extracted features information in talker identification. In *Proc. Int. Conf. Acoust., Speech, Signal Proc.*

[9] BRISTOW, G. (1986). *Electronic speech recognition: Techniques, Technology & Applications*. McGraw-Hill Book Company, USA.

[10] CHEN S. –H., AND WANG, Y. –R.(1995). Tone recognition of continuous Mandarin speech based on neural networks. *IEEE Trans. Acoustic, Speech, Signal Processing*, 3: pp. 146-150.

[11] CHOU, W., AND CHEN, H. H. (1995). Speech recognition for image animation and coding. In *Proc. Int. Conf. Acoust., Speech, Signal Proc.*

[12] COHEN, J. R. (1989). Application of an auditory model to speech recognition. *J. Acout. Soc. Amer.*, 85: pp. 2623-2629.

[13] COLE, R. (1995). The Challenge of Spoken Language Systems: Research Directions for the Nineties. . *IEEE Trans. Acoust., Speech, Signal Processing*, 3.2: pp. 1-29.

[14] COLTON, L. D. (1997). *Confidence and rejection in automatic speech recognition,* Ph.D. thesis, Oregon Graduate Institute of Science and Technology, Oregon.

[15] DAUTRICH, B., RABINER, L. R., AND MARTIN, T. B. (1983). On the effects of varying filter bank parameters on isolated word recognition. *IEEE Trans. Acoust., Speech, Signal Processing*, 31: pp. 793-806.

[16] DAVIS, S. B., AND MERMELSTEIN, P. (1980). A comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Trans. Acoust., Speech, Signal Processing,* 128.4: pp. 357-366.

[17] DEMPSTER, A. P., LAHD, N.M, AND RUBIN, D.B. (1977). Maximum-likelihood from incomplete data via the EM algorithm. *J. Roy. Stat. Soc.*, 39: pp. 1-38.

[18] DODDINGTON, G. (1985). Speaker recognition-Identifying people by their voices. *Proc. IEEE*, 73, November.

[19] DRAKE, A. W. (1967). *Fundamentals of Applied Theory.* McGraw-Hill.

[20] DUDA, R. O. (1978). *Pattern classification and scene analysis.* John Wiley & Sons, New York.

[21] ELLIOT, J. A., AND CONKIE, A. D. (1994). Speech technology helps the police with enquiries. *Computing and Control Engineering Journal*, 5.4: pp.165-171.

[22] EPHRAIM, Y. (1992). Gain-adapted hidden Markov models for recognition of clean and noisy speech. . *IEEE Trans. Acoust., Speech, Signal Processing*, 40.6: pp. 1306-1316.

[23] ERELL, A., AND CHEN, W. -Y. (1993). Generalized minimal distortion segmentation for ANN based speech recognition. *IEEE Trans. Acoustic, Speech, Signal Processing,* pp. 68-76.

[24] FIHER, W. M. , DODDINGTON, G. R., AND KATHLEEN, M. (1986). The DARPA Speech Recognition research database: Specifications and status. *In Proc. DARPA Speech Recognition Workshop*, Palo Alto, pp. 93-99.

[25] FLOCH, J. L., MONTACIE, C. AND CARATY, M. J. (1994). Investigation on speaker characterization from Orphee system technics. In *Proc. Int. Conf. Acoust., Speech, Signal Proc.,* I: pp.149-152.

[26] FUKUNAGA, K. (1972). *Introduction to statistical pattern recognition.* New York: Academic.

[27] FURUI, S. (1981). Cepstral analysis technique for automatic speaker

verification. *IEEE Trans. Acoust., Speech, Signal Processing,* 29: pp. 254-272.

[28] FURUI, S. (1986). Speaker-independent isolated word recognition using dynamic features of speech spectrum. *IEEE Trans. on Speech and Audio processing*, 35 : pp. 52-59.

[29] FURUI, S. (1988). A VQ-based preprocessor using cepstral dynamic features for speaker-independent large vocabulary word recognition. *IEEE Trans. Acoust., Speech, Signal Processing,* pp. 980-987.

[30] FURUI, S. (1989). *Digital speech processing, synthesis, and recognition.* Marcel Dekker, New York.

[31] FURUI, S. (1990). On the use of hierarchical spectral dynamics in speech recognition. In *Proc. Int. Conf. Acoust., Speech, Signal Proc.* Albuquerque.

[32] FURUI, S. (1992). Speaker-independent and speaker-adaptive recognition techniques. In *Advances in Speech signal processing,* S. Furui and M. M. Sondhi, Eds, pp. 597-622, Marcel Dekker.

[33] GHITZA, O. (1986). Auditory nerve representation as a front-end for speech recognition in a noisy environment. *Computer speech and Language*, pp. 109-130.

[34] GHITZA, O. (1992). Auditory nerve representation as a basis for speech processing. In *Advances in Speech Signal Processing*, S. Furui and M.M. Sondhi, Eds. Marcel Dekker, Inc. pp. 453-485.

[35] GHITZA, O. (1993). Adequacy of auditory models to predict human internal representation of speech sounds. *J. Acoust. Soc. Am.*, pp. 2160-2171.

[36] GHITZA, O. (1994). Auditory models and human performance in tasks related to speech coding and speech recognition. *IEEE Trans. Acoust., Speech, Signal Processing,* 2: pp. 115-131.

[37] GISH, H., KARNOFSKY, K., KRASNER, M. ROUCOS, S. SCHWARTZ, R., AND WOLF, J. (1985). Investigation of text-independent speaker identification over telephone channels. In *Proc. Int. Conf. Acoust., Speech, Signal Proc.*, pp. 379-382.

[38] GISH, H., KRASNER, M., RUSSEL, W., AND WOLF, J. (1986). Methods and experiments for text-independent speaker recognition over telephone channels. In *Proc. Int. Conf. Acoust., Speech, Signal Proc.*, 1: pp. 865-868.

[39] GISH, H., CHOW, Y. L., AND ROHLICEK, J. R. (1990). Probabilistic vector mapping of noisy speech parameters for HMM word spotting. In *Proc. Int. Conf. Acoust., Speech, Signal Proc.*, pp. 117-120.

[40] GISH, H., SCHMIDT, M, and MIELKE, A. (1994). A robust, segmental method for text independent speaker identification. In *Proc. Int. Conf. Acoust., Speech,*

Signal Proc., I : pp. 145-152.

[41] GODFREY, J., HOLIMAN, E., AND McDANIEL, J. (1992). Switchboard: Telephone speech corpus for research and development. In *Proc. Int. Conf. Acoust., Speech, Signal Processing,* pp. 517-520.

[42] GODFREY, J., GRAFF, D. AND MARTIN, A. (1994). Public databases for speaker recognition and verification, *in Proc. ECSA Workshop Automat. Speaker Recognition, Identification, Verification*, pp. 39–42.

[43] HANSEN, B., AND WAKITA, H. (1987). Spectral slope distance measures with linear prediction analysis for word recognition in noise. *IEEE Trans. on Speech and Audio processing*, pp. 968-973.

[44] HANSEN, J. H. L., AND BOU-GHAZALE, S.E. (1994). Morphological constrained feature enhancement with adaptive cepstral compensation for speech recognition in noise and Lombard effect. *IEEE Trans. on Speech and Audio processing*, pp. 598-614.

[45] HANSEN, J. H. L., AND ARSLAN, L. M. (1995). Robust feature-estimation and objective quality assessment for noisy speech recognition using the credit card corpus. *IEEE Trans. on Speech and Audio processing*, pp. 415-421.

[46] HANSEN, J. H. L., AND WOMACK, B. D. (1996). Feature analysis and neural network-based classing of speech under stress. *IEEE Trans. on Speech and Audio processing*, pp. 307-313.

[47] HARRIS, D. M., AND DALLOS, P. (1979). Forward masking of auditory nerve fiber responses. *Journal of Neurophysiology,* 42: pp. 1083-1107.

[48] HERMANSKY, H., AND HANSON, H. J. (1985). Perceptually based linear prediction analysis of speech. In *Proc. Int. Conf. Acoust., Speech, Signal Proc.*, pp. 509-512.

[49] HERMANSKY, H. (1990). Compensation for the effects of the communication channel in auditory-like analysis of speech. *In Proc. 2nd European Conf. Speech, Commun., Technol.*, Italy, pp. 1367-1370.

[50] HERMANSKY, H. (1992). RASTA-PLP speech analysis technique. *Int. Conf. Acoust., Speech, Signal Processing,* I: pp. 121-124.

[51] ITAKURA, F., AND SAITO, S. (1970). A statistical method for estimation of speech spectral density and formant frequencies. *Electronics and Communications in Japan.* 53A: pp. 36-43.

[52] ITAKURA, F. (1975). Minimum prediction residual principle applied to speech recognition. *IEEE Trans. Acoust., Speech, Signal Proc*essing, 23: pp. 67-72.

[53] JANKOWSKI, C. R. KALYANSWAMY, A. BASSON, S., AND SPITZ, J. (1990). NTIMIT: A phonetically balanced, continuous speech, telephone

bandwidth speech database. In *Proc. Int. Conf. Acoust., Speech, Signal Proc.*, pp. 109-112.

[54] JANKOWSKI, C. R., VO H. -D. H, AND LIPPMANN R. P. (1995). A comparison of signal processing front ends for automatic word recognition, *IEEE Trans. on Speech and Audio processing*, 3.4: pp. 286-293.

[55] JANKOWSKI, C. R, QUATIERI, T. F., REYNOLDS, D. A. (1996) Fine structure for speaker identification. In *Proc. Int. Conf. Acoust., Speech, Signal Processing,* pp. 689-692.

[56] JELINEK, F. (1985). The development of an experimental Discrete Dictation Recognizer. *Proc. IEEE* 73: pp. 1616-1624.

[57] JUNQUA, J. C. (1989). *Towards robustness in isolated-word automatic speech recognition*. Ph.D. dissertation, Univ. Nancy I, France.

[58] JUANG, J. –H., WAKITA, H., AND HERMANSKY, H. (1993). Evaluation and optimization of perceptually-based ASR front-end. *IEEE Trans on Speech and Audio processing*, 1: pp. 39-48.

[59] KAJITA, S., AND ITAKURA, F. (1994). Speech analysis and speech recognition using Subband-autocorrelation analysis. *J. Acoust. Soc. Jpn.*

[60] KIM, D., LEE, S., KIL, R.M, AND ZHU, X. (1997). Auditory model for robust speech recognition in real world noisy environments. *Electronics letters*, 33.1: pp. 12-13.

[61] KIM, D., LEE, S., KIL, R.M, AND ZHU, X. (1997). Feature extraction based on auditory representations for robust speech recognition. *Electronics letters,* 33.1: pp. 15-16.

[62] KIM, D., LEE, S., AND KIL, R.M. (1997). Auditory processing of speech signal for robust speech recognition in real-world noisy environments. *IEEE Trans. on Speech and Audio processing*, 7.1: pp. 55-69.

[63] KRASNER, M. (1984). Investigation of text-independent speaker identification techniques under conditions of variable data. In *Proc. Int. Conf. Acoust., Speech, Signal Proc.,* 18B.5.1: pp. 1-4.

[64] LAMEL, L. F., AND CAUVAIN, J. L. (1992). Continuous speech recognition at LIMSI. *DARPA Continuous Speech Recognition Workshop.*

[65] LAMEL, L. F., AND CAUVAIN, J. L. (1993). Cross-lingual experiments with phone recognition. In *Proc. Int. Conf. Acoust., Speech, Signal Processing,* pp. 100-109.

[66] LEE, K. -F., AND HON, H. –W. (1988). Large-vocabulary speaker-independent continuous speech recognition using HMM. In *Proc. Int. Conf. Acoust., Speech, Signal Proc.*

[67] LEE, K. -F. (1989). *Automatics Speech Recognition – The Development of the SPHINIX–System*. Kluwer Academic Publishers.

[68] LEE, K. -F., JON, H. –W., ANDEDDY, R. (1990). An Overview of the SPHINX speech recognition system. *IEEE Trans. on Speech and Audio processing*, 38: pp. 35-44.

[69] LEE, C. -H., RABINER, L.R., AND PIERACCINI, R. (1992). Speech independent continuous speech recognition using continuous density hidden Markov models. In *speech and understanding: recent advances*, P. Laface and R. D. Mori, Eds., vol. F 75 of NATO ASI. Springer-Verlang, pp. 135-163.

[70] LEUNG, H. C., CHIGIER, B., AND GLASS, J. R. (1993). A comparative study of signal representations and classification techniques for speech recognition. In *Proc. Int. Conf. Acoust., Speech, Signal Proc.*, pp. 680-683.

[71] LIBERMAN, M.C. (1978). Auditory-nerve response from cats raised in low-noise chamber. *J. Acoust. Soc. America*, 63: pp. 442-455.

[72] LIBERMAN, M.C. (1982). Single-neuron labeling in the cat auditory nerve. *Sci.*, 16: pp. 1239-1241.

[73] LIU, K. -F. (1994). Phoneme-based speaker-adaptive Mandarin syllable recognition using segmental Bayesian networks. *Communications of COLIPS*, 4: pp. 151-158.

[74] MAMMONE, R. J. ZHANG, X. AND RAMACHANDRAN, R. P. (1996). Robust speaker recognition. *IEEE Signal processing magazine*, pp. 58-71.

[75] MANSOUR, D., AND JUANG, B. H. (1989). The short-time modified coherence representation and noisy speech recognition. *IEEE Trans. Acoust., Speech, Signal Proc.*, 37: pp. 795-804.

[76] MARIANI, J. (1989). Recent advances in speech processing. *Int. Conf. Acoust., Speech, Signal Proc.*, pp. 429-440.

[77] MASHAO, D. J. (1996). *Computations and Evaluations of an optimal feature-set for an HMM-based Recognizer*. PhD Thesis, Brown University, Providence, Rhode Island, USA.

[78] MASHAO, D. J. (1997). Development of the LEMS speech recognizer: Improving performance using feature-sets. *COMSIG*, pp. 157-160.

[79] MATSUI, T., AND FURUI, S. (1992). Comparison of text-independent speaker recognition using VQ-distortion and discrete/continuous HMMs. In *Proc. Int. Conf. Acoustic, Speech, Signal Proc*, II: pp. 157-164.

[80] McLACHLAN, G. (1988). *Mixture models*. New York: Marcel Dekker.

[81] MISTRETTA, B. MORGAN, D., AND RIECK, L. (1990). Experiments with open set speaker identification. *Tech. Rep. VCI-5, Sanders.*

[82] MOORE, B. C. J., AND GLASBERG, B. R. (1983). Suggested formula for calculating auditory-filter bandwidth and excitation patterns. *J. Acoust. Soc. America*, 74: pp. 750-753.

[83] MURTHY, H. A., BEAUFAYS, F., HECK, L. P., WEINTRAUB, M. (1999). Robust text-independent speaker identification over telephone channels. *IEEE Trans. on Speech and audio Processing*, 7.5 : pp. 554-569.

[84] NEWNEYER, L., AND WEINTRAUB, M. (1994). Probabilistic optimum filtering for robust speech recognition. In *Proc. Int. Conf. Acoustic, Speech, Signal Proc.*, I: pp. 417-420.

[85] NIST Speaker recognition workshop, Johns Hopkins University, Baltimore, MD, June 1995.

[86] NOLAZCO FLORES, J. A., AND YOUNG, S. J. (1994). Continuous speech recognition using spectral subtraction and HMM adaptation. In *Proc. Int. Conf. Acoustic, Speech, Signal Proc.*, I: pp. 409-412.

[87] PARSONS, T. W. (1986). *Voice and speech processing*. McGraw-Hill, Inc.

[88] PAUL, D. B. (1984). The Lincoln robust continuous speech recognizer. In *Proc. Int. Conf. Acoustic, Speech, Signal Proc.*, 18A.2: pp. 1-4.

[89] PELLOM, B. L., AND HANSEN, H. L. (1998). An efficient scoring algorithm for Gaussian mixture model based speaker identification. *IEEE signal processing letters*, 5.11: pp. 281-284.

[90] PLUMPE, M. D., QUATIERI, AND REYNOLDS, D. A. (1999). Modeling of the glottal flow derivative waveform with application to speaker identification. *IEEE Trans. on Speech and audio Proc.*, 7.5: pp.569-585.

[91] POLS, L. C. W. (1966). *Spectral analysis and identification of dutch vowels in monosyllabic words.* Ph.D. Thesis, Free Univ., Amsterdam.

[92] PORTER, J. E., AND BOLL, S. F. (1984). Optimal estimators for spectral restoration of noisy speech. In *Proc. Int. Conf. Acoust., Speech, Signal Processing*, 18A.2: pp. 1-4.

[93] PRATT, L. CEBULKA, K. D., AND CLITHEROW, P. (1990). Residual speech signal in the practical application of neural networks technology. *Proc. of the third int. Conf. of Industrial and Engineering Applications of Artificial Intelligence and Expert Systems*, II: pp. 1063-1072.

[94] PROAKIS, J.G. (1983). *Digital communications*. New York: McGraw-Hill series in Electrical Engineering.

[95] RABINER, L.R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. of the IEEE,* 77.2: pp.257-286.

[96] RABINER, L.R, AND LEVISON, S. E. (1981). Isolated and connected word recognition-theory and selected applications. *IEEE Trans on Communications*, 29.5 : pp. 621-659.

[97] RABINER, L.R., AND JUANG, B. –H. (1993). *Fundamentals of Speech Recognition*, Prentice Hall, Englewood Cliffs.

[98] RAJESKARAN, P. J., DODDINGTON, G. R., AND PICONE, J. W. (1986). Recognition of speech under stress and in noise, In *Proc. Int. Conf. Acoust., Speech, Signal Processing,* pp. 733-736.

[99] REETZ, H. (1989). A fast expert program for pitch extraction. *In European Conference on Speech Communication and Technology,* Paris, European Speech Communication Association, 1: pp. 467-479.

[100] REYNOLDS, D. A. (1992*). A Gausssian mixture modeling approach to text-independent speaker identification*. PhD thesis, Georgia Institute of Technology, Antlanta.

[101] REYNOLDS, D. A. (1994). Effects of population size and telephone degradations on speaker identification performance. In *Proc. SPIE Conference on Automatic Systems for the Identification and Inspection of Humans*, July.

[102] REYNOLDS, D. A. (1994). Speaker identification and verification using gaussian mixture speaker models. In *Proc. SPIE Conference on Automatic systems for the identification and Inspections of humans*, pp. 27-30.

[103] REYNOLDS, D. A. (1994). Experimental evaluation of features for robust speaker identification. *IEEE Trans. on Speech and Audio processing*, 2.4: pp. 639-642.

[104] REYNOLDS, D., AND ROSE, R. (1995). Robust text-independent speaker identification using Gaussian mixture models. *IEEE Trans. on Speech and Audio processing*, 3: pp.72-83.

[105] REYNOLDS, D. A. (1995). Large population speaker identification using clean and telephone speech. *IEEE signal processing letters*, 2.3: pp. 46-48.

[106] ROGINSKI, K. (1991). *A neural network phonetic classifier for telephone speech.* M.S thesis, Oregon Graduate Institute of Science and Technology, Oregon.

[107] ROLL, S. F. (1979). Suppression of acoustic noise using spectral subtraction. *IEEE Trans. Acoust., Speech, Signal Proc.*, 27: pp. 254-572.

[108] ROSENBERG, A. (1976). Automatic speaker verification. *Proc. IEEE,* 64:

pp. 475- 487.

[109] ROSE, R. C., AND REYNOLD, D. A. (1990). Text-independent speaker identification using automatic acoustic segmentation. In *Proc. Int. Conf. Acoust., Speech, Signal Processing,* pp. 293-296.

[110] ROSENBERG, A. E. (1992). Recent research in automatic speaker recognition. In *Advances in Speech signal processing,* S. Furui and M. M. Sondhi, Eds, pp. 701-738, Marcel Dekker.

[111] SANDHU, S., AND GHITZA, O. (1995). A comparative study of mel cepstra and EIH for phone classification under adverse conditions. In *Proc. Int. Conf. Acoust., Speech, Signal Processing,* pp. 409-412.

[112] SENEFF, S. (1984). Pitch and spectral estimation of speech based on auditory synchrony model. In *Proc. Int. Conf. Acoust., Speech, Signal Proc.*

[113] SENEFF, S. A. (1986). A computational model for the peripheral auditory system: application to speech recognition research. In *Proc. Int. Conf. Acoust., Speech, Signal Proc.*, pp. 1983-1986.

[114] SENEFF, S. A. (1988). A joint synchrony/mean-rate model of auditory speech Processing. *Journal of Phonetics,* 16: pp. 55-76.

[115] O'SHAUGHNESSY, D. (1986). Speaker recognition. *IEEE ASSP Magazine,* pp. 4-17.

[116] SMITH, R., AND SONDHI, M .M. (1975). Short-term adaptation and incremental responses of single auditory-nerve fibers. *Biological Cybernetics,* 17: pp.169-182.

[117] SOONG, F. K., ROSENBERG, A. E. (1988). On the use of instantaneous and transitional spectral information in speaker recognition. *IEEE Trans. Acoustic, Speech, Signal Processing,* 36: pp. 871-879.

[118] SOONG, F. K., AND SONDHI, M. M. (1988). A frequency-weighted Itakura spectral distortion measure and its applications to speech recognition in noise. *IEEE Trans. Acoustic, Speech, Signal Processing,* pp. 625-628.

[119] STEENEKEN, H. J.M, AND GUERTSEN, F. W. M. (1988). Description of the RSG-10 noise database, *Tech. Rep. IZF 3, TNO Inst. for perception,* Soesteberg, Netherlands.

[120] STERN, R. M., LIU, F. –H., SULLIVAN, T. M., AND ACERO, A. (1992). Multiple approaches to robust speech recognition. *In Proc. DARPA Speech and Natural language Workshop*, Harriman, NY, pp. 274-279.

[121] STEVENS, S. S, AND VOLKMANN, J. (1940). The relation of pitch of frequency: A revised scale. *American J. of Psychology 53*, pp. 85-88.

[122] TOSI, O. (1979). *Voice identification: Theory and legal applications.* Baltimore, MD, Univ. Park Press.

[123] VERGIN, R., AND O'SHAUGHNESSY, D. (1999). Generalized mel frequency cepstral coefficients for large-vocabulary speaker-independent continuous-speech recognition. *IEEE Trans. on Speech and Audio processing*, 7.5, pp. 525-532.

[124] WAIBEL, A., HANAZAWA, T., HINTON, G., AND LANG, K.J. (1989). Phoneme recognition using time-delay neural networks. *IEEE Trans. on Speech and Audio processing*, 37.3: pp. 328-339.

[125] WEBB, J. J., AND RISSANEN, E. L. (1993). Speaker identification experiments using HMMs. In *Proc. Int. Conf. Acoust., Speech, Signal Proc.*

[126] WESTALL, F. A., JOHNTON, R. D., AND LEWIS, A. V. (1998). *Speech technology for telecommunications.* Chapman & Hall.

[127] YOUNG, S. (1992). The general use of tying in phoneme-based HMM speech recognizers. In *Proc. Int. Conf. Acoust., Speech, Signal Proc.*, pp. 569-572.

[128] YOUNG, S. (1996). A review of large-vocabulary continuous-speech recognition. *IEEE Signal Processing Magazine*, pp. 45-57.

[129] ZUE, V., GLASS, J. R., PHILLIPS, M., and SENEFF, S. (1989). The MIT SUMMIT speech recognition system: A progress report. *Proc. DARPA Speech and Natural.*

[130] ZWICKER, E., AND TERHART, E. (1980). Analytical expressions for critical bandwidth as a function of frequency. *J. Acoust. Soc. Amer.*